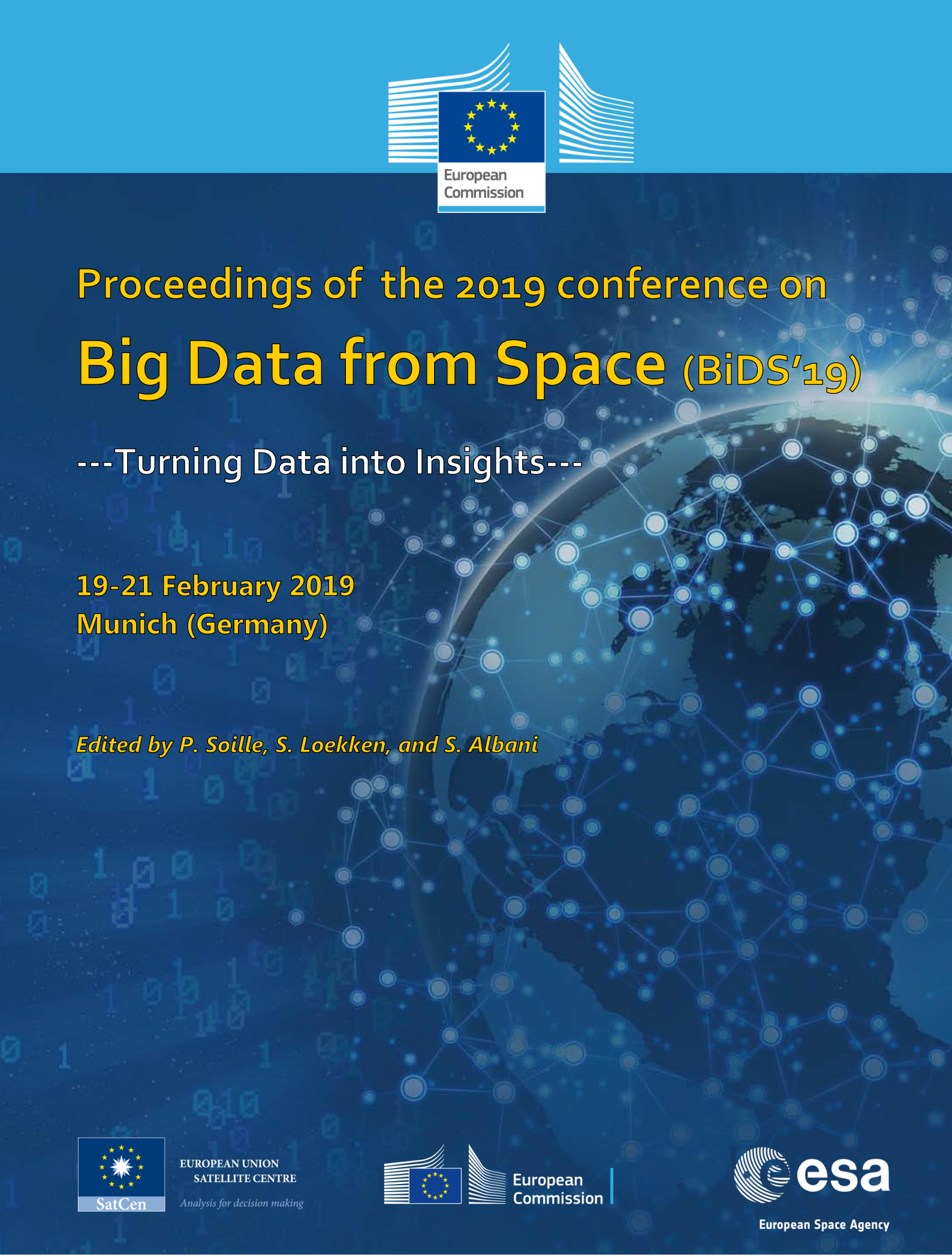


Proceedings of the 2019 conference on

Big Data from Space (BiDS'19)

---Turning Data into Insights---

19-21 February 2019
Munich (Germany)

Edited by P. Soille, S. Loekken, and S. Albani

SatCen
EUROPEAN UNION SATELLITE CENTRE
Analysis for decision making

European Commission

esa
European Space Agency

Joint Research Centre

EUR 29660 EN

Contact information
Name: Pierre Soille
Address: European Commission, Joint Research Centre,
Via Enrico Fermi 2749, TP 267, I-21027 Ispra (VA), Italy
Email: Pierre.Soille@ec.europa.eu
Tel.: +39 0332 78 9111

JRC Science Hub
https://ec.europa.eu/jrc

JRC115761

EUR 29660 EN

PDF ISBN 978-92-76-00034-1 ISSN 1831-9424 doi: 10.2760/848593

Luxembourg: Publications Office of the European Union, 2019

How to cite these proceedings: P. Soille, S. Loekken, and S. Albani (Eds.) Proc. of the 2019 conference on Big Data from Space (BiDS'2019), EUR 29660 EN, Publications Office of the European Union, Luxembourg, 2019, ISBN 978-92-76-00034-1 , doi:10.2760/848593, 2019.

Preface

Big Data from Space refers to the massive spatio-temporal Earth and Space observation data collected by a variety of sensors - ranging from ground based to space-borne - and the synergy with data coming from other sources and communities. This domain is currently facing sharp developments with numerous new initiatives and breakthroughs from intelligent sensors' networks to data science application. These developments are empowering new approaches and applications in various and diverse domains influencing life on earth and societal aspects, from sensing cities, monitoring human settlements and urban areas to climate change and security.

The main objectives of the BiDS'19 Conference are:

- Focus on new paradigms of data intelligence addressing the entire value chain: data processing to extract information, the information analysis to gather knowledge, and knowledge transformation in value;
- Maximise the uptake and impact of multi-source space data;
- Promote the use of platforms and analytical methods to maximise the value extracted for scientific exploration and discovery, societal benefits, commercial exploitation and operational applications;
- Bring together major European actors, including research, industry, institutions, and users, to strengthen the communication and transfer of requirements, methods and technologies, and to reinforce an interdisciplinary approach;
- Promote research and applications in innovative/disruptive data analysis methods;
- Advance the upscale of new solutions from research and innovation to operational use (e.g. for the security domain);
- Promote cross-fertilisation with similar works in other data intensive domains (e.g. high-energy physics, microbiology, social media, etc.).

The BiDS'19 Conference is co-organised by the European Space Agency (ESA), the Joint Research Centre (JRC) of the European Commission, and the European Union Satellite Centre (SatCen). It is hosted by the German Aerospace Center (DLR) in Munich, one of the key European cities with numerous activities focused on space and aerospace developments and applications.

These proceedings consist of a collection of 75 short papers accepted for oral or poster presentation at the conference as a result of the peer-review process by the conference programme committee. The papers are lined up around the topics matching the oral sessions as well as the poster session, also organised by topics.

This 4th edition of the Big Data from Space conference is directed towards 'Turning Data into Insights'. Indeed, while the first editions of the conference concentrated on technologies and platforms capable of sustaining the sharp increase of data streams originating from space sensors, the development of efficient and effective methodologies and algorithms capable of extracting insights from these data is gradually becoming the main challenge. In this context, artificial intelligence and machine learning

techniques have started to play a key role as illustrated by numerous papers of this conference edition. Methodological developments are motivated by the pressing need to extract information on large areas and/or over long time series to better understand the dynamics of the processes that are shaping our planet and indeed our universe in the case of data collected by telescopes. The topic of analysis ready data has also emerged since the last edition and is closely linked with the development of new data cube representations. Big Data from Space is also introducing some new legal challenges and the need for further developments of standards and interoperable interfaces between the growing number of platforms hosting multi-petabyte scale data co-located with processing capabilities. All these topics as well as other generic key aspects of big data are mirrored in dedicated sections of these proceedings. They provide a snapshot of the current research activities, developments, and initiatives in Big Data from Space.

Further to the regular oral and poster contributions, the conference has been enriched by 5 enlightening invited keynote lectures addressing various big data topics of interest to Big Data from Space:

1. *Artificial Intelligence and Data Science in Earth Observation*
   by Xiaoxiang Zhu (DLR-IMF, Head of Department "EO Data Science")

2. *Mosaics in Big Data: Stratosphere, Apache Flink, and Beyond*
   by Volker Markl (Technical University Berlin, Data Analytics Lab & DFKI)

3. *Overview of JPL data science for Earth science*
   by Tomas Huang (NASA JPL, Computer Science for Data Intensive Applications)

4. *European Data Relay System Achievements and Capabilities*
   by Harald Hauschildt (European Space Agency, Telecom & Integrated Applications Dept.)

5. *Machine learning in Earth Observation data analysis*
   by Gustau Camps-Valls (Universitat de València, Dpt. Enginyeria Electrònica (DIE), Image Processing Laboratory (IPL))

Additional conference materials such as electronic version of the slides presented at the conference, including those regarding the opening session talks and keynote lectures, are available on the conference website: www.bigdatafromspace2019.org.

Great thanks goes to all authors and presenters of BiDS'19 as well as the numerous participants (over 600 registrations from more than 50 different countries). Together, they have ensured the success of the 2019 conference on Big Data from Space. Special thanks goes to the Programme Committee members and the additional reviewers for their thorough reviews and detailed comments that were taken into account by the authors when preparing the final version of their paper included in these proceedings.

This edition of the BiDS conference is deeply grateful to the German Aerospace Center (DLR) for its strong support in having BiDS'19 hosted in Munich.

Pierre Soille, Sveinung Loekken, and Sergio Albani

Conference Chairs

General Chair

Sergio Albani	European Union Satellite Centre (SatCen)

Co-Chairs

Jutta Graf	German Aerospace Center (DLR), Germany
Pierre Soille	European Commission, Joint Research Centre (JRC)
Sveinung Loekken	European Space Agency (ESA), ESRIN, Italy

Organising Committee

Sergio Albani	European Union Satellite Centre (SatCen)
Simon Baillarin	Centre National d'Études Spatiales (CNES), France
Peter Baumann	Jacobs University Bremen, Germany
Lorenzo Bruzzone	University of Trento, Italy
Mihai Datcu	German Aerospace Center (DLR), Germany
Raffaella Franco	European Space Agency (ESA), ESTEC, The Netherlands
Jean-Pierre Gleyzes	Centre National d'Études Spatiales (CNES), France
Jutta Graf	German Aerospace Center (DLR), Germany
Michele Iapaolo	European Space Agency (ESA), ESRIN, Italy
Pieter Kempeneers	European Commission, Joint Research Centre (JRC)
Sveinung Loekken	European Space Agency (ESA), ESRIN, Italy
Vicente Navarro	European Space Agency (ESA), ESAC, Spain
Pierre Soille	European Commission, Joint Research Centre (JRC)
Juan Luis Valero	European Union Satellite Centre (SatCen)
Raffaelle Vitulli	European Space Agency (ESA), ESTEC, The Netherlands

Local Organising Committee

Doris Klein	German Aerospace Center (DLR), Germany
Katrin Molch	German Aerospace Center (DLR), Germany
Johannes Schmidt	German Aerospace Center (DLR), Germany

Programme Committee

Selim Aksoy	Bilkent University, Turkey
Mirko Albani	European Space Agency
Sergio Albani	European Union Satellite Centre (SatCen)
Christophe Arviset	European Space Agency (ESA), ESAC, Spain

Simon Baillarin	Centre National d'Études Spatiales (CNES), France
Peter Baumann	Jacobs University Bremen
Francesca Bovolo	Fondazione Bruno Kessler
Lorenzo Bruzzone	University of Trento
Francesco Casu	IREA, National Research Council (CNR), Italy
Esther Conway	Science and Technology Facilities Council (UK)
Christina Corbane	European Commission, Joint Research Centre (JRC)
Mihai Datcu	German Aerospace Center (DLR), Germany
Begum Demir	TU Berlin
Yves-Louis Desnos	European Space Agency (ESA), ESRIN, Italy
Liang Feng	Universuty of Edinburgh
Raffaella Franco	European Space Agency (ESA), ESTEC, The Netherlands
Steffen Fritz	International Institute for Applied Systems Analysis (IIASA), Austria
Paolo Gamba	University of Pavia, Italy
Jean-Pierre Gleyzes	Centre National d'Études Spatiales (CNES), France
Jutta Graf	German Aerospace Center (DLR), Germany
Jacopo Grazzini	DG ESTAT (Eurostat) - European Commission
Harm Greidanus	European Commission, Joint Research Centre (JRC)
Steve Groom	IPAC/Caltech, USA
Michele Iapaolo	European Space Agency (ESA), ESRIN, Italy
Jordi Inglada	Centre National d'Études Spatiales (CNES)–CESBIO, France
Francois Jocteur Monrozier	Centre National d'Études Spatiales (CNES), France
Pieter Kempeneers	European Commission, Joint Research Centre (JRC)
Doris Klein	German Aerospace Center (DLR), Germany
Riccardo Lanari	IREA, National Research Council (CNR), Italy
Henri Laur	European Space Agency (ESA), ESRIN, Italy
Samantha Lavender	Pixalytics Ltd
Michele Lazzarini	European Union Satellite Centre (SatCen)
Jacqueline Le Moigne	National Aeronautics and Space Administration (NASA), USA
Sébastien Lefèvre	Université de Bretagne Sud, France
Guido Lemoine	European Commission, Joint Research Centre (JRC)
Sveinung Loekken	European Space Agency (ESA), ESRIN, Italy
Michele Manunta	IREA, National Research Council (CNR), Italy
Lewis Mcgibbney	National Aeronautics and Space Administration (NASA), USA
Katrin Molch	German Aerospace Center (DLR), Germany
Vicente Navarro	European Space Agency (ESA), ESAC, Spain
Allan A. Nielsen	Technical University of Denmark
Simon Oliver	Geoscience Australia
Frank Ostermann	Faculty of Geo-Information Science and Earth Observation (ITC), The Netherlands
Edzer Pebesma	Inst. for geoinformatics, Univ of Muenster, Germany
Jean-François Pekel	European Commission, Joint Research Centre (JRC)
Rui Santos	European Space Agency (ESA), ESOC, Germany
Michael Schick	EUMETSAT

John Schnase	National Aeronautics and Space Administration (NASA), USA
Pierre Soille	European Commission, Joint Research Centre (JRC)
Peter Strobl	European Commission, Joint Research Centre (JRC)
Vasileios Syrris	European Commission, Joint Research Centre (JRC)
Corina Vaduva	University Politehnica of Bucharest, Romania
Juan Luis Valero	European Union Satellite Centre (SatCen)
Joost van Bemmelen	European Space Agency (ESA), ESRIN, Italy
Raffaele Vitulli	European Space Agency (ESA), ESTEC, The Netherlands
Julia Wagemann	European Centre for Medium-Range Weather Forecasts, U.K.
Wolfgang Wagner	Vienna University of Technology, Austria
Gui-Song Xia	Wuhan University, China
Xiaoxiang Zhu	German Aerospace Center (DLR), Germany& Technical University of Munich, Germany

Additional Reviewers

Beck, Pieter	European Commission, Joint Research Centre (JRC)
Chen, Wei	Institute of Remote Sensing and Digital Earth (RADI), China
Della Vecchia, Andrea	European Space Agency (ESA)
Lazzarini, Michele	European Union Satellite Centre (SatCen)
Maggio, Iolanda	European Space Agency (ESA)
Mirmahboub, Behzad	Université de Bretagne Sud, France
Mou, Lichao	German Aerospace Center (DLR), Germany
Popescu, Anca	European Union Satellite Centre (SatCen)
Verhegghen, Astrid	European Commission, Joint Research Centre (JRC)
Wang, Yuanyuan	Technical University of Munich (TUM)

Table of Contents

Posters: Processing Platforms and Cloud Computing

Posters: Data Hubs, Catalog, and Preservation

ROAD PASSABILITY ESTIMATION USING DEEP NEURAL NETWORKS AND SATELLITE IMAGE PATCHES

Anastasia Moumtzidou, Marios Bakratsas, Stelios Andreadis, Ilias Gialampoukidis, Stefanos Vrochidis, Ioannis Kompatsiaris *

Centre for Research & Technology Hellas
Information Technologies Institute
Thessaloniki, Greece

ABSTRACT

Artificial Intelligence (AI) technologies are getting deeper and deeper into remote sensing and satellite image processing offering value-added products and services in a real-time manner. Deep learning techniques applied on visual content are able to infer accurate decisions about concepts and events in an automatic way, based on Deep Convolutional Neural Networks which are trained on very large external image collections in order to transfer knowledge from them to the considered task. Existing emergency management services focus on the detection of flooded areas, without the possibility to infer if a road from point A to a point B is passable or not. To that end, we propose an automatic road passability service that is able to deliver the parts of the road network which are not passable, using satellite image patches. Experiments and fine-tuning on an annotated benchmark collection indicates the most suitable model among several Deep Convolutional Neural Networks.

Index Terms— Road passability, Deep Convolutional Neural Networks, Crisis Management, Road Network

1. INTRODUCTION

The high applicability of the Artificial Intelligence (AI) has led to the utilization of its technologies in order to develop and advance numerous other fields, among them remote sensing. Applying deep learning techniques on satellite images can offer an automatic identification of concepts or events. More specifically, we are based on Deep Convolutional Neural Networks (DCNNs) that are pre-trained on an external dataset of millions of images and use them to classify satellite imagery, a technique known also as transfer learning.

Our field of application is the Emergency Management applications, a managerial function that seeks to cope with hazards and disasters. While state-of-the-art mainly focuses on the detection of flooded areas in general, we target an explicit problem: starting from a point A to a point B, is a road passable or not due to a flood? Therefore, we introduce a road passability method that can automatically decide whether a roadway depicted in a satellite image is clear and able to be traveled.

The paper is structured as follows. In Section 2 we examine the existing works that are related to the problems of road extraction and flood detection. Section 3 describes the methodology, while Section 4 concerns the experiments and presents the results. Finally, Section 5 concludes and discusses future enhancements.

2. RELATED WORK

Road passability relies on two major sub-problems of remote sensing, being a combination of road extraction and flood detection procedures, with the most recent trends based on the exploitation of neural networks' capabilities. In the following we present the recent advances in both directions.

Road extraction detects road segments, as also defined in [1] where it is proposed to extract the road components from satellite images using Laplacian of Gaussian operator. The image is pre-processed to identify the color space components. At start, a panchromatic and a multispectral image of an area are combined (fused) to obtain more details of the image. Then, objects are identified using HSY color models components. Trying to distinguish roads from sandy regions, hue and luminance may have similar values but can be distinguished using saturation. A morphological method is applied to remove the unwanted objects in the image. In a more recent approach, the work of [3] explores 3 different Fully-Convolutional Neural Networks (FCNNs): FCN-8s with a VGG-19 backbone, Deep Residual U-Net0 and DeepLabv3+ for semantic segmentation. All networks were trained from scratch, where a considerable performance drop is noticed when using weights pretrained on ImageNet, due to the different nature of SAR images compared to optical ones. Adjusting the object segmentation, the task changes from a binary classification to a binary regression model, and instead of predicting each pixel as either road or background, the network weighs how likely it is for each pixel

*This work was supported by EOPEN project, partially funded by the European Commission, under the contract number H2020-776019.

Fig. 1. Part of a Web application that exploits the road passability service, developed for the purposes of H2020-EOPEN project.

to be a road. Due to Object awareness in FCNN, the predicted roads are sometimes disconnected at intersections, requiring re-connection of loose segments. In another work, in order to improve performance at heterogeneous areas (cars, trees on the road) a method on Generative Adversarial Networks (GAN) [6] is proposed to handle road detection. For the segmentation model, the so-called "Segnet" is used to generate a pixel-wise classification map. The GAN defines two models; the generative model, which is used to stimulate the data probability distribution, and the discriminative model, which is used to find whether a sample is coming from the generative model or the ground truth map. The generative and the discriminative models together form an adversarial network. Contrary to these approaches, we aim to infer whether a satellite image patch contains a passable road segment or not, without the need to segment the image patch into "road" and "no-road" regions.

Flood detection has been a popular problem in the remote sensing community, while nowadays the focus is on the use of Neural Networks, such as in [4], where the Fully-Convolutional Network (FCN), a variant of VGG16 on Gaofen-3 SAR images, is utilized for flood mapping. FCN demonstrates robustness to speckle noise in SAR images. Speckle noise is not filtered, making the deep learning model more universal (data augmentation). To make the model less complex, 7 x 7 kernels are replaced with 3 x 3 kernels greatly reducing conv6 parameters. In [5] the most widely used criteria performances, namely coefficient of determination (R2), sum squared error (SSE), mean squared error (MSE), and root mean squared error (RMSE) are used to optimize the performance of the Artificial Neural Network (ANN). Each method is estimated from the ANN predicted values and the measured discharges (targets). Seven input nodes, each representing flood causative parameters, including rainfall, slope, elevation, soil, geology, flow accumulation, and land use are used during the ANN modeling. There is little variation in maximum and minimum connection weights between the input and the hidden layers nodes except from the rainfall parameter. Rainfall factor is the main factor in the training of the neural network. The sensitivity analysis has shown that the elevation is the most important factor for flood susceptibility mapping. The approach in [8] is based on the segmentation of a single SAR image using self-organizing Kohonen maps (SOMs) and further image classication using auxiliary information on water bodies that could be derived, from optical satellite images. A moving window is applied to process the image and spatial connection between the image pixels is taken into account. Neural networks weights are adjusted automatically using ground-truth training data. In contrast, we propose a unifying approach to infer whether a road is passable or not, due to a severe flood event. We examine state-of-the-art classification methods with transfer learning, aiming to develop an effective road passability estimation service for the case of flooded road networks.

3. METHODOLOGY

3.1. Road passability service

In order to showcase the applicability of the proposed road passability service, we demonstrate a Web user interface that involves a classification service that adopts a DCNN architecture. As seen in Figure 1, the user is presented with a collection of satellite images, which are accompanied by their metadata (i.e., date, location, type). When an image is clicked, it is partitioned to smaller pieces and the classification method is performed to every piece. If a passable road is detected, then a green border appears around the image segment. Otherwise, a red border indicates the detection of a non-passable road. In case that no roads are recognized inside the image, no border is shown. With the results clearly illustrated, one can easily evaluate the effectiveness as well as the usefulness of the service.

3.2. Model selection and implementation

In order to classify satellite images to the class "road passability" we build models by using pretrained Convolutional Neural Networks (CNN). Namely, we experimented with the following models: VGG-19 [7], Inception-v3 [9], and ResNet [2]. VGG was originally developed for the ImageNet dataset by the Visual Geometry Group at the University of Oxford. The model involves 19 layers and it has as input images of size 224 x 224. Inception-v3 is another ImageNet-optimized model. It is developed by Google and has a strong emphasis on making scaling to deep networks computationally efficient, having as input 299 x 299 images. Finally, ResNet-50 is a model developed by Microsoft Research using a structure that uses residual functions to help add considerable stability to deep networks, using as input 224 x 224 images. For each of the aforementioned networks, we performed fine-tuning which involved removing the last pooling layers and replacing it with a new pooling layer with a softmax activation function with size 2 given that our aim is to recognize whether there is evidence of road passability or not.

For the implementation we used TensorFlow[1] and the open-source neural network Python package Keras[2] for developing our models. In general, Keras package simplifies the training of new CNN networks by modifying easily the network structure and the pre-trained weights, freezing the weights in the imported network and eventually training the weights in the newly added layers, in order to combine existing knowledge from the imported weights with the gained knowledge from the domain-specific collection of satellite images with ground-truth annotation on road passability.

4. EXPERIMENTS

4.1. Dataset description

The dataset consists of 1,437 satellite images provided for the MediaEval 2018 Satellite Task "Emergency Response for Flooding Events"[3] - data for "Flood detection in satellite images". These are satellite image patches of flooded areas that were manually annotated with a single label to indicate whether the road depicted is passable or not due to floods. The dataset was randomly split into a training a validation set. The training set contained 1,000 images, while the validation set the remaining 437 images.

4.2. Settings

Several experiments were run in order to find the best performing model. The parameters that were tuned concern the learning rate, the batch size and the optimizer function. Specifically, the values considered for the aforementioned parameters were the following: learning rate values = $\{0.001, 0.01, 0.1\}$, batch size values = $\{32, 64, 128, 256\}$, and the optimizer functions = {Adam, Stochastic Gradient Descent (SGD)}. Finally, the epoch was set to 35 and the loss function considered was the sparse_categorical_crossentropy.

4.3. Results

To evaluate the performance of the different networks we considered accuracy as the evaluation metric. The results of our analysis are shown in Tables 1 and 2 and in general they present the accuracy of the train and the validation set for the four networks (i.e. VGG-19, Inception_v3, ResNet-50, ResNet-101) for two widely used optimizers, i.e. Adam and SGD. Specifically, Table 1 shows how the learning parameter affects the performance of the networks. After a careful observation we can deduce that the networks perform better for the lower values of the learning rate, as they reach an average accuracy of 81.2% and 78.5% for learning rates 0.001 and 0.01 respectively.

In the sequel, we experimented with the batch size parameter and observed the impact on the networks accuracy (Table 2). The conclusion that rises from this experiment is that the increase of the batch size generally improves the accuracy. The best values of accuracy are achieved by ResNet-50 for batch size 256 and Adam optimizer (88.2%) and the Inception_v3 for batch size 128 and Adam optimizer (89.9%) (highlighted in bold). However, the accuracy of the validation set for the Inception_v3 is significantly lower than that of ResNet-50, probably due to over-fitting reasons.

5. CONCLUSION AND FUTURE WORK

In this work we presented our approach in road passability from satellite images using the recent advances in Deep Neural Networks. Tweaking the core settings of the network significant improvement in accuracy can be achieved. Better results appear with lower values of the learning ratio, while increasing the batch size improves accuracy, up to a certain level so as to avoid over-fitting. Additionally, this work highlights the necessity to evaluate alternative ways of fine-tuning pre-trained networks to compare performance differentiation.

Future work includes the combination of our approach with RCNN region proposal neural networks, to inherently perform semantic segmentation, as also described in [10], fusing heterogeneous data sources, to also highlight the road segments, in case they are not available through an external source as a GIS layer or any other format.

[1] https://www.tensorflow.org/
[2] https://keras.io/
[3] http://www.multimediaeval.org/mediaeval2018/multimediasatellite/

Table 1. Neural networks accuracy for different learning rate values.

		Learning rate 0.001		Learning rate 0.01		Learning rate 0.1	
DCNN	Optimizer	Dev. Set Acc.	Valid. Set Acc.	Dev. Set Acc.	Valid. Set Acc.	Dev. Set Acc.	Valid. Set Acc.
VGG-19	Adam	0,85	0,6911	0,5640	0,4851	0,5700	0,5904
VGG-19	SGD	0,8600	0,7140	0,8380	0,7277	-	-
Inception_v3	Adam	0,796	0,6018	0,8120	0,5973	0,4190	0,4348
Inception_v3	SGD	0,7050	0,6590	0,8100	0,6499	0,8040	0,5995
ResNet-50	Adam	0,872	0,6247	0,8400	0,6453	0,5670	0,5973
ResNet-50	SGD	0,789	0,6865	0,8470	0,5538	0,8060	0,6796
ResNet-101	Adam	0,866	0,5515	0,8450	0,4668	0,7470	0,6590
ResNet-101	SGD	0,799	0,6041	0,8700	0,4668	0,8450	0,5858

Table 2. Neural networks accuracy for different batch size values for best performing learning rates.

			Batch size 32		Batch size 64		Batch size 128		Batch size 256	
DCNN	Learning rate	Optimizer	Dev. Set Acc.	Valid. Set Acc.	Dev. Set Acc.	Valid. Set Acc.	Dev. Set Acc.	Valid. Set Acc.	Dev. Set Acc.	Valid. Set Acc.
VGG-19	0,01	Adam	0,861	0,7666	0,8610	0,7666	0,8610	**0,7667**	-	-
VGG-19	0,001	SGD	0,876	0,7071	0,8630	0,7117	0,8740	0,7162	-	-
Inception_v3	0,01	Adam	0,788	0,6247	0,8610	0,5789	0,8990	0,5629	0,8800	0,5378
Inception_v3	0,001	SGD	0,792	0,5950	0,8330	0,6224	0,8480	0,5973	0,8550	0,5995
ResNet-50	0,01	Adam	0,833	0,4943	0,8640	0,6957	0,8720	0,7094	0,8820	0,7323
ResNet-50	0,001	SGD	0,804	0,6911	0,8310	0,7094	0,8390	0,7140	0,8390	0,7185
ResNet-101	0,1	Adam	0,86	0,5492	0,8710	0,5126	0,8850	0,5126	0,8890	0,4989
ResNet-101	0,001	SGD	0,789	0,5835	0,8260	0,5995	0,8380	0,5881	0,8390	0,5812

REFERENCES

[1] Reshma Suresh Babu, B Radhakrishnan, and L Padma Suresh. Detection and extraction of roads from satellite images based on laplacian of gaussian operator. In *Emerging Technological Trends (ICETT), International Conference on*, pages 1–7. IEEE, 2016.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] Corentin Henry, Seyed Majid Azimi, and Nina Merkle. Road segmentation in sar satellite images with deep fully-convolutional neural networks. *arXiv preprint arXiv:1802.01445*, 2018.

[4] Wenchao Kang, Yuming Xiang, Feng Wang, Ling Wan, and Hongjian You. Flood detection in gaofen-3 sar images via fully convolutional networks. *Sensors*, 18(9): 2915, 2018.

[5] Masoud Bakhtyari Kia, Saied Pirasteh, Biswajeet Pradhan, Ahmad Rodzi Mahmud, Wan Nor Azmin Sulaiman, and Abbas Moradi. An artificial neural network model for flood simulation using gis: Johor river basin, malaysia. *Environmental Earth Sciences*, 67(1):251–264, 2012.

[6] Qian Shi, Xiaoping Liu, and Xia Li. Road detection from remote sensing images by generative adversarial networks. *IEEE access*, 6:25486–25494, 2018.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL http://arxiv.org/abs/1409.1556.

[8] Sergii Skakun. A neural network approach to flood mapping using satellite imagery. *Computing and Informatics*, 29(6):1013–1024, 2012.

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[10] Wei Yao, Dimitrios Marmanis, and Mihai Datcu. Semantic segmentation using deep neural networks for sar and optical image pairs. In *Proceedings of the ESA Big data from space conference*, pages 1–4, 2017.

MULTI-TASK DEEP LEARNING FROM SENTINEL-1 SAR: SHIP DETECTION, CLASSIFICATION AND LENGTH ESTIMATION

C. Dechesne [1], *S. Lefèvre* [2], *R. Vadaine*[3], *G. Hajduch*[3], *R. Fablet* [1]

[1] IMT Atlantique – Lab-STICC, UMR CNRS 6285, Brest, FR
[2] Univ. Bretagne Sud – IRISA, UMR CNRS 6074, Vannes, FR
[3] Collecte Localisation Satellites, Brest, FR

ABSTRACT

The detection of inshore and offshore ships is an important issue in both military and civilian fields. It helps monitoring fisheries, managing maritime traffics, ensuring safety of coast and sea, etc. In operational contexts, ship detection is traditionally performed by a human observer who identifies all kind of ships from visual analysis on remote sensing images. Such a task is very time consuming and cannot be conducted at a very large scale, while Sentinel-1 SAR data now provides regular, worldwide coverage. Meanwhile, with the emergence of GPUs, deep learning methods are now established as a state-of-the-art solution for computer vision, replacing human intervention in many contexts. They have been shown to be adapted for ship detection and recognition, most often with very high resolution SAR or optical imagery. In this paper, we go one step further and propose a deep neural network for the joint detection, classification and length estimation of ships from SAR Sentinel-1 data. We benefit from synergies between AIS (Automatic Identification System) and Sentinel-1 data to build significant training datasets. We then design a multi-task neural network architecture composed of one joint convolutional network connected to three networks dedicated to the different tasks, namely ship detection, classification and length estimation. Experimental assessment showed our network provides satisfactory results, with relevant ship presence probability maps, accurate classification and length estimation.

Index Terms— Deep neural network, Sentinel-1 SAR images, Ship characterization, Multi-task learning

1. INTRODUCTION

Deep learning is considered as one of the major breakthrough related to big data and computer vision [8]. It has become very popular and successful in many fields including remote sensing [14]. Deep learning is a paradigm for representation learning and is based on multiple levels of information. When applied on visual data such as images, it is usually achieved by means of convolutional neural networks. These networks consist of multiple layers (such as convolution, pooling, fully connected and normalization layers) aiming to transform original data (raw input) into higher semantics representation. With the composition of enough such operations, very complex functions can be learned. For classification tasks, higher representation layers amplify aspects of the input that are important for discrimination and discard irrelevant variations. For humans, it is simple through visual inspection to know what objects are in an image, where they are, and how they interact in a very fast and accurate way, allowing to perform complex tasks. Fast, accurate algorithms for object detection are thus sought to allow computers to perform such tasks, at a much larger scale than humans can achieve.

Sentinel-1 SAR images are well adapted for ship detection. Almost all coastal zones and shipping routes are covered by Interferometric Wide Swath Mode (IW), while Extra-Wide Swath Mode (EW) acquires data over open oceans, providing a global coverage for sea-oriented applications. Such images, combined with the Automatic Identification System (AIS), represent a large amount of data that can be employed for deep learning models. AIS provides meaningful and relevant information about ships (such as position, type, length, rate of turn, speed over ground, etc.). Combining these two data sources could ease accurate detection and estimation of ship parameters from SAR images, which remains a very challenging task. Indeed, detecting inshore and offshore ships is critical in both military and civilian fields (e.g. for monitoring of fisheries, management of maritime traffics, safety of coast and sea, etc). In operational contexts, the approaches used so far still rely on manual visual interpretations that are time-consuming, possibly error-prone, and definitely not able to cope with big data issues. On the contrary, the availability of satellite data such as Sentinel-1 SAR makes possible efficient and accurate ship detection.

Among existing methods for ship detection from SAR images, Constant False Alarm Rate (CFAR)-based methods have been widely used to detect ships in the sea [9, 1]. The advantage of such methods is their reliability and high efficiency. As the choice of features has an impact on the performance of discrimination, deep neural networks took the lead thanks to their ability to extract (or learn) features that are richer

than hand-crafted (or expert) features. In [10], a framework named Sea-Land Segmentation-based Convolutional Neural Network (SLS-CNN) was proposed for ship detection, combined with the use of saliency computation. A modified Faster R-CNN based on CFAR algorithm for SAR ship detection was proposed in [4] with good detection performances. In [6], a method categorizing ship targets from SAR images using texture features in artificial neural networks (TF-ANN) was proposed. The TF-ANN method selects an appropriate texture feature for SAR images and uses the feature as the input of neural network to extract ship pixels from sea ones. [12] employed highway network for ship detection in SAR images and achieved good results, especially in detecting correctly false positive. These state-of-the-art approaches focused on ship detection in SAR images. In this paper, we rather aim to achieve the recognition of ship types and their length estimation, which to our knowledge has not been dealt with before. Furthermore, our network is able to provide pixelwise probability map of ship presence, that can be further threshold to also allow ship detection.

2. PROPOSED APPROACH

2.1. Creation of reference datasets

With a view to implement deep learning strategies, we first address the creation of reference datasets from the synergy between AIS data and Sentinel-1 SAR data. AIS data are interpolated in order to know the ship location when the SAR images have been captured. Thus it is possible to know the precise location of the ship in the SAR image and its related information (in our case, length and type). The footprint of the ship is obtained by thresholding the SAR image in the area where it is located.

2.2. Proposed framework

The proposed multi-task framework is based on two stages, with a first common part and then three task-oriented branches for ship detection, classification and length estimation, respectively (see Fig. 1). The first part is a convolutional network made of 5 layers. It is followed by the task-oriented branches. For the detection task, the output consists in a pixel-wise probability of presence of ship. It is composed of 4 convolutional and 1 fully connected layer. For the classification task, we consider 4 ship classes (Cargo, Tanker, Fishing and Passenger). The branch is composed of 4 convolutional and 2 fully connected layers. The last task is related to the length estimation. The related branch is composed of 4 convolutional and 5 fully connected layers.

Such settings are commonly employed in deep learning methods [11]. All the activations of the convolutional layers and fully connected layers are ReLu [7]. Other activation functions are employed for the output layers: a sigmoid for the detection, a softmax activation for the classification, and a linear activation is employed for the length estimation.

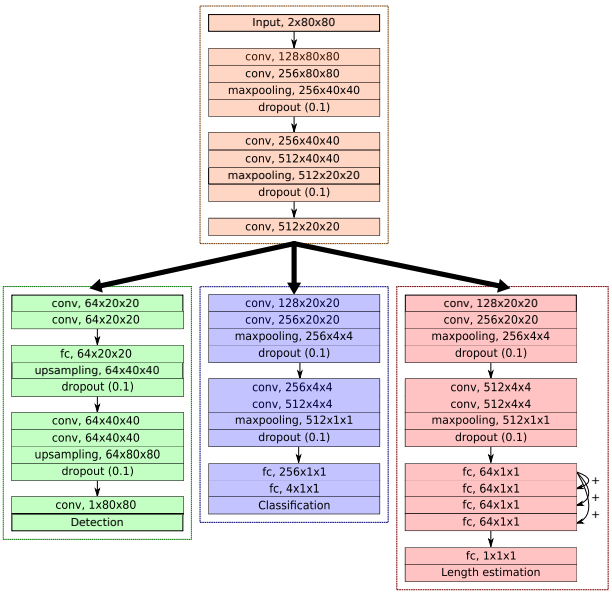


Fig. 1: Proposed multi-task architecture for ship detection, classification and length estimation from Sentinel-1 SAR.

2.3. Loss functions

2.3.1. *Detection*

The detection output is the probability of ship presence. We thus employ a binary cross-entropy loss, which is defined by:

$$L_{det} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k\in I}(y_k \log(p(k)) + (1-y_k)\log(1-p(k))), \tag{1}$$

where N is the number of samples, k is a pixel of the output detection image I, y_k is the ground truth of ship presence (0 or 1), and $p(k)$ is the predicted probability of ship presence.

2.3.2. *Classification*

The output for the last classification layer is the probability that the input image corresponds to one of the considered ship types. We use here the categorical cross-entropy loss:

$$L_{class} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{n_c}(y_{o,c}\log(p_{o,c})), \tag{2}$$

where N is the number of samples, n_c is the number of classes (here, $n_c = 4$), $y_{o,c}$ is a binary indicator (0 or 1) if class label c is the correct classification for observation o and $p_{o,c}$ is the predicted probability for the observation o to belong to c.

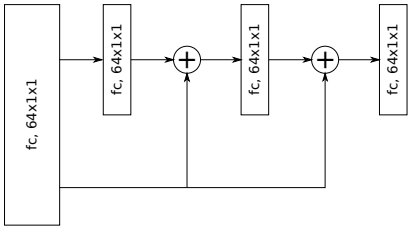


Fig. 2: Difference propagation flowchart in the fully-connected layers.

2.3.3. *Length*

In the length estimation network, the 4 fully-connected layers of shape (64×1×1) are connected to each other (see Fig. 2). The idea is to propagate the difference between the first layer and the current layer and is related to residual learning [3]. We use here the mean squared error defined as

$$L_{length} = \frac{1}{N}\sum_{n=1}^{N}(l_{pred} - l_{true})^2, \tag{3}$$

where N is the number of samples, l_{pred} is the predicted length and l_{true} is the true length.

2.3.4. *End-to-end training*

We define the loss function of the whole network as

$$L = L_{det} + L_{class} + L_{length}. \tag{4}$$

Each specific loss employed to design the loss of the whole network could have been weighted. But even if the range is not uniform among the different losses, we observed no effect on the optimization process. Thus, we have decided to equally weight each specific loss in order perform each task with the same importance. Our network is trained end-to-end using RMSProp optimizer [13]. The weights of the network are updated by using a learning rate of 1e-4 and a learning rate decay over each update of 1e-6 over the 500 iterations.

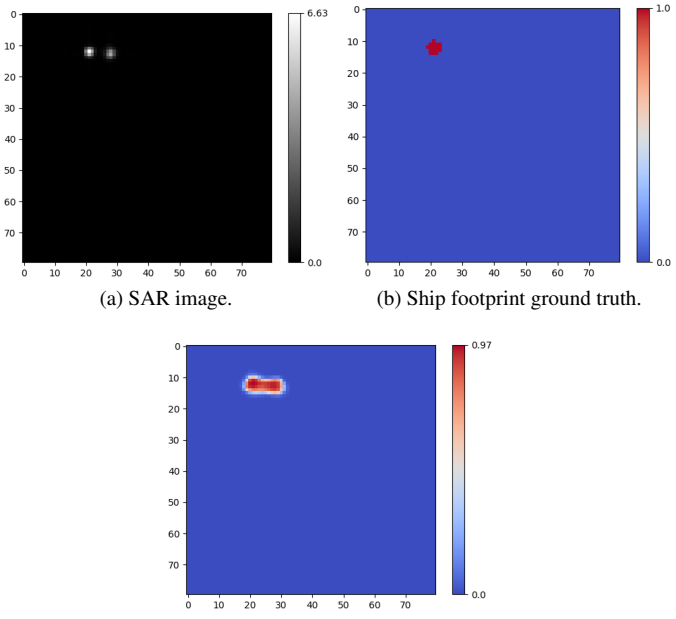
3. DATA

In our experiments, we consider a dataset composed of 18,894 SAR images of size 400×400 pixels and having a 10 m resolution. Each image is accompanied with the incidence angle since it impacts the backscatter intensity of the signal. We rely on Automatic Identification System (AIS) to extract images that contain a ship in their center. Furthermore, AIS also provides us with information about the ship type and length. The dataset is very unbalanced (10,430 instances of Tanker and only 1,071 instances of Passenger), thus requiring dedicated strategy [5]. Here we simply decided to enlarge our dataset through data augmentation with translations and rotations in order to have 20,000 balanced images. The images employed to train our network are 80×80 images containing ships (not necessarily in the center). The ship footprint ground truth is generated by thresholding the SAR image since we precisely know the location of the ship (i.e. it is the brightest pixel of the SAR image). The obtained footprint is not perfect (see Fig. 3b) but is sufficient in order to train the network. Let us note that a CFAR approach could have been employed in order to extract more precisely the ship footprint [9].

4. RESULTS

We train and test our network on a PC with a single NVIDIA GTX 1080 Ti, an Intel Xeon W-2145 CPU 3.70GHz and 64GB RAM (with a Keras [2] implementation). For a 80×80 image, our method can run at 55 frames per second.

The network is trained using 16,000 images from the augmented dataset and the remaining 4,000 images are used for validation. Since our goal is to correctly estimate ships parameters (namely length and type), only results on small images are presented. Accurate evaluation of ship detection is difficult, so we conduct a visual inspection to confirm that the detection is well performed by our network (see Fig. 3). Let us note that the detection task has been widely addressed in the literature [10, 4, 6] and is not our main purpose here.



(a) SAR image. (b) Ship footprint ground truth.

(c) Detection result.

Fig. 3: SAR image (with backscatter intensity), the generated ground truth and result of detection from the network.

To our knowledge, the length estimation is a task that has never been investigated using learning-based schemes yet. Our framework performs well with very promising results. The length is slightly under-estimated: -2.4 m ± 9.5 m, which is very good regarding the spatial resolution of the Sentinel-1 SAR data. Indeed, having only the ship footprint and the spatial resolution of the image is not sufficient and often leads to an over-estimation of the length. The classification task is

of high importance. Table 1 gives the confusion matrix, and several accuracy metrics are also presented in Table 2. The confusion matrix shows some light confusions for passenger ships, decreasing slightly the precision for this class. Some fishing ships are classified as passenger ships impacting the recall for this class. For the tanker and cargo ships, the classification is very accurate. The accuracy metrics confirm these satisfactory results with an overall accuracy and a mean F-score of 95.4%.

Label	Tanker	Cargo	Fishing	Passenger	Recall
Tanker	978	7	6	9	97.8
Cargo	8	946	7	39	94.6
Fishing	1	15	934	50	93.4
Passenger	5	13	24	958	95.8
Precision	98.6	96.4	96.2	90.7	

Table 1: Confusion matrix of ship classification.

Label	Tanker	Cargo	Fishing	Passenger	Overall
IoU	96.5	91.4	90.1	87.3	91.3
F-score	98.2	95.5	94.8	93.2	95.4
Accuracy	99.1	97.8	97.4	96.5	95.4
κ	0.98	0.94	0.93	0.91	0.95

Table 2: Accuracy metrics of ship classification.

5. CONCLUSION

In this paper, a multi-task deep neural network approach was introduced to address joint detection (i.e. pixelwise probability maps), classification and length estimation for ships in Sentinel-1 SAR images. We exploit AIS-Sentinel-1 synergies to automatically build reference datasets for training and evaluation purposes. Regarding the considered architecture, a mutual convolutional branch transforms raw inputs into meaningful information. Such information is fed into specific branches for each of the three considered tasks. In our context, ship detection cannot be totally assessed, but a visual inspection still shows our network achieved good performances. As expected, we reach state-of-the-art performance for the detection task but jointly deliver relevant performance for ship classification (above 90% of correct classification) and length estimation (relative bias and standard deviation below 10%). We may point out that the considered residual architecture for length estimation seems to be a critical feature to reach good estimation performance, but it should be further investigated in order to confirm its relevance.

Further improvements will be investigated. Using false positive in the dataset would allow to evaluate the relevance of our detection network. We also consider to increase the number of classes and see if our network is robust to more complex scenarios.

REFERENCES

[1] Wentao An, Chunhua Xie, and Xinzhe Yuan. An improved iterative censoring scheme for CFAR ship detection with SAR imagery. *IEEE TGRS*, 52(8):4585–4595, 2014.

[2] François Chollet et al. Keras. https://keras.io, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. on CVPR*, pages 770–778, 2016.

[4] Miao Kang, Xiangguang Leng, Zhao Lin, and Kefeng Ji. A modified faster R-CNN based on CFAR algorithm for SAR ship detection. In *International Workshop on Remote Sensing with Intelligent Processing*, pages 1–4, 2017.

[5] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. *RSE*, 216:139–153, 2018.

[6] E Khesali, H Enayati, M Modiri, and M Mohseni Aref. Automatic ship detection in single-pol SAR images using texture features in artificial neural networks. *The International Archives of ISPRS*, 40(1):395, 2015.

[7] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pages 971–980, 2017.

[8] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[9] Mingsheng Liao, Changcheng Wang, Yong Wang, and Liming Jiang. Using SAR images to detect ships from sea clutter. *IEEE GRSL*, 5(2):194–198, 2008.

[10] Yang Liu, Miao-hui Zhang, Peng Xu, and Zheng-wei Guo. SAR ship detection using sea-land segmentation-based convolutional neural network. In *International Workshop on Remote Sensing with Intelligent Processing*, pages 1–4, 2017.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. on CVPR*, pages 3431–3440, 2015.

[12] Colin P Schwegmann, Waldo Kleynhans, Brian P Salmon, Lizwe W Mdakane, and Rory GV Meyer. Very deep learning for ship discrimination in synthetic aperture radar imagery. In *IEEE IGARSS*, pages 104–107, 2016.

[13] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. Technical report, 2012.

[14] Liangpei Zhang, Lefei Zhang, and Bo Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016.

EUCLID – AI IN THE DARK SPACE

Maurice Poncet[1], Antoine Basset[1],
Samuel Farrens[2], Alexandre Bruckert[2],
M. Gray[3], D. Vibert[3], A. Schmitt[3], Sara Jamal[3], V. Le Brun[3], O. Le Fèvre[3], C. Surace[3],
Marc Huertas-Company[4],
Hervé Dole[5], Elie Soubrié[5], Raphael Peralta[5],
Rémi Cabanac[6].

[1] CNES - French Space Agency, Toulouse, France
[2] CEA - French Alternative Energies and Atomic Energy Commission, Saclay, France
[3] Aix Marseille Univ., CNRS, LAM, Laboratoire d'Astrophysique de Marseille, Marseille, France
[4] OBSPM - Observatoire de Paris, Paris, France
[5] IAS - Institut d'Astrophysique Spatiale, Orsay, France
[6] IRAP - Institut de Recherche en Astrophysique et Planétologie, Toulouse, France

ABSTRACT

Euclid is a high-precision survey space mission developed in the frame of the Cosmic Vision Program of ESA in order to study the nature of Dark Energy and Dark Matter. Its Science Ground Segment (SGS) will have to deal with around 175 PB of data both coming from Euclid satellite data, complex pipeline processing, external ground based observations or simulations, and with an output catalog containing the description of around 10 billion objects with hundreds of attributes. Thus, the implementation of the SGS [1] is a real challenge in terms of architecture and organization. This paper focuses on the Euclid processing challenge and how Machine Learning (ML) and Deep Learning (DL) approaches may solve some key issues though some concrete examples.

Index Terms— Euclid, Astronomy, Astrophysics, Data Processing, Machine Learning, Deep Learning

1. INTRODUCTION

One of the challenges that the Euclid project has to tackle with is the unpreceded amount of data (> 100 PB) and number of objects (> 10 billions) that it will have to process, store and distribute. This has to be done through a complex multilevel processing pipeline ending with the generation of ready for science high level products. Some of the associated points have already been solved. For example, the requested processing and storage capacities will be provided by the federation of 9 Science Data Centers (SDCs) – 8 in Europe + 1 in US. This architecture will allow to deploy and run a kind of distributed map-reduce Euclid pipeline thus taking benefits of any available SDC resources in parallel, through an optimized allocation of sky areas to the SDCs. Most part of the Euclid pipeline relies on "legacy" algorithms that have to be optimized and improved in order to fit this SGS architecture and to allow to deliver the Euclid scientific products in respect to the scientific requirements and within the requested time slot.

However, some Euclid process cannot be efficiently automated by conventional algorithms and still requires human expertise and manual validation/rejection, e.g.:

- Deblending of galaxy sources,
- Galaxy structures and classification,
- Galaxy/Star distinction,
- Anomalies detections (outliers),
- Redshift assessment,
- Point Spread Function (PSF) modeling,
- Image deconvolution,
- Modified Gravity Cosmological Model Discrimination.

This is not compatible with Euclid constraints, since it would take much many years to achieve the whole process. For example, the Galaxy Zoo project (https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/), that aims at galaxy classification, involves ~15 000 volunteers who achieved 968 133 in months. We can easily infer how much time it would take to classify billions of objects in such a way!
Thus, we have to find innovative ways in order to solve this issue. A segment of the Artificial Intelligence (AI) domain – Machine Learning (ML) and Deep Learning (DL) – is a very promising approach in this case. Coming back to the previous example, some tests show that the same amount of galaxy classification could be achieved efficiently in some hours

with the ML/DL approach after the appropriate training phase.
This paper details four examples and current works in the framework of the Euclid Project involving ML/DL technologies at the French side, among many other current initiatives inside the Euclid Consortium:

- CEA - Identification of blended galaxy sources,
- LAM - Redshift reliability assessment,
- OBSMP - Morphology and structure of galaxies,
- IAS/IRAP - Stars & galaxies separation.

2. CEA- IDENTIFICATION OF BLENDED GALAXY SOURCES

Upcoming astrophysical surveys such as Canada-France Imaging Survey (CFIS) and Euclid aim to constrain cosmological parameters using properties derived from galaxy images, in particular their shapes via weak gravitational lensing. However, blending of sources (i.e. the overlap of extended objects) has a significant impact on the measurement of the morphological and structural properties of galaxies. It is therefore essential to develop effective and reliable methods for identifying blended sources in survey data and establishing appropriate means of dealing with them. This problem is even more complicated for surveys like CFIS that lack the colour information that could otherwise be used to help distinguish different galaxies from one another.

Machine learning techniques have been shown to be incredibly successful when applied to complex classification problems (see e.g. [2]). The effectiveness of these tools, however, can be difficult to gauge without reliable labelled data, which is the case for real images. Therefore, circumventing these issues requires an innovative implementation of these tools.

We have used Convolutional Neural Networks (CNNs), specifically the VGG16 network [3], to identify blended sources in CFIS simulations and have demonstrated extremely positive results (see fig 1 Blended & Non Blended Object). The network accurately identifies around 90% of blended sources in the simulations. The small percentage of failures generally correspond to cases in which it is virtually impossible to identify a given source as blended due to a complete overlap of the two galaxies.

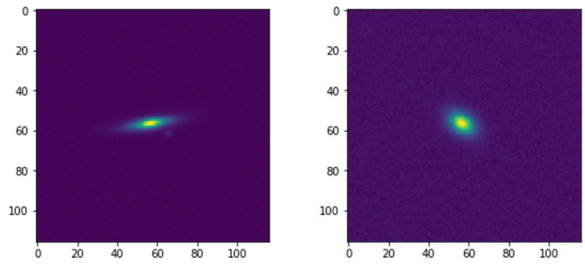


Figure 1 – Non Blended & Blending Object

This network has been trained in a way that attempts to minimize the potential overfitting of the network to the simulated data by freezing most of the initial VGG16 layers. This is done in order to make the network more robust to unseen properties in real images. Initial results on real CFIS images are very promising, where network labels correspond well with visually identifiable blends.

3. LAM - ML FOR REDSHIFT RELIABILITY ASSESSMENT

As the size of massive surveys in Astronomy continues to expand, assessing the redshifts reliability becomes increasingly challenging. The need for fully automated reliability assessment methods is now part of the requirements for future surveys as ESA Euclid mission which will provide a large set of galaxy redshifts (more than one million). It is justified by the fact that automation provides predictable and consistent performances (while the behavior of a human operator remains unpredictable), and by the need for automation in order to deal with the orders-of-magnitude increase in the total number of spectra that will be processed. We propose to automate this assessment of a spectroscopic redshift reliability flag, by exploiting key features in the redshift posterior Probability Density Function (PDF) and Machine Learning (ML) algorithms based on a preliminary step of clustering (unsupervised classification) and a second step of classical classification.

3.1. Introduction

A Bayesian framework for the spectroscopic redshift estimation, incorporating all sources of information and uncertainties of the estimation process (prior, data-model hypothesis), enables to produce a full spectroscopic redshift posterior PDF (Jamal et al. [4]). It is the starting point of our automated reliability flag definition that we describe in the following sections. Indeed, in galaxy surveys, a key issue often overlooked is the necessary evaluation of the quality of a redshift measurement because spectroscopic redshift measurement methods may be affected by a number of known or unknown observational biases that may produce some errors in the output redshift, ranging all the way to a catastrophic measurement far from the real galaxy redshift. All previous methods imply subjective information, either by selecting adequate thresholds from a constructed sample or by involving a human operator within the (visual) verification process that becomes largely unfeasible for samples over millions of galaxies. For massive spectroscopic surveys such as Euclid, there is a critical need for a fully automated reliability flag, that will adapt to the observed data and display a greater use of all available information.

Spectroscopic redshift measurements are obtained from maximization of the posterior probability p(z|D, I) in a Bayesian inference. We exploit some characteristics of the posterior PDF to build a discretized descriptor space that will

doi:10.2760/848593

be the entry point for ML techniques to predict a reliability label.
Our approach aims to build the "experience" of an automated system in order to assess the quality of a redshift measurement from the redshift PDF.

3.2. Initial clustering step

In machine learning, the typical entries of the model are a response vector Y and a feature matrix X. In the case of clustering (unsupervised classification), the response vector Y of the model is unknown. Therefore, the prediction of a label yj uses only the distribution of the feature matrix X in the features space. It will unveil the intricate structure and bring into light some properties (in the descriptor space) of the used reference dataset. We choose 8 key-descriptors of the zPDFs in order to construct the features matrix X:

- the quantity P(zMAP |D, I)
- the number of significant modes of the PDF.
- the difference in probability of the first two best redshift solutions: P(zMAP |D, I) − P(z2 |D, I)
- the dispersion $\sigma = [\int (z - \bar{z})2\, p(z)\, dz\,]1/2$, with $\bar{z} = \int z\, p(z)\, dz$.
- the cumulative probability in the region R2∗.
- the 3 characteristics of the CR∗ (restricted version of the Credibility Region with 95% in probability): number of z candidates, width Δz and cumulative probability.

By using the 8 previous listed descriptors, we expect that the main features of the zPDF can be inferred. This design is not immutable. Supplementing the feature matrix with additional information about the observed spectra, or designing a different feature selection, will also be explored.

In clustering, prior knowledge about class memberships is unavailable. Partitioning the descriptor space into K manifolds is realized by applying separation rules only to the feature matrix X. We used 3 different unsupervised classification algorithms:

- Fuzzy C-Means (FCM) with bi-partitioning (2 kinds of groups have been defined and we choose to reapply a bi-partitioning to decompose the all data in a dichotomized pattern). This partitioning strategy is applied to the entire descriptor components,
- Classical FCM (see Bezdek et al. 1984 [5]). Using this classic clustering algorithm FCM to minimize the intraclass variance, the final groups identify distinct partitions in the feature space,
- K-means.

In this study, the selection of the total number of clusters is an empirical process by testing different configurations and by evaluating the performances of the final results after the classification step using the purity-completeness curve.

3.3. Classification step

At the end of the previous step, we obtain a subset of spectra to which a reliability label is known to belong to one of the cluster that have been defined by the unsupervised clustering algorithm. In this step we use these labels as the response vector Ytrain in a supervised classification, to train a classifier to predict redshift reliability labels for new unlabeled data. The chosen classifier is the Support Vector machine (SVM) with a Gaussian kernel. To evaluate its performances two tests have been conducted: Resubstitution and Test prediction.

In the Resubstitution test, the "Training set" is reused as the "Test set" during the prediction phase. Extremely low prediction errors are expected (< 1% classification error rate): if a bijective relation exists between the observables Xtrain and the response vector Ytrain, the generated mapping from the training phase is supposedly accurate. The predicted labels Ypred in resubstitution tests are therefore expected to resemble the true labels Ytrain with high accuracy.

In the Prediction test, the "Training set" and the "Test set" are two different subsets of the initial dataset and the performances are evaluated on the Test set which has not been used during the classification step in order to define the SVM predictor.

Performances which have been observed for this SVM classifier in the resubstitution test give extremely low off-diagonal elements in the confusion matrices and an average per-class error rate less 0.1%. By having low resubstitution errors, the mapping is deemed a reliable reproduction of the input data, and the prediction of Xtest can be examined. In the Prediction test, we find that the confusion matrices for the SVM classifiers (with a Gaussian kernel) still testify of a good predictive power.

3.4. Preliminary results

To evaluate the whole method, we are using at the unsupervised clustering step a preliminary Euclid simulated dataset (COSMOS-MCDR) of 34221 spectra from mock simulations for the Euclid mission. An end-to-end simulation pipeline is currently under development using catalogs of realistic input sources with spectrophotometric information and an instrumental model for the spectrophotometer, using the pixel simulator software TIPS (Zoubian et al. [6]), 1D spectra are obtained from 2D dispersed images.
In order to evaluate the performances of the supervised classifier trained with the obtained labels we test its prediction of the reliability labels on a different preliminary dataset composed by 973 simulated spectra obtained from Euclid-SPV. For each spectrum, we estimate the redshift and compute the corresponding zPDF from which reliability labels are computed using the trained classifier described above. The obtained labels are then used to compute the

purity-completeness curve defined in the context of Euclid redshift estimation document (see Vibert et al. [7]), by rejecting sequentially the spectra belonging to the less reliable remaining class (see Fig 2).

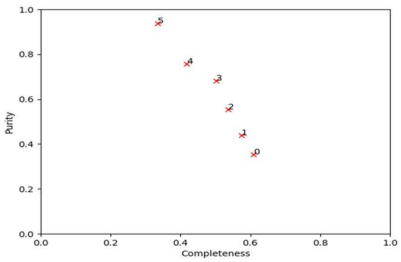


Figure 2: Purity-Completeness curve with the FC-Means algorithm & 6 clusters

A random rejection would produce a flat curve, decreasing the completeness while keeping the same purity level. We see that the classifier is already able to reject bad spectra while keeping better ones with thus a strong benefit in purity at the price of losing completeness.

This work is still in progress and many improvements can still be worked out: the descriptors space could be extended/refined, other algorithms for the clustering could be tested.

4. OBSPM - MORPHOLOGY AND STRUCTURE OF GALAXIES

Euclid will observe an area of 15.000 square degrees in the optical bands with a spatial resolution similar to the one delivered by the Hubble Space Telescope. The estimation of galaxy morphologies and structural properties for billions of galaxies spanning the last 10 billion years of the cosmic time will be one of the many legacy products of Euclid. The main challenge is to develop accurate and fast algorithms capable of dealing with the unprecedented amount of data.

4.1. Visual classifications with supervised learning & knowledge transferring

Deep learning appears as a very powerful way to overcome these issues. In two recent works (Huertas-Company+15, Dominguez-Sanchez+18) we have shown that CNNs reach an agreement close to 95% with morphological classifications of galaxies performed through visual inspection both a low redshift (SDSS) and high redshift (CANDELS). Extending this approach to EUCLID is a promising way towards providing a detailed morphological classification of billions of galaxies. One important limitation of a machine learning based approach for morphological classification is the need of a large training set. The aforementioned works benefitted indeed from significant efforts be community to visually inspect a large number of galaxies (typically tens of thousands) both by astronomers (e.g. Kartaltepe et al. 2016 [8]) and by citizens (GZOO, Lintott et al. 2011 [9]). Since different surveys have also different spatial resolutions and depths, it is not obvious that a network model trained on a given dataset can be used to classify another one. We have explored the issue of transferring knowledge between surveys in Dominguez-Sanchez et al. 2018b. We used a CNN model trained on 20.000 SDSS galaxies to reproduce the Galaxy Zoo classification from Linottt et al. 2011 [9] to classify similar galaxies observed in the framework of the Dark Energy Survey (DES). The model is a simple CNN architecture with 5 convolutional layers followed by 2 fully connected layers. We performed two basic tests. First, we used exactly the SDSS model to classify DES galaxies. The impact is that the accuracy typically drops from 95% to 85% as shown in figure 3.

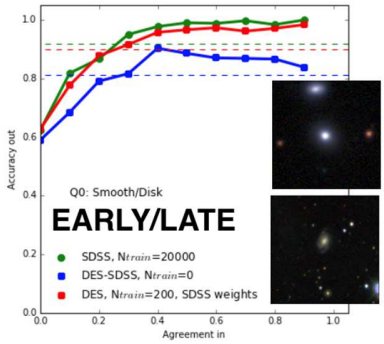


Figure 3: Knowledge transferring for morphological classification. Adapted from Dominguez-Sanchez et al. 2018b. The figure shows the accuracy obtained in reproducing the Galaxy Zoo Classification as a function of the agreement between classifiers. The green line shows the accuracy with an SDSS training of 20.000 galaxies. The blue line is the result when the same model is applied to SDSS. The red line is the accuracy obtained after a transfer learning with 200 objects.

The second exercise consisted in using a small dataset (200 galaxies) of visually classified DES galaxies to continue the training of the already trained model. We tried both retraining the full model and only the fully connected layers as usually done when transferring models. The results in our case were similar given the simplicity of the architecture in the first place. Figure 3 shows that this transferring is enough to reach again the same accuracy as with SDSS. This is an important result which suggests network models for galaxy morphology can be easily transferred between datasets without the need of visually classifying a large number of objects. Applied to Euclid, it means that one could achieve an accurate classification of billions of galaxies with a minor effort in terms of visual inspection. It is worth noticing though, that the previous result is based on two fairly similar datasets (SDSS and DES) in terms of spatial resolution and depth. The transferring between surveys with major differences still needs to be addressed.

4.2. Unsupervised morphologies with generative models

An alternative approach to morphology classification is to use unsupervised learning which would automatically classify galaxies based on their similarities. The advantage is obviously that no training set is needed. The main problem is that the interpretability of the obtained classification is not always straightforward. Generative model such as VAEs or GANs offer an interesting path to explore. Figure 4 shows a preliminary attempt to use VAEs to reconstruct H-band (F160W) images of galaxies from the CANDELS survey observed with HST.

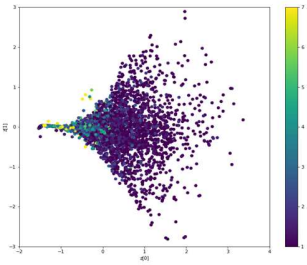


Figure 4. Two dimensional latent space representation of high redshift galaxies observed with HST in the H-band obtained with a VAE. The color code is the Sersic index of the galaxies obtained independently. High Sersic index galaxies (i.e. early-type galaxies) are preferentially located in the left part f the plane while low Sersic index galaxies (e.g. late-type) are in the right side of the diagram.

In this preliminary result, 128*128 pixel images are projected into a 2 dimensional latent space using a classical VAE. The figure shows the distribution of galaxies in the plane color coded by the Sersic index. At first approximation, the VAE is successfully splitting galaxies in groups of high Sersic index (typically early-type morphologies) and low Sersic index (typically late-type morphologies) with no pre-assumption. This could constitute an alternative approach to classify galaxies in the Euclid survey without the need of a training set. One of the problems in the current approach is that more detailed morphological structures such as clumps or other irregularities common at high redshift are not currently well captured by the VAE. A possible solution could be to increase the dimensionality of the latent space and/or to use generative adversarial models (GANs) instead which should be able to synthesize more realistic images.

4.3. Bulge-disc decompositions of galaxies with encoder-decoder networks

Another important quantity, are structural parameters of galaxies (effective radii, axis ratios, bulge-to-total fractions). The classical approach to estimate these quantities from images is based on model fitting with codes such as GALFIT or GIM2D. These codes have not been conceived however to deal with big data surveys such as Euclid. Several efforts have been made to automatize their use for catalogue compilation in large survey applications. GALAPAGOS, programmed by Barden et al. (2012) [10], combines SEXTRACTOR (Bertin & Arnouts 1996 [11]) for source detection and extraction, and then makes use of GALFIT for modeling Sersic profiles. The computing time is still very large and probably incompatible with the data volumes Euclid will handle. An alternative approach is based on neural networks. In a recent publication (Tuccillo et al. 2018 [12]), we showed that CNNs in a regression configuration can achieve similar accuracies than standard approaches in deriving the sizes and Sersic indices of galaxies but 3000 times faster. This gain in computing time makes it doable to “fit” large famous of galaxies. The limitation of the mentioned approach is that no images are delivered (as opposed to a traditional model fitting where a best model for each galaxy is produced). We are currently exploring an alternative based on U-nets (Ronnerberger et al. 2015) that aims at extracting from an image the bulge and disc components in two separate images. The results are still very preliminary (Fig. 5) but seem to indicate that the U-net is able to split light between the different components without the need of model fitting.

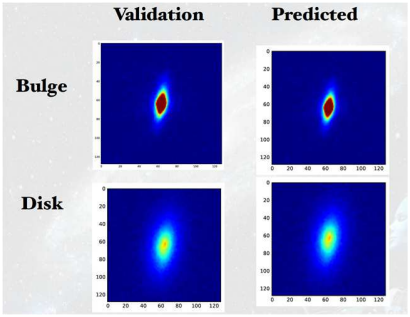


Figure 5. Preliminary example of the output of a U-net to perform bulge/disc decomposition of galaxies. The left column shows the input simulated bulge and disc components and the right column shows the predictions of the U-net for the same object.

5. IAS/IRAP – STARS & GALAXIES SEPARATION

Cosmology space missions require an extreme control on systematics, among which a proper separation between resolved objects such as galaxies (used to measure estimators relevant for cosmology) and unresolved objects such as stars (used to measure the properties of the optical system). Machine Learning can perform in principle a very efficient classification based on images. A major difficulty, however, is the building of a reliable training set based on observed data (not simulations).

Using ground-based data and an already star and galaxy classification obtained with a widely used method (such as SExtractor software, Bertin & Arnouts [11]) from public

surveys, namely the CFHTLS (McCracken et al., 2008 [14]; Ilbert et al., 2006 [15]) and the DES (Dark Energy Survey Collaboration, 2016 [16]), we trained a CNN with a few layers, as described in [17]. We reach an accuracy comparable with the widely used method.

Our difficulty is currently to build a reliable training set. We plan to use simulated data, but real data closer to the Euclid expected properties are thought to be a better direction in order to be more realistic and reliable. We will review our progress and difficulties.

6. CONCLUSIONS

Machine Learning and Deep Learning technologies are very promising approaches in the framework of the Euclid project. Euclid teams are already exploring these solutions in many fields of the Euclid project where conventional approaches are not relevant. The first results are conclusive and the next years should confirm Euclid ML/DL adoption.

References

[1] M. Poncet & al, Euclid Science Ground Segment (SGS) Processing Operations Concept, SpaceOps, Marseille, 2018

[2] Kotsiantis, Sotiris, Supervised Machine Learning: A Review of Classification Techniques, Informatica (Slovenia) volume 31 p249-268, 2007

[3] K. Simonyan & A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, ICLR, 2015

[4] S. Jamal et al., Automated reliability assessment for spectroscopic redshift measurements, A&A 611, A53 (2018).

[5] Bezdek,et al (1984), FCM: The fuzzy c-means clustering algorithm, Computers & Geosciences, 10(2-3), 191-203.

[6] Zoubian, J., Kümmel, M., Kermiche, S., et al. (2014), in ASP Conf. Ser., 485, 509.

[7] D. Vibert et al., Euclid statistics: Purity and Completeness estimation, (internal document) EUCL-LAM-8-0001, April 2018.

[8] Kartaltepe et al. 2016, Candels visual classification: scheme, data release, and first results, The Astrophysical Journal Supplement Series, 221:11(17pp), 2016 November.

[9] Lintott et al, Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies'. Monthly Notices of the Royal Astronomical Society 410 1, pp. 166–178, 2011.

[10] Barden et al, GALAPAGOS: from pixels to parameters, Monthly Notices of the Royal Astronomical Society, Volume 422, Issue 1, pp. 449-468, 2012.

[11] Bertin & Arnouts, SExtractor: Software for source extraction, Astronomy and Astrophysics Supplement, v.117, p.393-404, 1996.

[12] Tuccillo & al, Deep learning for galaxy surface brightness profile fitting, 2018.

[13] Ronnerberger et al., U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015.

[14] McCracken et al., Clustering properties of a type-selected volume-limited sample of galaxies in the CFHTLS, A&A, 479, 321, 2008.

[15] Ilbert et al., Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey, A&A, 457, 851, 2006.

[16] Dark Energy Survey Collaboration, The Dark Energy Survey: more than dark energy – an overview, MNRAS, 460, 1270, 2016.

[17] Kim, Edward J.; Brunner, Robert J. ; Star-galaxy Classification Using Deep Convolutional Neural Networks ; MNRAS, Volume 464, Issue 4, p.4463-4475.

A NEW LARGE-SCALE SENTINEL-2 BENCHMARK ARCHIVE AND A THREE-BRANCH CNN FOR CLASSIFICATION OF SENTINEL-2 IMAGES

Gencer Sumbul [1]*, Begüm Demir* [1]*, Volker Markl* [1,2]

[1]Technische Universität Berlin
[2]DFKI GmbH
e-mail: gencer.suembuel@tu-berlin.de,
demir@tu-berlin.de, volker.markl@tu-berlin.de

ABSTRACT

This paper presents a new large-scale Sentinel-2 benchmark archive together with a three-branch convolutional deep neural network (TB-CNN) designed for accurate characterization of Sentinel-2 images. Our archive, named BigEarthNet, consists of 590,326 Sentinel-2 image patches that are accompanied by multiple land-cover annotations. The BigEarthNet is 20 times larger than existing archives in remote sensing and thus is much more convenient to be used as a training source in the framework of deep learning. Our TB-CNN with its specifically designed three branch convolutional neural network architecture considers all bands of Sentinel-2 images, having different spatial resolutions in order to extract complementary representation from different band combinations. Both the BigEarthNet and the TB-CNN open up promising directions to advance the research on image classification in massive Sentinel-2 image archives.

Index Terms— Sentinel-2 image archive, classification, convolutional neural network, remote sensing

1. INTRODUCTION

The huge number of recent Earth observation satellite missions has led to a significant growth of remote sensing (RS) image archives. One of the most challenging and emerging applications in RS is related to efficient and effective classification of RS image scenes present in such archives. The performance of classification systems highly depends on the capability of the RS image features in characterizing the content of the images. Recent advances in deep learning have attracted great attention in RS due to high capability of deep networks (e.g., Convolutional Neural Network, Recurrent Neural Networks, Generative Adversarial Networks) to model high-level semantic content of RS images. To train such networks, very large training sets are needed with a high number of annotated images in order to learn effective models with several different parameters.

To date, publicly available RS image archives contain only a small number of annotated images and a large-scale annotated archive does not yet exists. As an example, the recently published EuroSAT archive [1] is composed of only 27,000 annotated Sentinel-2 images from which accurately learning the large number of parameters in deep learning models is not feasible, as the models may overfit dramatically when using small training sets. Thus, the lack of large training sets is an important bottleneck that prevents the use of deep learning in RS. In order to address this problem, fine-tuning deep networks pre-trained on large-scale computer vision archives (e.g., ImageNet) are considered. However, such an approach has several limitations considering differences on the characteristics of images in computer-vision and RS. As an example, Sentinel-2 images are different with respect to the high resolution RGB images in computer vision on which modern deep networks trained on. Sentinel-2 images have not only lower spatial resolutions with respect to those images but also varying resolutions for different spectral bands. Furthermore, using pre-trained models generally prevents to use more than three channels, whereas accurate characterization of the other spectral bands having different spatial resolutions can further increase the performance of deep networks if used in a convenient way.

To overcome all these problems, we introduce a new large-scale Sentinel-2 archive, named BigEarthNet, that contains 590,326 Sentinel-2 image patches with multiple annotations. Moreover, we design a deep neural network that includes three different convolutional branches, which are specifically designed for different spatial resolutions to benefit from all Sentinel-2 bands having 10m, 20m and 60m spatial resolutions. Our network also contains feature-level fusion of different representations extracted for different spatial resolutions in order to map fused image patch embedding into multi-label clases. It is worth noting that the BigEarthNet and pre-trained model weights of TB-CNN will be made publicly available and we believe that it will make a significant advancement in terms of developments of algorithms for the analysis of large-scale RS image archives. Our contributions for this paper can be summarized as follows: 1) we introduce a large-scale annotated Sentinel-2 benchmark archive and 2)

doi:10.2760/848593

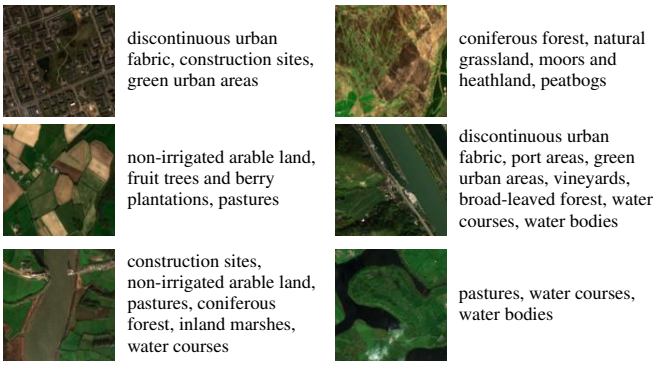


Fig. 1: Example Sentinel-2 images and their multi-labels in our BigEarthNet archive.

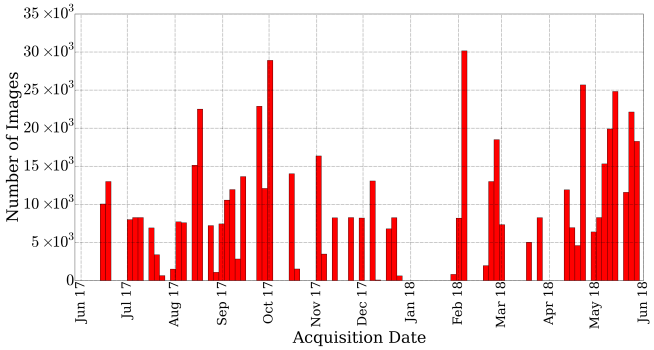


Fig. 2: The number of Sentinel-2 images with respect to acquisition date.

we design a deep network particularly designed for modelling Sentinel-2 images with 13 bands.

2. THE BIGEARTHNET ARCHIVE

The BigEarthNet has been constructed by selecting 125 Sentinel-2 tiles distributed over 10 European countries (Austria, Belgium, Finland, Ireland, Kosovo, Lithuania, Luxembourg, Portugal, Serbia, Switzerland) and acquired between June 2017 and May 2018. The reason of selecting tiles from these countries and acquisition dates is that European Environment Agency has very recently received the updated land-cover products of these countries, for which annotation process has been carried out for the period of 2017-2018. In other words, CORINE land cover products has been recently updated as CORINE Land Cover (CLC) 2018 for these countries. In details, all the considered image tiles are associated to cloud cover percentage less than 1% and they were atmospherically corrected by the use of Sentinel-2 Level 2A product generation and formatting tool of ESA [2]. All spectral bands except for the 10th band, in which surface information is not embodied, were included. The tiles were divided into non-overlapping image patches with the size of 120x120 for 10 meter bands, 60x60 for 20 meter bands and 20x20 for 60 meter bands. Each image patch (denoted as image hereafter) in the archive has been annotated with one or more labels based on 43 classes of CLC 2018 and then visual inspection has been also carried out in order to apply a quality check. The considered class labels are: continuous urban fabric, discontinuous urban fabric, industrial or commercial units, road and rail networks and associated land, port areas, airports, mineral extraction sites, dump sites, construction sites, green urban areas, sport and leisure facilities, non-irrigated arable land, permanently irrigated land, rice fields, vineyards, fruit trees and berry plantations, olive groves, pastures, annual crops associated with permanent crops, complex cultivation patterns, land principally occupied by agriculture, with significant areas of natural vegetation, agro-forestry areas, broad-leaved forest, coniferous forest, mixed forest, natural grassland, moors and heathland, sclerophyllous vegetation, transitional woodland/shrub, beaches, dunes, sands, bare rock, sparsely vegetated areas, burnt areas, inland marshes, peatbogs, salt marshes, salines, intertidal flats, water courses, water bodies, coastal lagoons, estuaries and sea and ocean.

Table 1: Seasonal distribution of Sentinel-2 images

Seasons	Autumn	Winter	Spring	Summer
Number of Images	154943	117156	189276	128951

Fig. 1 shows examples of images and the multi-labels associated with them, while Fig. 2 shows the number of Sentinel-2 images with respect to acquisition date. The number of labels associated with each image varies between 1 and 12, whereas 95% of images have at most 5 multi-labels. It is worth noting that we aimed to represent each considered geographic location with images acquired in all different seasons. However, due to the difficulties of collecting Sentinel-2 images with lower cloud cover percentage within a narrow time interval, for some areas it was not possible. The number of images for each season is listed in Table 1.

We would like to note that in the existing archives (e.g., EuroSAT) each training image is annotated by a single label associated to the most signicant semantic content of the image. However, this assumption does not fit well with the complexity of Sentinel-2 images, where an image might have multiple land-cover classes (i.e., multi-labels). The BigEarthNet is very promising since according to our knowledge it is the first large-scale Sentinel-2 benchmark archive that includes multi-labels. We would like to also note that the BigEarthNet will be publicly available at http://bigearth.net.

3. OUR THREE BRANCH DEEP CONVOLUTIONAL NEURAL NETWORK

In this section, we introduce our approach for classification of Sentinel-2 images considering all bands with different spatial resolutions. Let $\mathcal{X} = \{x_1, x_2, ..., x_M\}$ be an archive that consist of M images where x_i is the i-th image (for the BigEarthNet, $M = 590,326$). We aim to find a mapping of x_i into

the one or more class labels $\{y_1, .., y_C\}$ while $y_j \in \mathcal{Y}$ is the set of all BigEarthNet class labels and C is the number of possible labels. To this end, we introduce a new deep neural network architecture developed considering the specifications of Sentinel-2 images and learnt end-to-end on BigEarthNet archive. In our approach, there are three different convolutional branches for each 10m, 20m and 60m resolution of Sentinel-2 images. Each branch acts as a feature extractor for different resolutions. To this end, unlike the existing deep networks for Sentinel-2 images that generate a single representation of x_i with only RGB channels, we aim to extract different image representations for each spatial resolution. The motivation behind using three branch convolutional neural network to use different CNNs specialized for the different spatial resolution. Although all branches have the same regime for the number of filters for each convolutional layer (starting with 32 filters, increasing with multiplication by 2 while going deeper into model and decreasing again with division by 2 and ending with 64 filters), filter size and applied operations between convolutional layers are different. We have decided to select decreasing filter sizes among different branches for capturing the textural regions having similar size while keeping them large enough. Thus, 5x5 filters for initial layers and 3x3 filters for deeper layers are used for the first branch, which accepts 10m resolution bands. For the second and third branches developed for 20m and 60m resolutions, 3x3 filters and 2x2 filters are used respectively throughout the layers. For all convolutional layers, stride of 1 and zero padding are used to prevent information deficiency and to retain the spatial dimensionality. Additionally, we apply max-pooling to have partial translation invariance for the first two branches. However, in order not to decrease spatial resolution more, it is not preferred for the last branch. For each branch, there is a fully connected (FC) layer that takes the output of the last convolutional layer and produces feature vector. Let ϕ^r be one of the branches that takes all image bands, x_{i_r}, having same spatial resolution and generates image representation for the spatial resolution r where $r \in \{10m, 20m, 60m\}$ is the set of all spatial resolutions for Sentinel-2 images. After obtaining different image features, all of them are concatenated into one vector, $\phi_{concat}(x_i)$, as follows:

$$\phi_{concat}(x_i) = [\phi^{10m}(x_{i_{10m}})^\top, \phi^{20m}(x_{i_{20m}})^\top, \phi^{60m}(x_{i_{60m}})^\top]^\top \quad (1)$$

where $\phi^{10m}, \phi^{20m}, \phi^{60m}$ are the result of different branches for different spatial resolutions while $x_{i_{10m}}, x_{i_{20m}}, x_{i_{60m}}$ are the different spectral band subsets in which spatial resolution is the same. Although all image features are represented as the single vector by cascading all vectors, using all information complementarily as the single feature vector for classification is provided with a new FC layer that takes concatenated vector and produces fused image representation as the final feature vector, $\phi_{fused}(x_i)$. By appyling feature-level fusion with FC layer, our model is more capable of extracting complementary visual features for contextual and spectral information. To this end, fusion is applied to cascaded vectors as follows in our approach:

$$\phi_{fused}(x_i) = W_{fusion}\phi_{concat}(x_i) \quad (2)$$

where W_{fusion} is the model weights of the fusion FC layer. After obtaining final image representation, last FC layer generates class scores, z_{y_j}, with respect to this fused feature vector for each class label y_j where $\forall j \in 1, ..., 43$ is one of the BigEarthNet class labels. Finally, class posterior probability of y_j for the image x_i is written as the *sigmoid* of its class score as follows: $P(y_j|x_i) = {}^1/_{(1+e^{-z_{y_j}})}$. After obtaining class labels, we define overall model loss as the cross entropy loss throughout all class labels and images as follows:

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{i=1}^{C} l_{ij} log(P(y_j|x_i)) + (1 - l_{ij}) log(1 - P(y_j|x_i)) \quad (3)$$

where l_{ij} takes 1 if y_j is the true class label of x_i, 0 otherwise. We should emphasize here that by using this loss function, which maximizes the predicted true class probabilities of multi-labels for all training examples, our model is capable of classifying images into multi-labels. Our approach is illustrated in Fig. 3.

4. EXPERIMENTAL RESULTS

To train our model for the multi-label classification of BigEarthNet images, we first randomly selected 10^5 training and 3×10^4 validation patches and shuffle them in order to prevent biases occured in archive preparation process. For each image, we splitted the bands into three subsets and in each subset we stacked bands into a single volume. First, second and third CNN branches accept bands 4, 3, 2 and 8 having 10m spatial resolution and 120×120 image pixel size, bands 5, 6, 7, 8A, 11 and 12 having 20m spatial resolution and 60×60 image pixel size, bands 1 and 9 having 60m spatial resolution and 20×20 image pixel size.

End-to-end training of all branches together is employed and thus we learn visual representations of spectral bands having different resolutions simultaneously. We train the whole model using Adam [3] method of Stochastic Gradient Descent in order to decrease the sigmoid cross entropy loss, which provides maximizing the log-likelihood of each label from the set of BigEarthNet classes throughout all training images. To this end, all model parameters were initialized with Xavier method [4] and initial learning rate of Adam is selected as 10^{-3}. The mini-batch size and the L2-regularization weight for layer-wise regularization of convolutional and fully connected layers are decided as 100, and 2×10^{-5} respectively with respect to initial tests. In addition to those, Batch Normalization [5] in order not to be affected from different spectral band statistics and Dropout regularization [6] with 20% dropping out probability in order to prevent overfitting over

doi:10.2760/848593

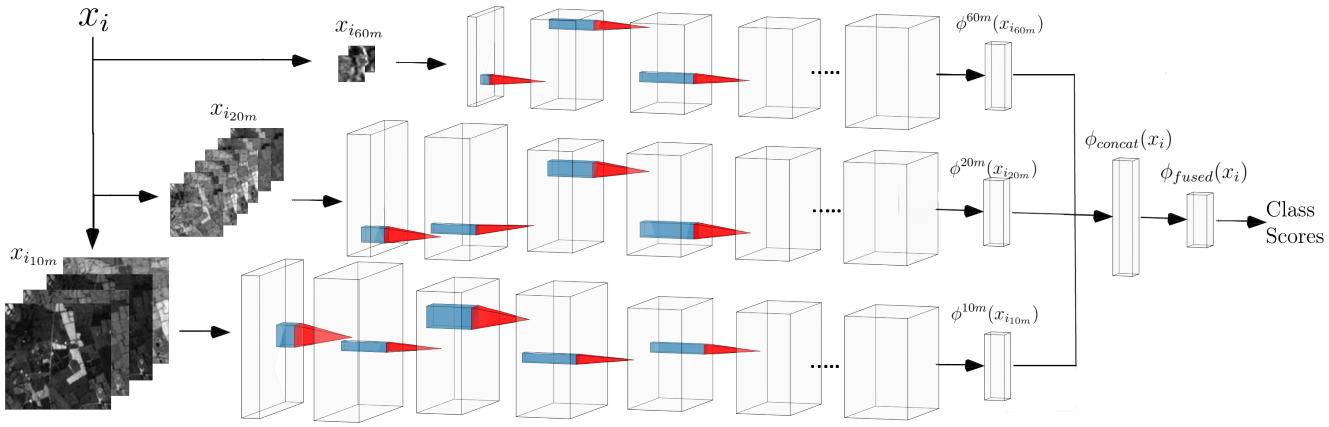


Fig. 3: Our TB-CNN in which each branch is developed for the different resolution bands of Sentinel-2 images.

training data were used in order to improve the training. For our experiments, we use average recall as the performance metric.

In order to evaluate our TB-CNN, we compare the results of our approach with single branch CNN that only uses RGB image channels. To use only RGB images, CNN branches for 20m and 60m spatial resolutions, cascading and fusing layer were excluded from our model that were retrained again from scratch within the same conditions. As we can see from Table 2, our TB-CNN provides 17.5% higher recall than single branch CNN. Accordingly, the proposed TB-CNN is much more suitable to be used on real Sentinel-2 image classification scenarios with respect to the single branch CNN that considers RGB bands only. We note that the promosing performance of our TB-CNN relies on: i) the high quantity of annotated training images; and also ii) the use of tree CNN branches that complementarily consider all the Sentinel-2 bands.

Table 2: Classification results for BigEarthNet images

Evaluation Metric	Single Branch CNN	**Our TB-CNN**
Average Recall	50.0%	**67.5%**

5. DISCUSSION AND CONCLUSION

This paper represents a large benchmark archive that consists of 590,326 Sentinel-2 image patches annotated by multi-labels for training and evaluating RS image classification algorithms. In addition, we introduce a deep neural network for the classification of Sentinel-2 images, which includes both three convolutional branch for each spatial resolution and fusion of different image representations obtained by these branches. Experimental results show the effectiveness of the TB-CNN trained on Sentinel-2 annotated images in our BigEartNet archive.

We would like to note that we plan to regularly enrich the BigEarthNet by increasing the number of annotated Sentinel-2 images. Our current system does not include any scalable architecture for the management of the BigEarthNet and also for the training of the TB-CNN model. This could create a limitation when BigEarthNet grows. However, we are currently working on designing and implementing a scalable architecture for massive processing and analysis of images in the BigEarthNet, which can be very beneficial for the complete analysis of Sentinel-2 archives with petabytes of images. In details, we give special emphasize on developing an architecture for: i) an efficient management of the BigEarthNet; ii) scheduling of large number of data-depending tasks; and iii) an efficient implementation of TC-CNN on hierarchical cluster-based parallel systems.

6. ACKNOWLEDGEMENTS

This work was supported by the European Research Council under the ERC Starting Grant BigEarth-759764. The authors would like to thank Dr. Marcela Charfuelan for many valuable discussions and also her support to prepare the web page of the BigEarthNet with scalable image search options.

REFERENCES

[1] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *arXiv preprint arXiv:1709.00029*, 2017.

[2] U. Muller-Wilm, J. Louis, R. Richter, F. Gascon, and M. Niezette, "Sentinel-2 level 2a prototype processor: Architecture, algorithms and first results," in *ESA Living Planet Symp.*, 2013, p. 98.

[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Intl. Conf. Learn. Represent.*, 2014, pp. 1–41.

[4] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Intl. Conf. Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[5] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Intl. Conf. Mach. Learn.*, 2015, pp. 448–456.

[6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

DATA SCIENCE WORKFLOWS FOR THE CANDELA PROJECT

Mihai Datcu[1], Corneliu Octavian Dumitru[1], Gottfried Schwarz[1], Fabien Castel[2], and Jose Lorenzo[3]

[1]Remote Sensing Technology Institute, German Aerospace Center, Wessling 82234, Germany (email: corneliu.dumitru@dlr.de; gottfried.schwarz@dlr.de; mihai.datcu@dlr.de)

[2]ATOS France SA, Les Espaces St Martin, 6 Impasse Alice Guy, 31024 Toulouse, France (email: fabien.castel@atos.net)

[3]ATOS SPAIN SA, Calle Albarracín 25, 28037 Madrid, Spain (email: jose.lorenzo@atos.net)

ABSTRACT

This paper describes CANDELA - Copernicus Access Platform Intermediate Layers Small Scale Demonstrator - a general platform for the handling, analysis, and interpretation of Earth observation satellite images, mainly exploiting big data of the European Copernicus Programme. Its workflow allows the selection of satellite images, the generation of local image patch descriptors, the ingestion of image and descriptor data into a common database, the assignment of semantic content labels to image patches, and the search and retrieval of similar content-related image patches.

***Index Terms*—**Data science, Earth observation, Copernicus data, data mining, data fusion.

1. INTRODUCTION

With the advent of the Copernicus Programme with its wealth of open data, the Earth observation (EO) application and service development domain is increasingly adopting big data technologies. This adoption is first related to efficient data storage and processing infrastructures, but most importantly to data analytics and application development concepts. Efficient data retrieval, mining augmented with machine learning techniques, and interoperability are key in order to fully benefit from the available assets, create more value and subsequently economic growth and development of the European member states [1-2].

CANDELA - Copernicus Access Platform Intermediate Layers Small Scale Demonstrator - aims at building a platform that delivers building blocks and services which enable users to quickly use, manipulate, explore and process Copernicus data. The main objective of CANDELA is to bridge the gap between big data technology and the EO data user community. While the objective is very ambitious, the pragmatic approach that we follow when building CANDELA makes it reachable. With the right blend between solid, operational existing assets and innovative tool integration, the platform shall help current and future Copernicus users to take a leap and profit from big data technology to maximize value creation.

In addition to an existing set of tools that our consortium already implemented for the platform [2], CANDELA will enable users to integrate already existing building blocks with a homogeneous, powerful and operational platform, opening up collaboration possibilities to research new approaches and offerings. This approach is highlighted by the development of scenarios such as urbanization, vineyard development, or forest disaster monitoring that will not only contribute as validation scenarios, but that constitutes real commercial, operational scenarios with existing customers.

The goal of one of CANDELA's work packages (WPs) is to develop the data science workflows and data analysis tools needed for implementing the functionality needed for the practical use cases of CANDELA. Each of the data science workflows will require configuration of data and optimization of the overall processes which will be performed through this task.

In this paper, we report about the data science workflows of the CANDELA project. The measured performance of the platform/system is beyond the scope of this paper.

2. DATA SET DESCRIPTION

Our main data sets extracted from different instruments are Earth surface images of the European Copernicus Programme, namely Sentinel-1 and Sentinel-2 images. While Sentinel-1 is an active twin satellite synthetic aperture radar (SAR) system, each of the Sentinel-2 twin satellites carries a passive optical multi-color imager. All instruments have been designed, calibrated, and are being operated by the European Space Agency (ESA) [3-4].

There are many reasons why we advocate the use of Sentinel-1 and Sentinel-2 images.

Firstly, we can recognize different target area details in overlapping radar and optical images complementing each other with rapid succession.

Secondly, individually selectable Sentinel-1 and Sentinel-2 images can be rectified and co-aligned by publicly available toolbox routines offered by ESA allowing a straightforward image comparison.

doi:10.2760/848593

Thirdly, all Sentinel instruments are well-documented, and typical data sets are already well understood within the remote sensing community. Many publications describe newly discovered Earth surface characteristics derived from the individual instruments.

Furthermore, the long-term operations of the Sentinel satellites allow the interpretation of image time series, or even the combination of time series data with external supplementary data via additional data mining/data fusion tools [5-7].

Besides these data sets, we include other 3rd party EO data sets as specified by CANDELA users (*e.g.*, TerraSAR-X, WordView, and Landsat).

3. DATA MINING AND DATA FUSION COMPONENTS OF THE CANDELA PLATFORM

The *CANDELA - Copernicus Access Platform Intermediate Layers Small Scale Demonstrator* - platform allows the prototyping of EO applications by applying interactive data mining and knowledge discovery functions to satellite images in order to select appropriate products in large data archives. It also helps to detect objects or structures, and to classify their land cover categories. The system allows ingesting Synthetic Aperture Radar (SAR) and multispectral images (*e.g.*, Sentinel-1, Sentinel-2, WorldView-2, TerraSAR-X). For other types of data, the main requirement is that the input data are provided in GeoTIFF format.

The main components of the CANDELA system are depicted in Figure 1 and are the following ones: Data Model Generation (DMG), DataBase Management System (DBMS), Image Search and Semantic Annotation, and Multi-Knowledge and Querying.

a) Data Model Generation

The **data mining** image ingestion can be seen as a processing chain, which, in our case, is managed by the Data Model Generation (DMG). This component is responsible for extracting the basic primitive features from the EO images (*e.g.*, SAR or multispectral images), generating tiles and their corresponding high-resolution visual quick-looks, and storing all the generated information into a database. The information is stored into an XML file. Finally, this file is automatically transformed into SQL statements and inserted into the DBMS.

It's important to mention here that for the DMG module, the input EO images should be GeoTIFF files (*e.g.*, Sentinel-1, TerraSAR-X), and their associated metadata XML files, widely used remote sensing data standards. An exception are the Sentinel-2 images, which are composed of several quadrants; here it is necessary to specify the folder of the quadrant-image that shall be processed.

Before starting the ingestion, it is necessary to verify that a freely available and thus popular MonetDB database installation is running. Once this has been verified, the next step is to select the **input images** and the **output location,** together with the size of the image patches and the **number of grid levels** (*e.g.*, 1, 2 or 3). The relation between the size of the image patches and their grid level is that, in case of grid level 1, an image is divided into patches with the specified size. If the grid level is 2, the same patch, from the previous step, is further divided into four patches with half of the previous size. Therefore, the number of patches of grid level 2 will by four times the number of patches of grid level 1. This procedure is repeated for grid level 3, etc.

In addition, for each generated image patch a visual quick-look file is created in JPEG format. In the case of SAR images, their brightness is adjusted to create this JPEG file. In the case of multispectral images, the RGB bands are used to create the JPEG quick-look file.

In case of Sentinel-2, since the data come in JPEG2000 format, an intermediary step is needed in order to convert the JPEG2000 format to GeoTIFF format necessary for the DMG input.

Within the ingestion, the **sensor type** can be chosen by the user from the following sensor types that correspond to the type of images one would like to ingest: **TSX** for TerraSAR-X data, **S1A** for Sentinel-1A/1B data, **S2A** for Sentinel-2A/2B data, or **OPT** for WorldView-2 data or other multispectral data (all in GeoTIFF format).

Depending on the type of input data and the envisaged application, different **feature extraction methods** can be applied. In the current version we implemented: **Gabor filters**, **Weber local descriptors**, and **histograms** [8]. The feature extraction methods are classified and used according to the input data. As for example, for **SAR images** one can use **Gabor filters** with the two options **Gabor linear moments** and **Gabor logarithmic cumulants** or **adaptive Weber local descriptors**, while for **multispectral images** one can use either **Weber local descriptors** or **multispectral histograms**. The size of the feature vector depends on the combination of the selected parameters (*e.g.*, for Gabor linear moments the mean and variance of 4 scales and 5 orientations gives us a feature vector of 2×4×5=40 parameters). Typically, data ingestion and patch tiling take together 1.5 ms per patch of 256×256 pixels, while feature extraction requires 2 ms per patch (for 40 parameters).

Inside the DMG module, there are a number of components that have been developed for **data fusion**. These fusion components are **data fusion ingestion**, **data fusion feature generation,** and **data fusion high-resolution quick-look**.

For each EO product within the data fusion component, there are a number of **features/descriptors** which can be selected by the user. For multispectral products (*e.g.*, Sentinel-2, WorldView) there are three feature algorithms being implemented, namely **multispectral histograms**, **Weber local descriptors**, and **Gabor linear moments** (computed for each band and concatenating the results). For SAR products, the same number of feature algorithms are implemented, namely **Gabor linear moments**, **Gabor log-cumulants**, and **adaptive Weber local descriptors**.

doi:10.2760/848593

Based on the available features, the user can select one type of feature per sensor (one for multispectral, and one for SAR data) and after that, the features are fused and normalized before being inserted into the DataBase Management System (DBMS).

b) DataBase Management System

Here in the DBMS, the inputs from data ingestion and from the semantic annotation are stored into a relational database structure and act as the core of the system interacting with all components and supporting their functionality. The database is used also for different types of queries.

c) Image Search and Semantic Annotation

This module is used to search for image content and to create semantic annotations of the ingested images. Our implementation is based on the Cascaded Active Learning for Object Retrieval (CALOR) algorithm [9], which contains a Support Vector Machine (SVM) as our active learning and relevance-feedback method in order to allow the inclusion of human expertise in the annotation.

The definition of image semantics is achieved using an interactive loop where human expertise is required to define the most appropriate semantic category and to terminate the loop [10]. The employed CALOR algorithm is based on active learning methods. The idea behind active learning is that a machine learning algorithm can achieve higher accuracy with fewer training labels if it's allowed to choose the data from which it learns.

Another important component is **data fusion**, the generation of high-resolution quick-look data which includes the DMG from each sensor as well as the corresponding high-resolution quick-look image, and generates a single fused quick-look image needed for the image search and semantic annotation module. This module has been developed for semantic annotation of the image content by using machine learning algorithms and human interaction; another benefit of this module is that it can be used for fusion of different semantic labels. These fused semantic labels are saved into the database management system from where the user can run a query using the Multi-Knowledge and Querying module.

d) Multi-Knowledge and Querying

This module reads the required information from all tables in the database.

The query module is an interactive component, which allows the user to better exploit the EO products (*e.g.*, image and metadata). Based on these two types of data (image content and metadata), there are two different queries being available: query by metadata, and query by semantics. These queries can be also combined.

The first type of query is exploiting the entire metadata of each EO product by extracting and storing the information into the DBMS. Depending on the user needs, different metadata parameters can be combined during querying.

The second querying option is a query by semantics. In this case, the user can select one or several labels from a given list in order to perform the query. Please note that the semantic labels are generated *a priori* via the Image Search and Semantic Annotation module. Using this module the quality of the semantic annotation can be verified after the interactive part has been finished. To query new data (Sentinel-1 /-2), these data need to be annotated before.

4. DATA SCIENCE WORKFLOWS

In this section, we explain the functionality of the **data mining** and **data fusion** modules. First, we start by presenting in Figure 2 and 3 the data mining capabilities for two functions, namely data mining exploration and data mining semantic annotation. Similar workflows can be created for data fusion. Our example refers to a single sensor and can be extended to multi-sensor configurations.

After the first stage has been completed (see Figure 2) and the user has a first idea of the content of the data set, now he/she can go ahead and can assign semantic labels to each retrieved category (see Figure 3).

Based on the output of these schemes (Figures 2 and 3), two possible scenarios can be imagined in close connection with known CANDELA use cases. The first one is a data mining query together with data analytics, and the second one is semantic sensor fusion.

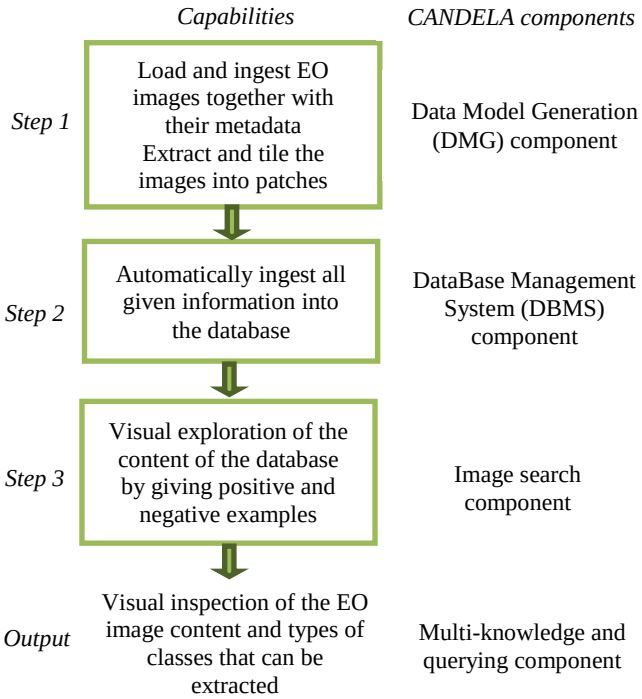


Figure 2. Data mining exploration.

5. DISCUSSIONS

After the final testing of our system has been completed, we will select a dataset for which the output is known and we will try to find similar systems, if such systems exist, to compare the results.

doi:10.2760/848593

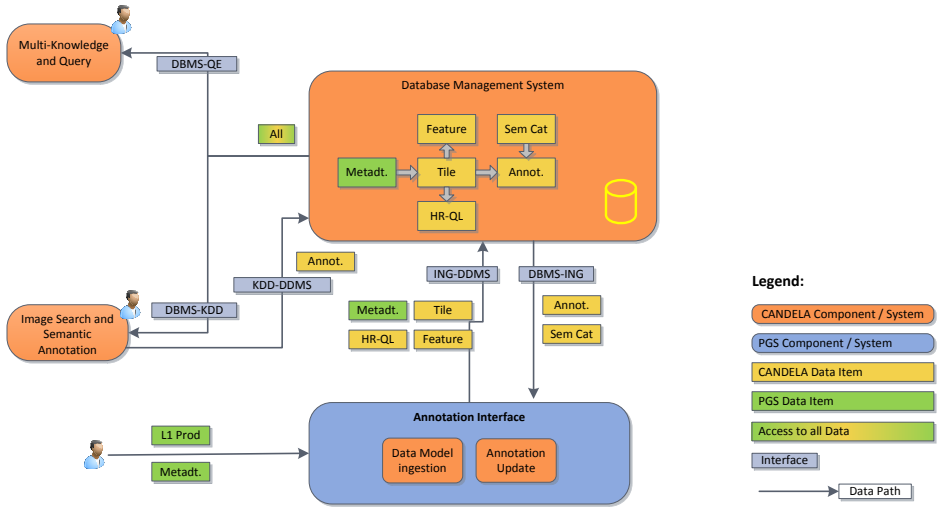


Figure 1. Components of the CANDELA - Copernicus Access Platform Intermediate Layers Small Scale Demonstrator - platform.

In addition, we are very much interested in comparable approaches implemented by other institutions.

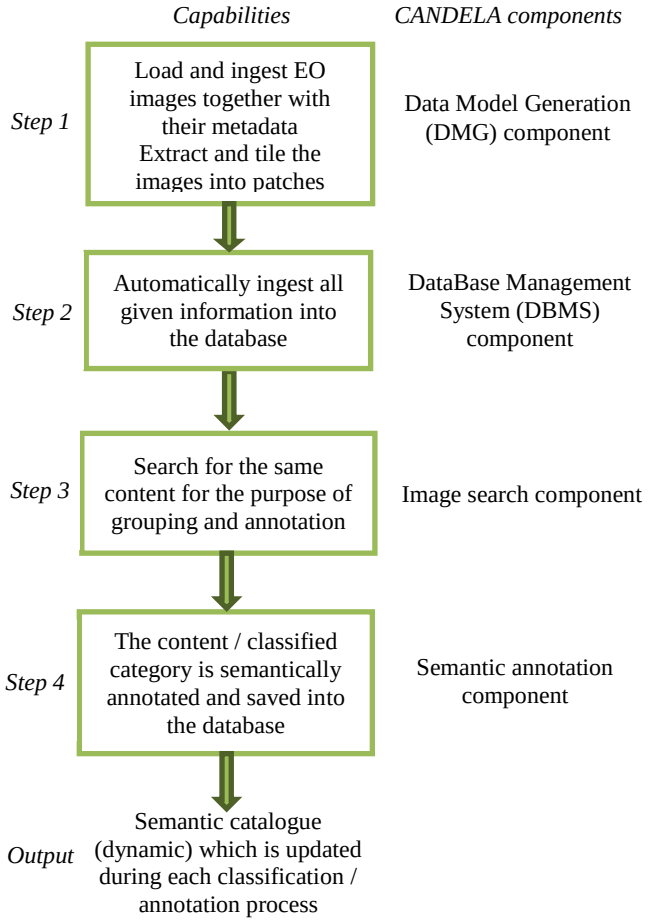


Figure 3. Data mining semantic annotation.

6. ACKNOWLEDGEMENTS

This work was supported by the CANDELA - Copernicus Access Platform Intermediate Layers Small Scale Demonstrator H2020 research and innovation project under grant agreement No. 776193.

7. REFERENCES

[1] CANDELA: Copernicus Access Platform Intermediate Layers Small Scale Demonstrator H2020 proposal, 2017.
[2] CANDELA project, 2018. [Online]. Available: http://www.candela-h2020.eu/.
[3] ESA Sentinel-1, 2018. [Online]. Available: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1.
[4] ESA Sentinel-2, 2018. [Online]. Available: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2.
[5] Living Planet Symposium, 2016. [Online]. Available: http://lps16.esa.int/.
[6] Big Data from Space Conference, 2017. [Online]. Available: https://earth.esa.int/web/guest/events/all-events/-/article/conference-on-big-data-from-space-bids-17.
[7] IGARSS, 2018. [Online]. Available: https://www.igarss2018.org/Tutorials.asp#FD-6.
[8] D. Espinoza-Molina, V. Manilici, S. Cui, Ch. Reck, M. Hofmann, C.O. Dumitru, G. Schwarz, H. Rotzoll, and M. Datcu, *"Data Mining and Knowledge Discovery for the TerraSAR-X Payload Ground Segment"*, PV 2015, Darmstadt, Germany, 2015.
[9] P. Blanchart, M. Ferecatu, S. Cui and M. Datcu, *"Pattern retrieval in large image databases using multiscale coarse-to-fine cascaded active learning,"*Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 4, pp. 1127-1141, 2014.
[10] C. Dumitru, G. Schwarz and M. Datcu, *"SAR Land Cover Datasets for Benchmarking,"* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 5, pp. 1571-1592, 2018.

HERE IS MY QUERY, WHERE ARE MY RESULTS? A SEARCH LOG ANALYSIS OF THE EOWEB® GEOPORTAL

Sirko Schindler, Marcus Paradies

German Aerospace Center DLR
Institute of Data Science
Jena, Germany
{sirko.schindler,marcus.paradies}@dlr.de

André Twele

German Aerospace Center DLR
German Remote Sensing Data Center
Oberpfaffenhofen, Germany
andre.twele@dlr.de

ABSTRACT

With the rapid growth of available earth observation data and the rising demand to offer web-based data portals, there is a growing need to offer powerful search capabilities to efficiently locate the data products of interest. Many such web-based data portals have been developed with vastly different search interfaces and capabilities. Up to now, there is no general consensus within the community how such a search interface should look like nor exists a detailed analysis of the user's search behavior when interacting with such a data portal.

In this paper we present a detailed analysis of user's search behavior based on a log analysis of a real earth observation data portal and generalize our findings to recommendations for future data portal search frontends to improve the overall user experience and increase the search quality.

Index Terms— Geoportal, Search, Query Log Mining

1 Introduction

Earth Observation (EO) data is growing rapidly in volume and increasingly scientific and commercial users demand efficient and intuitive access to value-added EO data products. EO geoportals offer such functionality by providing advanced search capabilities over the archived data. With the increasing diversification of potential user groups and different levels of EO domain knowledge, this poses a tremendous challenge to offer an intuitive yet powerful search interface that can be successfully operated by multiple user groups. Missing domain knowledge or different vocabularies often lead to underspecified queries (too many results) or unsuccessful searches (no/wrong results).

To better serve user groups with different levels of domain knowledge and experience it is essential to better understand the user's search behavior based on an analysis of a real-world system with real user queries. Real EO data portals typically log all incoming search requests, effectively providing value information about the most commonly used search keywords and additional temporal and geospatial constraints.

In the past, log data from NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC)[1] has already been analysed [3]. However, the authors focused on the technical aspects of processing large amounts of heterogeneous log data and provide only little information about the results of their analysis. Regarding the constraints used by their users, only a list of top ten keywords and their frequency is given.

In this paper we give a detailed analysis over 6 months (April–October 2018) of log data from DLR's EOWeb GeoPortal.[2] We summarize our findings and provide practical guidelines for the development or enhancement of the search interface of EO data portals.

The remainder of the paper is structured as follows. In Section 2 we introduce the EO data portal that was used for the analysis. In Section 3 we describe our analysis setup before we present our log analysis in Section 4. We summarize our findings and sketch specific search enhancement possibilities in Section 5, before we conclude the paper in Section 6.

2 EOWeb GeoPortal

The German Satellite Data Archive (D-SDA) consists of a large collection of Earth Observation (EO) data from both national and international missions maintained by the German Aerospace Center (DLR). The EOWeb GeoPortal (EGP) has been developed as a multi-mission web portal for accessing the heterogeneous data sources of the D-SDA [4]. It provides access via a set of services compliant with the standards of the Open Geospatial Consortium (OGC) as well as an open web interface that allows users to query the archive for its products and services. Users can express their information need using multiple constraints:

Collections: They can achieve a catalog-like browsing of datasets by restricting their search to single or multiple of *collections*, e.g., spotlight images from the TerraSAR-X mission. In addition, collections are ordered in a hierarchical fashion, so users may directly select all TerraSAR-X collections without the need to iterate through all of them manually.

Geospatial: Users can restrict the *spatial* extent of their

[1] https://podaac.jpl.nasa.gov
[2] https://geoservice.dlr.de/egp/

search via a map interface to a specific region of the world. They can either draw a bounding box within the map, upload an own area of interest as a Shape or KML-file, or select one from a predefined list of regions. The list covers most countries in the world as well as selected regions like central Africa.

Temporal: The third option is to restrict the search by a *time* period, which reflects the acquisition time of the satellite scene. Similar to spatial restrictions, a number of predefined values can be selected. Besides including generic time intervals like "last week", this also allows setting the time frame to the life time of a specific mission like SRTM.

Keyword: EGP allows the specification of *keywords* matching datasets' content. Predefined keywords can guide novice users through the portal and may provide experienced ones with shortcuts in their workflow. The offered options include keywords derived from thesauri such as the INSPIRE Spatial Data Themes (e.g., "Atmospheric conditions"), uniform resource names (URNs), such as, e.g., "urn:eop:DLR:EOWEB:GOME.TC", and mission related terms like "MERIS". The keyword search is modeled as a full-text search over the collection metadata provided by a OGC-compliant CSW (Catalog Service for the Web) interface.

Type: Users can also restrict their search to a specific type of result. EGP offers not only EO collections, so users can focus their search on either datasets, dataset series, or services.

The search result for EO collections can be further limited through additional filter criteria, which are derived from the product metadata. For example, this allows restricting the search to a cloud coverage of less than 20% in case of optical satellite data or a HH-polarization in case of Synthetic Aperture Radar (SAR) data.

After users identified the products of interest, they can directly order them through the EGP interface. For some products there are access restrictions in place, but most of the products are freely accessible after registration and can be downloaded or retrieved via one of the OGC-compliant services (e.g., WMS, WFS, WMTS, and WCS).

3 Methodology

We performed an offline analysis of the EGP log files for the time period between April and October 2018 of the EGP as the main source of information about current users' requirements. Before the actual analysis, the log files were stripped of any personal identifying information. They were then parsed and stored using an Elastic stack[3] pipeline. In this process each log entry was classified and deconstructed into its components like identifying spatial constraints used or an anonymous session-id to connect different requests of a single search session. This preprocessing allowed us an easy and efficient access to the various aspects of the query log.

Neither ELASTICSEARCH[4] nor KIBANA[5] supports the full extent of our analysis out of the box. Nonetheless, we tried to automate most of the data extraction tasks using manually created scripts and queries.

[3] https://www.elastic.co
[4] https://www.elastic.co/products/elasticsearch
[5] https://www.elastic.co/products/kibana

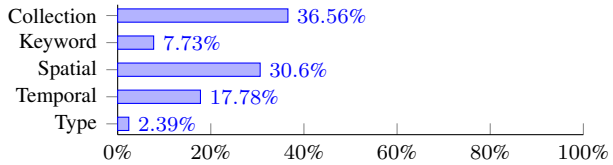


Fig. 1. Frequency of queries using at least the given constraint.

Keyword constraints were decomposed into used concepts. Note at this point the difference between terms, which are most commonly used in query log analysis, and concepts. A concept may consist of a single term, but can also contain an n-gram of terms, e.g., "digital elevation model". While the individual terms refer to rather general concepts, in conjunction this concept denotes a specific type of value-added product.

Individual concepts were subsequently categorized using information extracted from publicly available resources. In particular, we made use of Wikidata [7], as it covered a wide range of appearing concepts. Further inspection revealed several uncategorized concepts. To increase coverage we employed stemming techniques as well as manually curated mapping files to mitigate the impact of typos and similar mistakes. Uncategorized concepts after this step contain single letters and other unidentifiable sequences of either letters or numbers. We collected those in a separate category "unknown".

The use of predefined values is not tracked separately in the log files. We attempt to gauge their usage by comparing the respective restrictions with the set of predefined values. Although users may enter the exact value directly, we believe this approach to be sufficiently precise for both keyword and spatial constraints. However, it is not applicable to temporal constraints, as here the generic options like "last week" will not translate to fixed values usable for comparison.

4 Patterns of Use

In our analysis we followed established approaches in query log mining. For an overview we refer the reader to [5]. Unless otherwise noted, the following results refer to initial search requests. Requests that arise from traversing the different result pages are excluded.

In addition to common keyword-based queries, EGP allows users to apply other constraints like spatial or temporal restrictions (cf. Section 2). As shown in Figure 1 almost 31% of all queries contain at least a spatial restriction and 18% a temporal one. On the other side only 8% of queries use keywords, while type restrictions contribute to only 2% overall. The high frequency of collection-based queries is predominantly caused by the order process that requires users to browse collections before ordering.

The use of predefined constraints varies. They are barely used for spatial restrictions: only 0.64% use predefined values. On the other hand, about 21.5% of keyword restrictions make

TerraSAR-X	FIREBIRD
*DLR	*Land Cover
SRTM	Elevation
TanDEM	Terra
DEM	*Climatology, meteorology, atmosphere

Table 1. Most frequent concepts used (* - predefined option).

use of the offered options. Note, that the predefined term "DLR" accounts for more than half of those.

Our main focus then shifted to the keywords used. Prior work [6] establishes rather low numbers for terms used per query. They give an average number of terms per query at about 2.4. For EGP we observed on average 1.02 concepts per query with a standard deviation of 0.17[6]. Also the distribution of terms is highly skewed: Only 1% of the unique concepts contribute to over 25% of the keyword-restricted requests. A list of the ten most frequent concepts is given in Table 1. Kindly recall, that we refer to concepts at this point.

As mentioned in Section 3 we had to cope with various abbreviations and different spellings for some concepts. The concept for TerraSAR-X mission was, e.g., labeled using the following terms (omitting several variations of upper-/lowercase): "TSX", "TerraSAR-X", "Terrasar--X", "TerraSar x", "TerraSAR", or "Terra SAR".

A more comprehensive impression of keyword usage can be gathered from the classes of concepts used. An overview of their respective frequencies is given in Figure 2. The dominant classes relate to the initial gathering of the data with slightly over half of the concepts used referring to specific missions. Furthermore, users looked for instruments (about 4.3%) or specific observational parameters of them (around 1.5%). Also part of these provenance-related classes are organizations (about 12%)[7] and types of products (about 9.6%).

Some users chose to search for applications data products can be used for (around 10.7%). Setting aside the predefined suggestions, the most frequent concept here is "elevation", followed by "snow", "flood", and "water".

Notable is also the share of location-related information entered as a keyword (about 7.5%). These concepts identify regions of varying size reaching from generic ones like "world" or "global", over countries and areas like "Romania" or "Baltic" to specific locations like "Moscow" or "Rome". Most location are referenced by their English name. However, there are some exceptions like the Polish "Warszawa" for Warsaw or the German "Kroatien" for Croatia.

A few users use the keyword field to enter coordinates directly (below 0.5%). Here we observed values like "51.75602 / 14.31971" pointing to a location in Cottbus, Germany, or "N39E068" near the border of Uzbekistan and Tajikistan.

The remainder of concepts includes system-specific IDs (around 2%) – presumably obtained in previous sessions – and

[6]Terms per query was slightly higher at an average of 1.20 terms per query with a standard deviation of 0.60.

[7]Note that this class is largely dominated by the concept "DLR", which is also part of the suggested keywords.

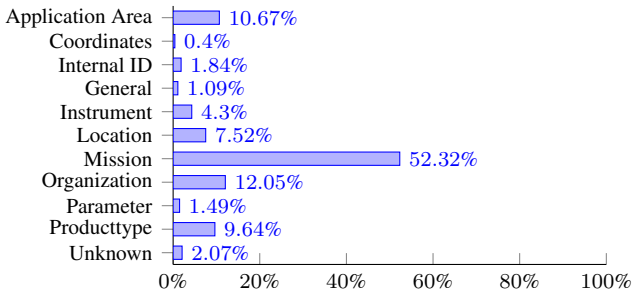


Fig. 2. Frequency of concept classes.

rather general terms (around 1%). The later consists of terms like "collection" or "data formats", which seem to indicate an information need that does not aim at a single data product. Finally, there is a number of concepts we could not assign to a specific class: fragments of terms and arbitrary numbers apparently not referring to any coordinate.

As mentioned before, the search logs showed quite some variation in the keywords used to describe a specific concept. In an ideal world, all those variations would lead to the same result set. However, for many variations we observe substantial differences. One example are the keywords "Ozone" and "O3". Both denote the same concept, but the former returns 31 results, while the latter one only matches 5.

5 Observations and Directions

The analysis of the EGP log files offered some interesting insights: First, although the predefined spatial restrictions are barely used, users seem to prefer the keyword input for the same purpose. Here, we observe a substantial amount of keywords relating to location or region names.

We can imagine different possible reasons. Independent of the actual techniques used to satisfy user requests, most popular search engines offer a single keyword input field as the default way of interaction. Users familiar with those interfaces might transfer that usage pattern to EGP and describe their information need primarily using keywords.

Another reason might be caused by the current user interface design. Predefined options for spatial restrictions are not available on the default interface itself, but need to be accessed via the "Advanced Map" menu. Users new to the system might not notice that and, hence, resort to the keyword input field.

The final possible reason concerns the selection of predefined options. While those options mostly define the scale of countries or other large regions, many location keywords refer to much smaller areas like specific cities. So users might be lacking the options there to express their search intent.

Most of the keyword-based location queries have no suitable result, as the metadata information does not include the respective terms. This problem could be tackled by adopting databases like OSMNames[8]. They offer a wide selection of geographical entities and their spatial extent. Transparently

[8]`https://osmnames.org`

translating from location-keywords to a spatial constraint could significantly improve the quality of search results here.

Another side-effect of using such databases is the support for different languages. Metadata information is usually restricted to a rather low number of languages. However, OSMNames and similar collections offer labels in a wide variety of languages. In particular, OSMNames uses OpenStreetMap [2] as a data source, which has crowdsourced the collection of geographic data including the names of locations. The result is a steady influx of updated data from a multilingual community.

Similar problems arise in other keyword classes as well. The backend currently employs a full-text search engine over the collection metadata. A collection needs to include the exact term as entered by users to appear in the results. This may fail for several reasons, thus preventing users from discovering the datasets they need. The examination of keywords found in the search logs suggests three different categories of reasons:

Typographical Issues: A first category is given by mere typos of different degree. It includes the inconsistent use of space, dash, and similar characters, as well as common misspellings. As example we refer to the different variations denoting the concept TerraSAR-X as mentioned before.

Abbreviation Issues: A second category consists of synonyms and similar relations. Besides traditional synonyms and abbreviations, we also include the various representations in different languages here. The aforementioned pair "O3" and "Ozone" is an example for this category.

Semantic Issues: The final category is comprised of semantically related terms. Metadata authors and end-users oftentimes have different backgrounds and, hence, use different terminologies. This results in a semantic gap that prevents users unfamiliar with the specific vocabulary used in the metadata from finding appropriate datasets. An example here is "height", which probably refers to the concept "elevation" as used throughout the metadata descriptions.

While all these categories will deteriorate a users search experience, there are different techniques to mitigate them. Misspellings can generally be addressed by the use of string similarity measures like Levenshtein distance or stemming/-lemmatization approaches. Similar to the aforementioned resolution of location-related terms, codelists can also counteract the effect of abbreviations by expanding the respective terms, so they concur with the usage within the metadata.

The most challenging category are semantically related terms. A brute-force approach using codelists will soon reach its limits given the vast amounts of terms and relations as well as the effort needed to maintain it. The Semantic Web [1] uses a graph-based knowledge base connecting terms using various relations. It promises to bridge the semantic gap between content creators and consumers beyond the capabilities of traditional search engines.

Beyond the aforementioned keyword-focused aspects, we recognize that other techniques can also improve users' search experience. This includes, but is not limited to using visualizations to represent the results, providing support for explorative search strategies, or recommender engines that are based on users' past interactions. However, we consider their discussion too broad and, hence, out of scope for this paper.

6 Conclusion

EO data grows rapidly in both size and topics addressed. With an increasingly broad range of possible usecases, geoportals serve as the entry point for a diverse group of users coming from a wide range of domains.

As a first step to cater to this expanded audience that might lack knowledge of terms and procedures used in the EO community, in this paper we analyzed the current user behavior in the EOWeb GeoPortal. Based on an analysis of the log files we described different usage patterns and highlighted existing issues with a focus on keyword-based queries. We outlined possible strategies to mitigate those issues and increase user satisfaction and efficiency at finding suitable EO products.

7 Acknowledgments

We would like to thank the EOWeb GeoPortal team for providing us with the log files and their ongoing support in this analysis. We especially thank Gina Campuzano Ortiz, Katrin Molch, Stephan Kiemle, and Daniele Dietrich for their valuable input and technical support.

REFERENCES

[1] T. Berners-Lee et al. The semantic web. *Scientific American*, 2001. doi: 10.1038/scientificamerican0501-34.

[2] M. Haklay and P. Weber. OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing*, 2008. doi: 10.1109/mprv.2008.80.

[3] Y. Jiang et al. Reconstructing sessions from data discovery and access logs to build a semantic knowledge base for improving data discovery. *ISPRS International Journal of Geo-Information*, 2016. doi: 10.3390/ijgi5050054.

[4] H. Rotzoll et al. From Discovery to Download - The EOWEB GeoPortal (EGP). In *PV 2015*, 2015.

[5] F. Silvestri et al. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 2010. doi: 10.1561/1500000013.

[6] A. Spink et al. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 2001. doi: 10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.0.CO;2-R.

[7] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 2014. doi: 10.1145/2629489.

EO ON-LINE DATA ACCESS IN THE BIG DATA ERA

Gaetano Pace (1), Michael Schick (2), Andrea Colapicchioni (1), Antonio Cuomo (1), Uwe Voges (3)

(1) CGI Italy, Frascati (RM), Italy, (2) EUMETSAT, Darmstadt, Germany, (3) Con terra, Münster, Germany.

ABSTRACT

EO On-Line Data Access (OLDA) in the Big Data era poses a number of challenges related to the need of providing access to large and diverse data holdings with high performance, scalability and reliability and an easy to use data discovery search and access method both from UI and standard interoperable APIs. The EUMETSAT OLDA addresses all the above points, using leading edge technology for storage management, discovery search and access, user management integration and scalability.

Index Terms— EUMETSAT, OLDA, Big Data, Kubernetes

1. INTRODUCTION

In 2017 CGI, who is representing a consortium consisting of Con terra GmbH, 52North, ask - Innovative Visualisierungslösungen GmbH, GeoSolutions and The Server Labs, has been awarded a contract by EUMETSAT [1] for engineering, development and maintenance of services in support to implementing the EUMETSAT Data Services Roadmap [2], including, among other things, the development of an EO On-Line Data Access (OLDA). This paper elaborates on the relevant developments CGI performed on behalf of and together with EUMETSAT.

Traditional EO data ordering in archive systems affects data access timeliness and is resource consuming. EO online data access in the big data era poses a number of challenges:

- High performance and fast response;
- Managing large data volumes, variability and granularity;
- A single access point providing easy data discovery, search and access both from UI and API;
- Guaranteed service quality and reliability.

Moreover, in terms of usability nowadays users' expectations are significantly influenced by popular content access services such as Netflix and Spotify. Last but not least, the solution must be cost effective:

- Reducing overall costs by: order handling automation; low cost storage and no data duplication; optimised processing resource usage;
- Minimising the refactoring of legacy infrastructures.

The very innovative EUMETSAT OLDA addresses all of the above with leading edge solutions for:

- Storage and data management
- Discovery search and access
- User management integration
- Service level scalability and reliability control

2. STORAGE AND DATA MANAGEMENT

Object storage technology can be used to implement inexpensive and easily scalable solutions. On the downside object storage typically performs worse than a distributed file system when multiple writes and reads are necessary, but this is not relevant for a data access system in which files are written once and read multiple times.

In EO multi-mission systems data often span across multiple storage domains. The OLDA storage management adopts a configurable rule engine to dynamically associate data to different Object Storage Providers, e.g. depending on mission, location, time, etc. This also supports elasticity between on-premise and external commercial cloud solutions and allows distributing data across systems with different service quality depending on the access frequency and concurrency.

One of the challenges when using object storage is to support variable data access granularity. EO products are provided in the form of Submission Information Packages (SIP), implemented by ZIP archives. Some users want to be able to access the whole package in one request, while others are only interested in a single file within the package (e.g. a single Sentinel-2 band). This issue has been tackled using metadata information of the ZIP files Central Directory (e.g. file name, byte offset and length) in combination with the range parameters of the Get Object S3 request. To support this, the following actions are performed at ingestion time:

- Identifying the target storage system based on the rule configurations;
- Retrieving metadata of the single files in the ZIP product and referencing these in the OLDA Catalogue;
- Storing the SIP into the appropriate storage system.

When retrieving the product, if the user wants to have access to a single file in the ZIP archive the data access engine uses the Central repository metadata to identify and download only the specific item from the package.

doi:10.2760/848593

3. DISCOVERY SEARCH AND ACCESS

Discovery search and access functions are implemented via a UI as well as through an API. Following the HATEOAS approach the application status is defined by the resource representation, materialized as links defining the options of the client. The UI discovery is implemented by the EUMETSAT *Product Navigator* [3], which allows browsing through the available data collections with a simple and intuitive presentation of metadata, sample pictures and general descriptions, as illustrated in Figure 1.

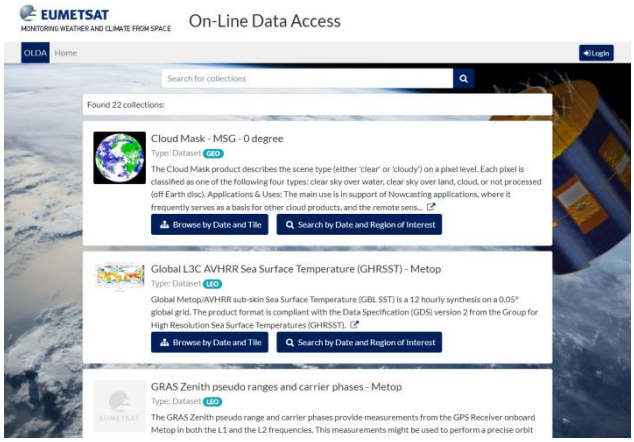


FIGURE 1 –OLDA PRODUCT NAVIGATOR

Once the user has selected a collection, it is possible to browse by date (year/month/day) and tile using Equi7grid [4]. This gridding system is based on the 7 independent continental zones where tiles are projected using Azimuthal Equidistant projections and is a particularly suitable tiling system for referencing high resolution EO data, due to hierarchical grid structure (3 tiling levels) and minimal distortions. The list of available products is dynamically updated and shown on the UI while the user interactively navigates into the tile structure, as shown in Figure 2.

Alternatively the user may also adopt the traditional AOI and TOI search.

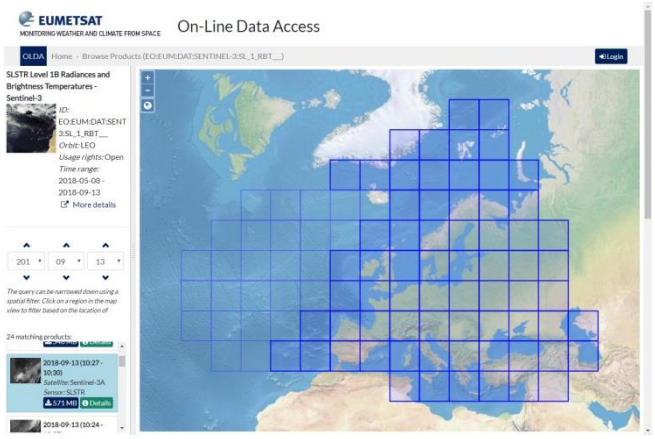


FIGURE 2 – OLDA UI USING THE EQUI7GRID

The OLDA Catalogue, indexing EO metadata, uses Elasticsearch, which supports scalability using shards and replicas. GeoHashing [5] is used for efficient AOI geographic searches.

The data access API, defined in OpenAPI (2.0/Swagger) [6] is implemented by 3 components:

- Browse API
- Download API
- OpenSearch-EO interface

The Browse API associates the needed resources to predefined URL paths, allowing implementing browsing and navigation similarly to the UI use case, like in the examples provided below.

URL request	GET Operation
/collections	Provides the metadata for every Collection available in OLDA in GeoJSON- or HTML-format depending on the requested format.
/collection{collectionId}/dates/{year}/{month}/{day}/tiles/{Zone}	The response provides the list of finer grained value ranges in the navigation.
/products/{productID}	Provides the EOP metadata of the identified product ID (see also Figure 3).
/footprints/{sensorMode}/{subSatLon}	Provides the footprint corresponding to sensor mode and sub-satellite longitude of a geostationary product.

TABLE 1- EXAMPLE REQUESTS FOR THE OLDA BROWSE API

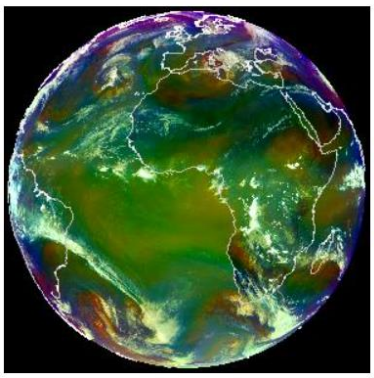
The Download API, leveraging the storage manager flexibility described above, allows retrieving the whole SIP product or parts of it. Some products and/or collections have restricted download access, which is controlled by the back-end service. The Download API enforces access control based on the JSON Web Token (JWT) provided in the HTTP header (see next section for details). The (Geo)JSON information provided by the Data Access API can be easily incorporated into programming environments such as Jupyter Notebook [7], etc.

The OpenSearch-EO interface represents an industry standard (OGC 13-026) based OpenSearch interface for search and discovery of EO collections and products. OLDA uses version 1.1 including the GeoJSON response encodings. All details can be found in OGC 13-026r9 [8], OGC 17-047 [9], OGC 17-003 [10]. The search on collection is based on text match on the collection title and/or description, while the search on products is based on mission, TOI and AOI.

doi:10.2760/848593

W_XX-EUMETSAT-Darmstadt,VIS+IR+IMAGERY,MSG3+...nc

Collection	EO:EUM:DAT:MSG:HRSEVIRI
Sensing start time	2017-11-14T11:00:09Z
Sensing end time	2017-11-14T11:12:40Z
Mission	MSG3
Instrument	SEVIRI
Size	143722



Download as SIP
EOP Metadata
EOP Metadata in JSON format

SIP Contents

- W_XX-EUMETSAT-Darmstadt,VIS+IR+IMAGERY,MSG3+...nc
- EOPMetadata.xml
- browse.jpg
- thumbnail.jpg
- manifest.xml

FIGURE 3 – EXAMPLE OF METADATA DISPLAY

4. USER MANAGEMENT INTEGRATION

Main user management challenges were:

- Supporting API access control, moving on from the UI-only SSO support;
- Integrating legacy systems with the associated user bases and diverse technologies.

The OLDA solution is based on an API Gateway, built using the WSO2 API Manager and Identity Server products [11], which is fully integrated with the legacy EUMETSAT user management system based on CAS [12] and complies with the following requirements:

- EUMETSAT users registered using CAS shall automatically be granted access to the OLDA services;
- Data collections access authorisation is based on pre-existing user attributes (e.g. organisation, type of license held) stored in the CAS as part of the registration process.

The OLDA implements the OAuth2 [13] workflow illustrated in Figure 4 for API access. In steps 1-4 the user accesses a web page on the API Gateway Store, is authenticated using the CAS pre-existing credentials and obtains API access credentials to be used in subsequent programmatic API requests.

The API access logic is represented in steps 5-8. The calling program uses the access credentials to obtain a short-lived token, which is embedded in the actual API request. The API Gateway translates this into a JSON Web Token (JWT) adding relevant user attributes (e.g. organization, license status) obtained previously from the CAS and injects it into the actual API request. The back-end service finally enforces authorization based on its internal logic.

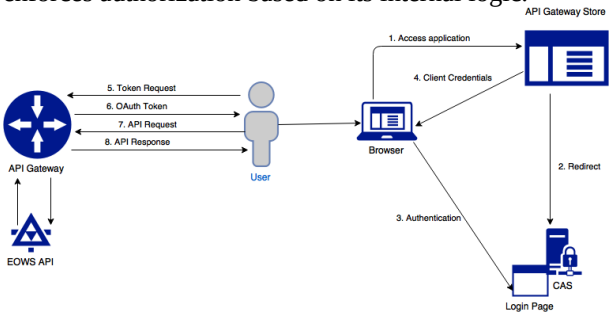


FIGURE 4 - API ACCESS SCENARIO

This paradigm has the following strengths:

- Adopts protocol transformation to simplify integration with legacy systems;
- Supports machine to machine authorisation protocols through the OAuth2 standard, opening to new possibilities for development of virtual marketplaces;
- Increases security because access tokens, differently from the access credentials, have configurable duration and scope.

5. SERVICE LEVEL SCALABILITY AND RELIABILITY CONTROL

Controlling resources dedicated to each user category allows guaranteeing predefined service levels to a large number of users, while optimizing resource costs.

Besides authentication and authorization, the API Gateway component also implements throttling, i.e. limiting the number of successful hits to an API for a given period of time, which allows to: protecting APIs from Denial Of Service security attacks; regulating traffic according to infrastructure availability; making an API, application or resource available to a consumer at different levels of service. Moreover some tests have been performed using NGINX to implement rule-based rate limiting features, based on the *leaky bucket algorithm* [14], which proved to be very efficient and will be integrated in the upcoming releases.

Scalability and reliability are based on Kubernetes [15], the world's most popular production-grade container orchestration platform, and the most important project of the Cloud Native Computing Foundation [16], which provides the following advantages:

- Velocity: evolving quickly, while staying available;
- Scalability: number of service replicas supports auto-scaling using pre-defined configurations;

- Abstraction from the infrastructure: applications deployed using Kubernetes APIs can be easily transferred between environments;
- Efficiency: applications can be co-located on the same machine without impacting the application themselves.

OLDA is deployed using the standard Kubernetes architecture, sketched below.

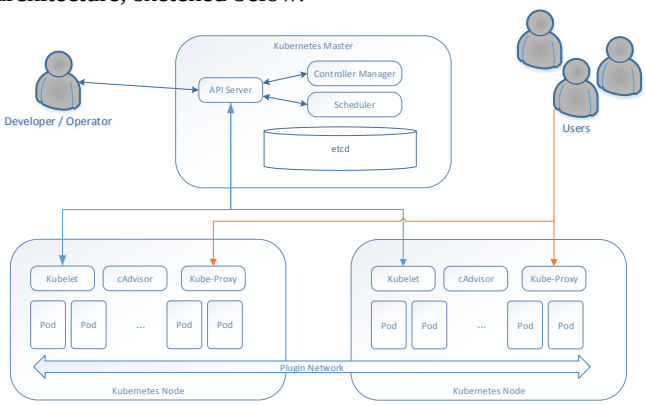


FIGURE 5 -KUBERNETES ARCHITECTURE

Kubernetes allows building self-healing systems, managing how to reach the desired state. Using Prometheus [17] it is possible to gather service performance metrics (e.g. number of calls per second of a certain web service) to be used together with the Kubernetes Horizontal Pod Autoscaler service to dynamically control the number of active replicas of a certain service depending on the service load (e.g. scaling application pods between 3 and 10 replicas, when the load of a pod exceed 100 calls per second). Health checks in Kubernetes can be implemented using liveliness probes, which are agents used to know when a container should be restarted.

6. CONCLUSIONS

The EUMETSAT OLDA uses state of the art and cost effective big data technologies to create a superior data access experience that meets the users' needs and expectations in the big data era, making it simple to find the data of interest, access it with the desired level of granularity, guaranteeing high service levels while minimizing and controlling resources costs.

A flexible data storage mechanism allows getting data efficiently and with the desired granularity without replicating storage. Kubernetes technology implements elasticity by autoscaling resources as needed. A simple modern UI navigates the user through the data providing the needed information at the right time. A data navigation and download API supports programmatic access and uses a standard OAuth2 protocol which could be easily extended in the future to replace the legacy identity server with any other identity provider (e.g. a Google or Facebook account). The OLDA is currently under service validation by EUMETSAT and is planned to be opened to EUMETSAT member state users in the course of 2019 via a pilot service phase.

7. REFERENCES

[1] EUMETSAT, "Main Website" [Online]. Available: https://www.eumetsat.int/website/home/index.html.
[2] EUMETSAT, "Data Services Roadmap" [Online]. Available: https://www.eumetsat.int/website/home/News/DAT_3112252.html.
[3] EUMETSAT, "Product Navigator" [Online]. Available: https://navigator.eumetsat.int/start.
[4] B. Bauer-Marschallinger, C. Paulik and S. Cao, "The Equi7 Grid – V13," *Grid and Tiling Definition Document – Issue 0.4*, 1 June 2016. Also available online at https://www.sciencedirect.com/science/article/pii/S0098300414001629.
[5] Geohash "Wikipedia Website" [Online]. Available: https://en.wikipedia.org/wiki/Geohash.
[6] "OpenAPI Swagger 2.0" [Online]. Available: https://swagger.io/docs/specification/2-0/basic-structure/.
[7] "The Jupiter Notebook" [Online]. Available: https://jupyter.org/.
[8] Open Geospatial Consortium, "OGC 13-026r9: OGC OpenSearch Extension for Earth Observation, V1.1", 2017.
[9] Open Geospatial Consortium, "OGC OpenSearch-EO GeoJSON(-LD) Response Encoding Standard, V 1.0.0, OGC doc 17-047 (under development)".
[10] Open Geospatial Consortium, "OGC EO Dataset Metadata GeoJSON(-LD) Encoding Standard, OGC doc 17-003, V 1.0".
[11] WSO2, "Main Website" [Online]. Available: https://wso2.com/.
[12] Apereo, "Central Authentication Service" [Online]. Available: https://www.apereo.org/projects/cas.
[13] Oauth2 "Main Website" [Online]. Available: https://oauth.net/2/.
[14] NGINX, "Rate Limiting with NGINX and NGINX Plus" [Online]. Available: https://www.nginx.com/blog/rate-limiting-nginx/.
[15] "Kubernetes" [Online]. Available at: https://kubernetes.io/.
[16] Cloud Native Computing Foundation "Main Website" [Online]. Available: https://www.cncf.io/.
[17] "Prometheus" [Online]. Available at: https://prometheus.io/.

ESA SPACE DATA AND ASSOCIATED INFORMATION LONG TERM PRESERVATION, DISCOVERY AND ACCESS

Razvan Cosac, Sergio Folco, Mirko Albani, Rosemarie Leone, Iolanda Maggio, Emilia Di Bernardo

Heritage Data Programme (LTDP+), European Space Research Institute (ESRIN), European Space Agency (ESA), Frascati, IT

ABSTRACT

Knowledge management practices ensure the identification, capture, organisation, preservation and sharing of core knowledge and information in order to continuously improve an organisation's effectiveness and efficiency in pursuing its mission [1]. In the context of Big Data, it is vital to manage the vast amount and variety of the knowledge around the data, as this knowledge facilitates our capacity to extract information and meaning from the data. The European Space Agency recognizes that knowledge represents the most valuable resource of the organisation, and therefore, that knowledge management represents a crucial aspect when considering the successful completion of the Agency's goals. This paper will give an overview of the European Space Agency's Heritage Data Programme (LTDP+) Earth Observation Data Preservation System and will discuss in more detail the ESA LTDP+ Knowledge Management System (KMS), which is composed of the OMNES platform and the Preserved Data Set Content (PDSC) Management System. These two environments focus on ensuring the long term preservation and discovery of ESA and Third Party Missions (TPM) Earth Observation knowledge and information.

Index Terms— associated information; knowledge management; preservation; discovery; access; PDSC; documentation; information; Knowledge Management System; OMNES

1. INTRODUCTION

In order to understand the present and to be able to shape the future, we need to know the past. Historical information and knowledge is key to making informed decisions. The European Space Agency has the mandate to assure the long term preservation, sharing and exploitation of space data and its associated knowledge. ESA aims to achieve these goals through the Heritage Data Programme, which is built on four main pillars:

- Preservation – preserve and manage ESA's Space Mission Data and Information
- Discovery – inventory and assure discoverability of all ESA Space Mission Data and Information
- Access – share ESA Space Mission Data and Information
- Value Adding – enhance the value of ESA's Space Mission Data and Information

In order to achieve these objectives an Earth Observation Data Preservation System (EO-DPS) was set up. This system is primarily composed of a Master Archive, a Cold Back-up data archive, and a Knowledge Management System (KMS). The main aim of the EO-DPS is to preserve the EO Mission/Sensor Data Set, which is comprised of the Data Records and the Associated Knowledge, acquired or procured by ESA EO and Third Party Missions (TPM).

Data Records include raw data and/or Level-0 data, higher-level products, browse images, auxiliary and ancillary data, calibration and validation data sets, and descriptive metadata. Associated Knowledge includes all the Tools used in the Data Records generation, quality control, visualization and value adding, and all the Information needed to make the Data Records understandable and usable by the designated community (e.g. mission architecture, products specifications, instruments characteristics, algorithms description, calibration and validation procedures, mission/instruments performances reports, quality related information). It also includes all the Data Records' representation information, packaging information and preservation descriptive information [2].

One of the main components of the ESA LTDP+ EO Data Preservation System (Fig. 1) represents the Knowledge Management System (KMS), a CCSDS (Consultative Committee for Space Data Systems) OAIS (Open Archival Information System) Reference Model compliant environment [3], which is composed of the OMNES platform (© DB Seret) and the Preserved Data Set Content (PDSC) Management System. Together, these two elements aim to ensure the long term preservation and discoverability of all ESA and TPM missions' Earth Observation space data knowledge and information.

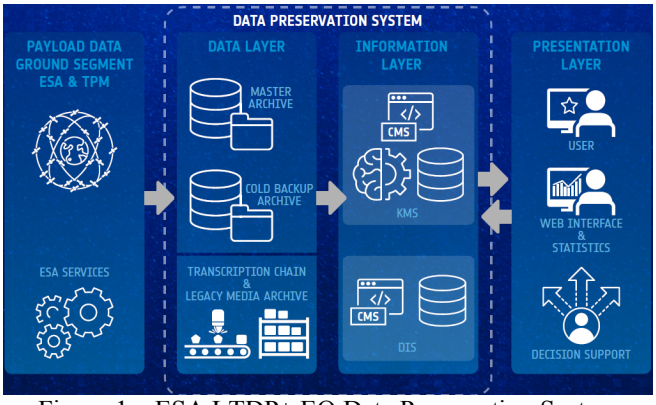


Figure 1 – ESA LTDP+ EO Data Preservation System

2. OMNES PLATFORM

The OMNES system is a complex software application used for the preservation, traceability, discovery, and access of digital resources [4] captured in various digital formats. It is designed to adopt an easy, fast and flexible ingestion process for archiving and information retrieval of digital content. The objective of the OMNES platform is to ingest digital information, such as documentation or images, and to preserve it in a digital repository, with an appropriate long term archive format.

The OMNES platform currently supports the following file formats as input:

- Native PDF (digital text and images)
- PDF with images (e.g. scanned document)
- Images in TIFF, JPG or PNG format
- MP3 audio file (currently only supported in the OMNES ISE system)
- MP4 video file (currently only supported in the OMNES ISE system)

The OMNES platform is composed of two sub-systems: OMNES ISE (Indexing and Search Engine) and OMNES LTDP (Long Term Digital Preservation).

OMNES ISE has the following characteristics:

- Conversion of digital content during ingestion process (e.g. searchable PDF/A creation, page previews, etc.)
- Full-text and metadata indexing using Optical Character Recognition (OCR) and Text Extraction processes
- Use of Dublin Core metadata for discoverability
- Web front-end for information retrieval, discovery and access (e.g. search metadata/resource and download)

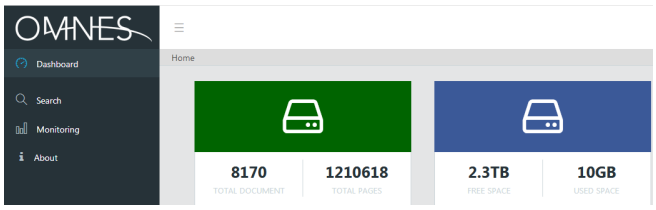


Figure 2 – OMNES ISE Dashboard

OMNES LTDP sub-system focuses on:

- Long Term Digital Preservation using FITS (Flexible Image Transport System) file format [4]
- Conversion manager to/from FITS archiving format
- Preservation Metadata Implementation Strategy (PREMIS) metadata model for long term preservation of digitalized documents and images

OMNES Ingestion Workflow (Fig. 3) comprises the following four stages:

- Pre-Ingestion
- Ingestion
- Indexing
- Long Term Digital Preservation

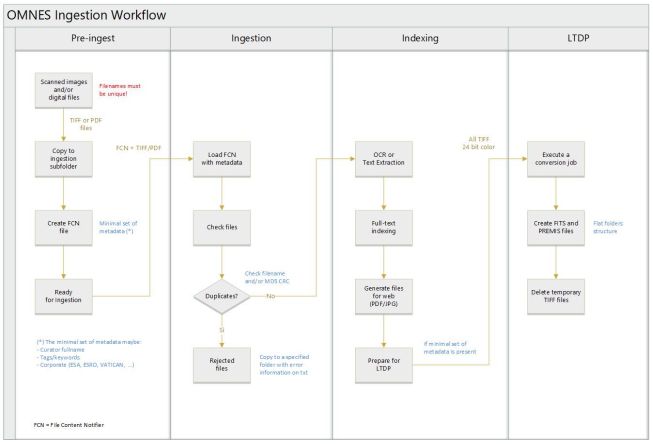


Figure 3 – OMNES Ingestion Workflow

As part of the pre-ingestion phase, documentation in digital format (e.g. PDF) or scanned images (e.g. TIFF) that are ready for ingestion are copied to the ingestion subfolder by an operator. Here, a File Content Notifier, containing a minimal set of metadata, is created for each digital resource. At this stage, additional Dublin Core metadata can be added by the operator to the FCN file. In the future it is envisaged to have templates used for the generation of documentation, which will facilitate automatic generation of Dublin Core metadata. It will be the responsibility of each document author/owner to fill in the correct corresponding information. During the ingestion stage, the files are loaded into the system, together with their corresponding FCN files. The system then checks if the digital resource has already been ingested and either rejects the file, reporting the reason for rejection, or ingests it and indexes the content. As part of the indexing phase, Optical Character Recognition (OCR) and Text Extraction processes occur, and the full text is indexed, allowing a user to search for the resource or its content, through the ISE platform. The OMNES system then prepares the resource for Long Term Digital Preservation. As part of this stage, the digital resources are converted to FITS files and PREMIS metadata is generated. The FITS files together with the PREMIS metadata represent what is preserved for the long term [5]. If, for a certain reason, a digital resource (e.g. PDF document) has to be regenerated, this can be performed starting from the FITS files and PREMIS metadata.

The OMNES platform is aimed at ensuring the preservation, discoverability and accessibility for all ESA EO Space Data Associated Knowledge, in line with LTDP+ objectives. It is initially intended for ESA internal use and will comprise the

information generated during each mission phase, from all ESA Earth Observation Programmes.

3. PRESERVED DATA SET CONTENT MANAGEMENT SYSTEM

The OMNES platform does a good job at managing digital information, such as documentation. However, as part of the Heritage Data Programme, ESA is also interested in managing and preserving the all the knowledge around the data. This is where the Preserved Data Set Content (PDSC) Management System steps in.

The PDSC Management System deals with the management of the Earth Observation Preserved Data Set Content. The platform aims to link ESA Earth Observation data records with related documentation and software, in order to be able to manage the Preserved Data Set Content [2] and trace the provenance of the data. The PDSC System has the objective of ensuring the long term preservation, discoverability and stewardship of ESA EO associated knowledge, in line with the LTDP+ Programme objectives.

The PDSC Management System is composed of an open source web enterprise environment, CMDBuild, and an Associated Knowledge repository, Alfresco. CMDBuild allows the management of IT assets' configuration so that the organisation may be able to keep under control the situation of data and associated digital resources, knowing at all times the composition, dislocation, functional relations, rules for updating over time. The CMDBuild is a flexible tool that has been customized for managing the complete EO mission associated information and data records life-cycle. The metadata used by the PDSC Management System are related to the Dublin Core metadata, while taking into consideration the particular needs for ESA Earth Observation information. The digital resources (e.g. documents) are ingested together with the corresponding metadata into the PDSC System and they are saved in the Alfresco repository, for discovery and access.

The ingested digital resources are then linked (Fig. 4) to the related data records and software tools, based on the associated metadata. The Data Curator covers an important role in this activity because he creates the link and the relations among the CMDBuild's records as well determining what the PDSC will preserve for the future. The criteria for ingesting content into the PDSC Management System is given by the tailoring of the Earth Observation Preserved Data Set Content for each Earth Observation mission/sensor/phase [2]. In this scenario the Data Curator accesses the OMNES platform to retrieve the associated information to be preserved, based on the EO mission tailoring. Other Data Curator uses cases are currently under definition, and these involve mainly the annotation, publication and presentation of data, in order to maintain the data over time for reuse and preservation.

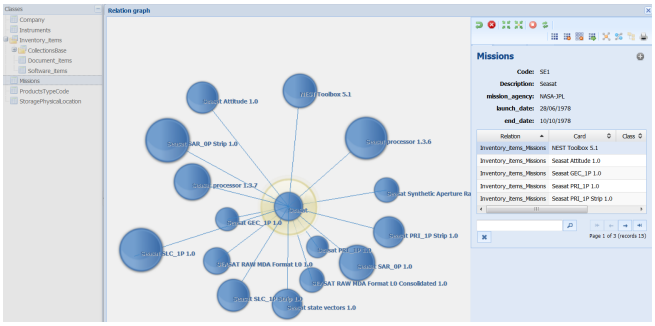


Figure 4 – PDSC data set relationship graph for the Seasat mission

Once the associated knowledge and tools are linked to the corresponding data records (including their physical location), the data set is considered consolidated from an information point of view. The CEOS PDSC Best Practices document [2] has been, and continues to represent, a crucial guide in the collection of information necessary for understanding, preserving and valorising Earth science data. At ESA's Centre for Earth Observation (ESRIN) a significant amount of associated information has been collected for each EO mission, and the PDSC Best Practices [2] have helped in identifying the missing documentation, some of which was recovered from the various ESA sites around Europe, and ingested into the PDSC Management System. The PDSC Best Practices [2] represent the fundamental criteria for identifying the necessary information to be preserved, thus avoiding individual choices and preferences.

It is important not to confuse the PDSC Management System with a catalogue. It is a system that enables EO heritage data and information discoverability, insight, exploitation and curation activities. The PDSC Management System facilitates the valorisation of Historical EO Missions Data Sets for long time data series and innovative applications, for multi-disciplinary use, favouring "cross-fertilization" with current and future missions, thus adding value to heritage data.

4. KNOWLEDGE MANAGEMENT SYSTEM AND OAIS REFERENCE MODEL

The ESA LTDP+ Knowledge Management System is compliant with the OAIS Reference Model developed by CCSDS. Considering the growing interest in the digitalisation of resources, for long term preservation, discovery and access, the OAIS Reference Model is being revised in order to adjust and extend it to new types of users and use cases, including non-space institutions such as libraries and other long term digital preservation organisations. In order to achieve this, the CCSDS DAI (Data Archive and Ingest) Working Group is finalising an OAIS architecture model and a use cases document which will embrace different types of institutions involved in long term preservation, discovery and access of data and information. Once these CCSDS

documents are approved, the ESA Knowledge Management System will be used to support the implementation of the CCSDS DAI OAIS resolutions and use cases.
These use cases include:

- Single mission: EO mission preservation and curation
- Sensor family missions: long-term data series and Fundamental Climate Records
- Space mission and in-situ data fusion
- Multi space domain missions: data cross-valorisation
- From heritage missions to new mission conception
- New space missions simulations and test bed

5. CONCLUSION

Knowledge and information are the most valuable resources of the European Space Agency, and therefore, knowledge management represents a crucial aspect when considering the successful completion of the Agency's goals. The ESA LTDP+ Programme, contributes to reaching these objectives through the established Knowledge Management System, ensuring the long term preservation, discovery and access of ESA Space Data and Associated Information.

6. REFERENCES

[1] ESA Director General's Office, *ESA Knowledge Management Policy,* European Space Agency, http://intramedia.sso.esa.int/public/corporate/ESA_KM_Policy_admin-ipol-know-2017-001e.pdf, 2017.

[2] M. Albani, R. Leone, I. Maggio, and R Cosac, *Long Term Preservation of Earth Observation Space Data – Earth Observation Preserved Data Set Content,* CEOS-WGISS Data Stewardship Interest Group, http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/EO%20Preserved%20Data%20Set%20Content_v1.0.pdf, 2015.

[3] Consultative Committee for Space Data Systems (CCSDS), *CCSDS Recommended Practice for an OAIS Reference Model*, CCSDS Secretariat, https://public.ccsds.org/pubs/650x0m2.pdf, 2012.

[4] DB Seret, *Enhancing Knowledge to seize Tomorrow's Opportunities*, DB Seret, http://www.dbseret.com/assets/pdf/DBSeret_ENG_Light.pdf.

[5] I. Maggio, *Associated Knowledge Preservation Best Practices*, CEOS-WGISS Data Stewardship Interest Group, http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS%20Best%20Practices/CEOS%20Associated%20Knowledge%20Preservation%20Best%20Practices_v1.0.pdf, 2017.

doi:10.2760/848593

A WEB OF DATA ANALYTICS SERVICES

Thomas Huang

Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109, U.S.A.

ABSTRACT

Cloud Computing has become the ubiquitous approach to our Big Data challenge. However, one will quickly discover that moving (a.k.a. forklifting) existing on-premise data analytic solutions to the Cloud doesn't always translate to costing saving and performance boost. The Cloud's elasticity, its availability, and its wide selection of computing options and selections of costing models making Cloud an attractive environment to tackle our Big Data challenge. Cloud by itself cannot tackle our daunting challenge need for analyze and derive scientific inferences through vast collections of multi-sensor measurements. We need an integrated analytic architectural solution that allows researcher to conduct science without the overhead of search and download. This paper describes the open source data analytic web architecture NASA is developing by infusing instances of Integrated Data Analytic systems next to the data. The goal is to create a connected web of analytics platforms.

Index Terms— Big Data, Distributed Analytics, Parallel Analytics, Cloud Computing, Ocean Science, OceanWorks, Apache SDAP, Apache NEXUS, CEOS, PO.DAAC, NASA

1. INTRODUCTION

Climate change is a defining issue of our time. It is touching on direct human and societal impacts. With increasing global temperature warming of the ocean and melting ice sheets and glaciers, the impacts can be observed from our coastline, and may involve drastic changes to marine ecosystems. While there is no lack of information and publications on climate change impacts on aspects such as sea level rise, floods, droughts, and hurricanes, understanding of ecosystem level impacts on and effects on flood security is critically important yet very poorly understood. Adding to the science and data integration challenges that understanding these impacts poses is the complexity of broader public and policy-maker engagement as stakeholders and fundamental determinants of future outcomes.

While much of the satellite observations from various disciplines are accessible from different data centers, the solution for analyzing decades of measurements and coordinating measurements collected from various instruments for time series analysis is both difficult and critical. Climate research is a big data problem that involves high data volume, measurements collected by various sources, methods for on-the-fly extraction and reduction to keep up with the speed and data volume, and the ability to address uncertainties from data collections, processing, and analysis.

For decades scientists have been relying on a common process flow, which includes scrape FTP sites, download data files to their local computing environment, and developing algorithms to analyze the downloaded data. Data center are only chartered to distribute file products. In this age of big data, our climate research community recognizes the traditional analytic workflow is unsustainable. While data centers do provide some tools for reduction, such as data subsetters, the size of the subsetted data may still be too large to download. A more efficient approach is to have large analytic solutions right next to the data holdings to eliminate data movement. With affordable Infrastructure as a Service (IaaS) of commercial Cloud and semantic web, we are still seeing much of the informatics community is in the business of building one-off, stovepipe tools. Users are finding themselves working with different disjoint tools and having to manually translate between different data formats and nomenclatures, often data have to be transformed into different representations to satisfy different tools requirements.

We need a web of Integrated Data Analytic systems that shares common taxonomy and provides common web-service API for access and analysis that allows the service providers to scale-up or scale-down the computing according to the requirements and user needs. The users of these services shouldn't have to be concerned about the physical computing and internal data management architecture. More importantly, these services share common taxonomy and nomenclature to enable federated analysis of different measurements.

2. DISTRIBUTED ARCHITECTURE

The Committee on Earth Observation Satellites (CEOS) Ocean Variables Enabling Research and Applications for GEOS (COVERAGE) initiative [8] is an international initiative that seeks to provide improved access to multi-agency ocean remote sensing that are better integrated with in-situ

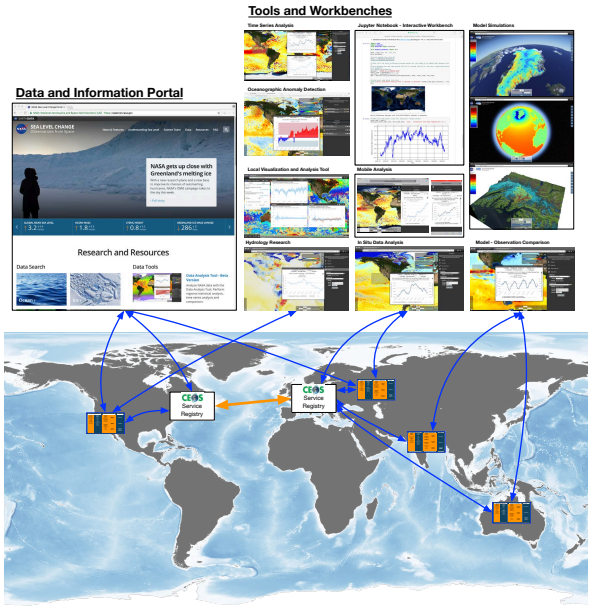


Fig. 1. Distributed Analytics Architecture

and biological observations, in support of oceanographic and decision support applications for societal benefit. While it would be ideal to have all data in one place, such as a common Cloud computing environment, such solution is unsustainable due to various factors including international policies between agencies and security requirements, access to subject matter or domain experts, and the overall cost for managing and providing open access to exabyte (EB) of data in one place. COVERAGE has taken on a distributed analytic architectural approach [1], Figure 1, where each data provider or agency can standup their own Integrated Data Analytic Platform for the data they manage. The services share common API, taxonomy, and metadata model. All analyzed results are packaged in JSON documents. This architectural approach reduces the need for unnecessary massive data movement between services and the client application will only have to develop logics to process the result JSON responses. CEOS Service Registry can be established according to the continents and/or agency alliance to serve as the data and services lookup and discovery access point.

Clients of COVERAGE include

- *Data portals* for data and climatological events discovery that link to relevant data, analytics services and published results.
- *GIS-based domain-specific data tools* that is tailored to specific science investigation and/or community. Examples of such tools include
  - NASA Sea Level Change Portal's Data Analysis Tool (http://sealevel.nasa.gov/data-analysis-tool/) [3] is an advanced data visualization and analysis tool for sea level rise research
  - The GRACE Data Analysis Tool (https://grace.jpl.nasa.gov/data-analysis-tool/) is an advanced analysis tool specifically for the GRACE data.
  - The NASA Physical Oceanography Distributed Active Archive Data Center (PO.DAAC)'s State of the Ocean Tool (https://podaac-tools.jpl.nasa.gov/soto/) is a web-based tool for physical oceanography data.
- *Domain-specific applications* which could ranges from simple scripts to advanced GIS-based programs in any programming languages (e.g. Python, Java, MATLAB, IDL, C/C++, etc.) to orchestrate search results and analytic operations.
- *Interactive workbench*, such as the popular Jupyter Notebook (https://jupyter.org), for researchers to interact with these services to create recipes to share with other researchers. Figure 2 is an example of an interactive workbench demonstrated at the 2018 CEOS SIT Technical Workshop at Darmstadt, Germany [1]. The demo generated coordinated time-series between river gauges and perception data from the Tropical Rainfall Measurement Mission (TRMM). The river time series was computed by an analytic service at the NASA JPL and the TRMM time series was produced by the analytic service hosted under the Amazon Web Services (AWS). This demo involved no data movement. The Jupyter Notebook was running on a typical laptop computer connected to the internet over WIFI. The demo shows the spike on river runoff after abnormal rate of rainfalls around February 1998 in the county of Los Angeles.

3. INTEGRATED DATA ANALYTIC PLATFORM

An Integrated Data Analytic Platform, also known as Analytic Center Framework (ACF), is an architectural concept to encapsulate the scalable computational and data infrastructures and to harmonize data, tools and computation resources to enable scientific investigations. The goal is to create a web-service platform for researchers and tools developers to discover, interact and analysis massive amount of related data without having to move data between systems over the internet. This platform must tackle both storage and software architecture together in order to fully leverage of its operating environment, such as the elastic cloud, without tying the users of the platform to a specific cloud provider and/or a specific underlying technology.

The Apache Science Data Analytics Platform (SDAP) (https://sdap.apache.org) is an open source implementation

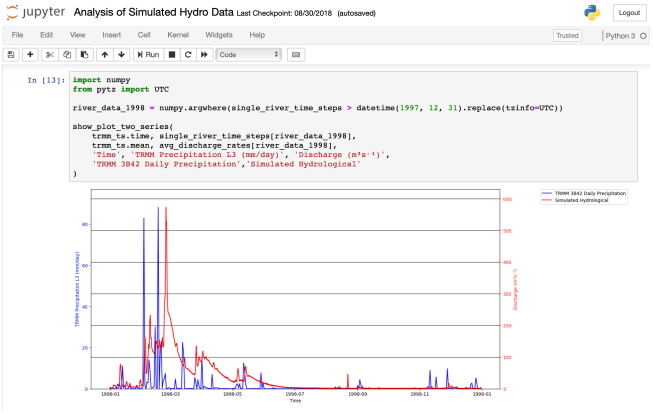


Fig. 2. Plotting Time Series between River and TRMM Precipitation Measurements. The two time series were computed by two, distributed services without data movement

of an Integrated Data Analytic Platform. The technology is the backend for the NASA's Sea Level Change Portal, NASA's GRACE science portal, and the core for the NASA's Advanced Information Systems Technology (AIST) OceanWorks technology, which will be the analytics solution for the NASA's Physical Oceanography Distributed Active Archive (https://podaac.jpl.nasa.gov) for the ocean science community. Figure 3 illustrates the architecture of an Integrated Data Analytic Platform [2]. Rather than aiming for creating a killer scientific application, the goal is to create a service platform to enable suite of scientific applications and systems. The platform can be divided into three tiers

1. *Tools and applications* – these are the clients of the platform. Their only binding to the platform is through RESTful APIs with all responds packaged in JSON documents. These clients can be implemented in any web-enabled programming languages, that is, able to make HTTP(S) calls and able to parse simple text response in JSON format. These clients have no knowledge of the physical hardware infrastructure and how the actual data is being stored.

2. *Services and Workflow* – these are the implementation of the data access and analytics webservices. They are the clients of the Analysis-Ready Storage tier. These services and workflow are built to leverage the parallel GIS-based data query and retrieval services provided by the Analysis-Ready Storage tier. It is a parallel analytic environment. The SDAP analytics services are implemented using Apache Spark for fast, in-memory MapReduce statistical analysis operations. It has no knowledge of how the data is physically stored and how the spatial indexes are being maintained. These services include area-averaged time series, climatological map, etc. The Workflow are for automated processing such as generation of climatology and large on-demand services.

3. *Analysis-Ready Storage* – it is more than a collection of disks and folders. The platform is designed for horizontal scaling, that is, to enable parallel fetching and conduct parallel analytic operations. This tier harmonizes different satellite observation data and its metadata to create a unified representation of information to simplify the development of analytic webservices and workflow systems. It is also equipped with its own workflow system to automate the discovery, transformation, and ingestion of various new observational and model data from different data providers.

The deployment of such big data analytics solution is no small task if done manually. As a horizontal-scale solution, depending on the volume and the kind of analysis, it involves orchestration of large number of compute nodes. Container deployment technology, such as Kubernetes and Docker, has matured over the years. SDAP packages all of its components and services into a collection of Docker containers where the deployment can be automated using Continuous Integration (CI) tool such as Jenkins or Atlassian Bamboo.

3.1. NASA's OceanWorks project and the Apache Science Data Analytics Platform (SDAP)

OceanWorks is an NASA Advanced Information Systems Technology (AIST) project to establish an Integrated Data Analytic Platform at the NASA PO.DAAC for big ocean science. It focuses on technology integration, advancement and maturity by bringing together several previous NASA-funded analytics projects as an effort to deliver a production-ready data science platform for the ocean science community. OceanWorks is a key part of PO.DAAC's solution for enabling big ocean science on the cloud. With NASA's upcoming Surface Water Ocean Topography (SWOT) mission that is expected to generate over 20PB of data, it is expected the future ocean science will be conducted on the Cloud to reduce unnecessary data movement. Recognizing the building blocks of OceanWorks can support multi-disciplinary Earth Science, the OceanWorks project collaborates with the Apache Software Foundation and established the Apache Science Data Analytics Platform (SDAP) (https://sdap.apache.org). The goal is to establish a community-driven and supported GIS-based big data analytics platform. The components of SDAP includes:

- NEXUS: the big data analytics engine. See the following subsection.

- Extensible Data Gateway Environment (EDGE) [4]: a GIS-based OpenSearch and metadata translation integration service for fast geospatial lookup of data and

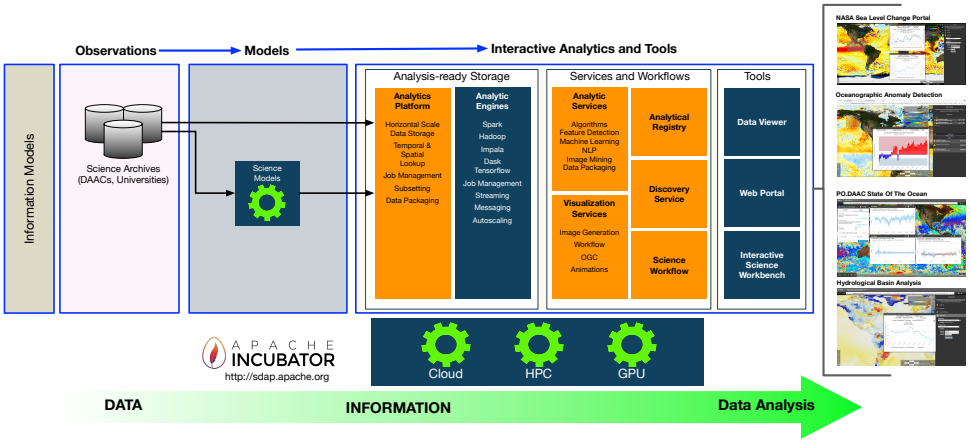


Fig. 3. Architecture for an Integrated Data Analytic Platform

translate metadata into various standards includes ISO-19115, DIF, UMM-C and UMM-G, etc.

- OceanXtremes [9]: a big data analytics solution for anomaly detection that enables to perform on-the-fly computation of daily difference by comparing observation against the climatology and provides tools for scientists to register anomalies and publish them using RSS feed.
- Distributed Oceanographic Matchup Service (DOMS) [7]: a big data analytics solution to perform on-the-fly matchup of in-situ measurements against satellite observation. To date, the in-situ data include SPURS I/II from JPL, SAMOS from the Center for Atmospheric Prediction Studies (COAPS) at Florida State University, and ICOADS from the National Center for Atmospheric Research (NCAR).
- Data relevancy [10] and event search: the data relevancy engine is a machine learning based technology to continuously analyze web search logs to dynamically rank the relevant datasets. The goal is to have the most relevant datasets listed in the beginning of the search results. The event search solution is to create relevant search respond that is encoded with space and time information. If a user searches for a specific hurricane, the responding datasets include URLs for the users to directly visualize and analyze the relevant data for a specific time period and location.

The Apache SDAP is currently under Apache Incubation process. It is in active development and infusion into various domain-specific environments.

3.2. Big data analytics engine

NEXUS, Figure 4, is an emerging data-intensive analytics framework. It takes a different approach on handling file-based observational temporal, geospatial artifacts in order to fully leveraging existing horizontal-scaling technologies like MapReduce and the elastic cloud environment. NEXUS breaks the original data file into tiles and stores tiled data in cloud-scaled databases with an added high-performance spatial lookup service. NEXUS provides the bridge between science data and horizontal-scaling data analysis. This platform simplifies development of big data analysis solutions by bridging the gap between files and MapReduce solutions.

In addition to delivering the typical analytics services such as area-averaged time series and coordination map, NEXUS is also the base analytic framework for OceanXtremes and DOMS.

NEXUS is designed to be adaptable to different deployment environments. It supports on-premise computing cluster and private/public cloud (such as AWS). It uses Apache Solr as its spatial registry for data tiles, metadata and precomputed tile statistics. For data tile management, NEXUS supports fast, cloud-based NoSQL databases like Apache Cassandra and ScyllaDB, and it also supports storing tiles in an object store like AWS S3. For data ingestion, it uses serverless architecture when operate on the AWS and uses an ingestion cluster when operate on local hardware. The goal is to create a GIS-based analytics framework that is flexible to the project needs. Since this is a webservice-based solution, the internal infrastructure is hidden from the users of this framework.

3.3. Performance

NEXUS is still evolving as the community continuously finding new ways to improve its architecture and performance. A recent benchmark was gathered to analyze 16 years of MODIS TERRA Aerosol Optical Depth 550nm [5] [6] on a point-based, regional, and global scale. The analysis involves subsetting 5790 daily files (2.9GB) and apply analysis on the subsetted data. Performance numbers were gathers between

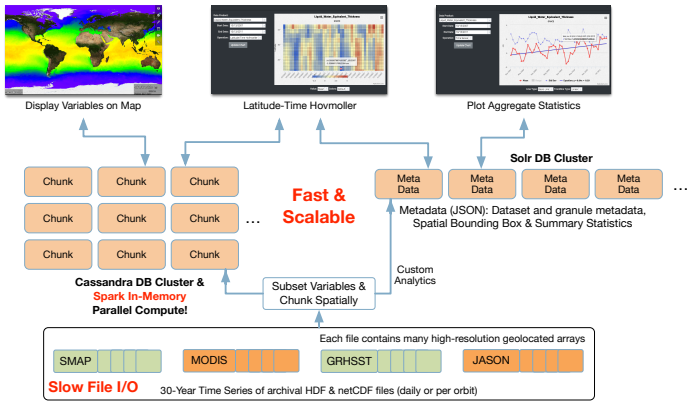


Fig. 4. NEXUS' Two-Database Architecture

NASA'G GIOVANNI, AWS EMR, and NEXUS. NASA's GIOVANNI is a popular web-based data analysis tool, that is built around file-based analysis, Figure 5, shows NEXUS outperforms the traditional analysis method by hundreds of times. What usually takes nearly 30 minutes to compute, it only took NEXUS less than 2-second to compute.

4. CONCLUSION

In the age of Big Data, we look to the Cloud as the solution to our challenge. We should consider Cloud as an instrument to our solution. In order to tackle our big data challenge and to deliver high performance analytic capacities to our climate researchers, we need to start with a scalable architecture. Our goal is to have our computing close to the data and deliver services for users to work with the data without the need of data download. The idea of Distributed Data Analytics relies on federated instances of Integrated Data Analytic systems. The demonstration and performance figures presented here have proven the importance of having a community-driven open source architecture for big data analytics in order to deliver end-to-end data management and horizontal-scale analytic services, which eliminates the need for massive data download and expensive hardware procurement for a domain-specific science investigation. The NASA OceanWorks will be infused into PO.DAAC to introduce on-the-fly capabilities to PO.DAAC's SOTO tool this year. Apache SDAP is expected to graduate from the Incubator this year as well.

5. ACKNOWLEDGEMENT

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

REFERENCES

[1] Huang, T. Architecture for Distributed Earth Science Data Analysis. In proceedings of the 2018 CEOS SIT Technical Workshop, Darmstadt, Germany, 2018.

[2] Huang, T., E.M. Armstrong, F.R. Greguska, J. Jacob, N. Quach, L. McGibbney, V. Tsontos, B. Wilson, S. Smith, M.A. Bourassa, J. Elya, S.J. Worley, T. Cram, and Z. Ji. High Performance Open-Source Big Ocean Science Platform. In proceedings of the 2018 Ocean Science Meeting, Portland, OR, 2018.

[3] Huang, T. NASA Sea Level Change Portal – It is not just another portal site. In proceedings of the 2017 American Geophysical Union Fall Meeting, New Orleans, LA, 2017.

[4] Huang, T., E.M. Armstrong, and N. Quach. Metadata-Centric Discovery Service. In proceedings of the 2012 Federation of Earth Science Information Partners (ESIP) Summer Meeting, Madison, WI, 2012.

[5] Jacob, J.C., F. R. Greguska, T. Huang, N. Quach, and B. D. Wilson. Design Patterns to Achieve 300x Speedup for Oceanographic Analytics in the Cloud. In proceedings of the 2017 American Geophysical Union Fall Meeting, New Orleans, LA, 2017.

[6] Lynnes, C., M. M. Little, T. Huang, J. C. Jacob, C. Yang, and K. Kuo. Benchmark Comparison of Cloud Analytics Methods Applied to Earth Observations. In proceedings of the 2016 American Geophysical Union Fall Meeting, San Francisco, CA., 2016.

[7] Smith, S.R., J. Elya, M.A. Bourassa, T. Huang, V. Tsontos, B. Holt, N. Quach, K. Gill, F. Greguska, S. Worley, and Z. Ji. The Distributed Oceanographic Match-Up Service. In proceedings of the 2018 Federation of Earth Science Information Partners (ESIP) Winter Meeting, Bethesda, MD, 2018.

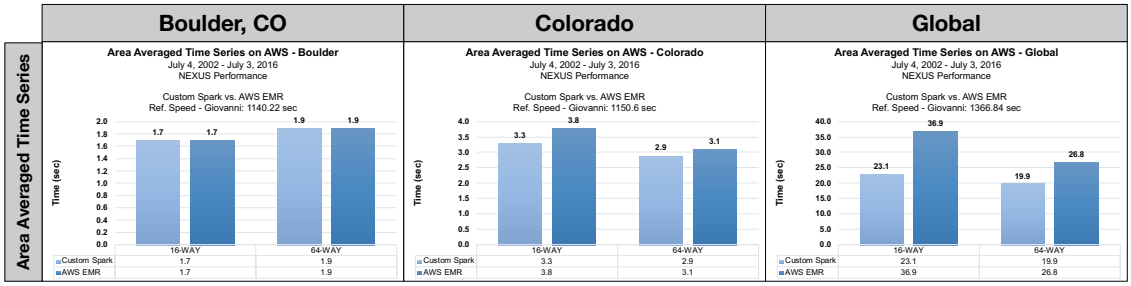


Fig. 5. NEXUS Performance compare to traditional method by the Giovanni Tool

[8] Vazquez, J., V. Tsontos, and E. Lindstrom. CEOS Ocean Variables Enabling Research and Applications for GEOS. In proceedings of the 19th International GHRSST Science Team Meeting (GHRSST XIX), Darmstadt, Germany, 2018.

[9] Wilson, B., E.M. Armstrong, T. Chin, K. Gill, F. Greguska, T. Huang, J. Jacob, and N. Quach. OceanXtremes: Scalable Anomaly Detection in Oceanographic Time-Series. In proceedings of the 2106 American Geophysical Union Fall Meeting, San Francisco, CA, 2016.

[10] Yang, C., E.M. Armstrong, M. Bambacus, K. Clarke, M. Cole, D. Duffy, S. Graves, W. Guan, Y. Jiang, K. Keiser, T. Huang, E. Law, Y. Li, Q. Liu, M. Little, D. Moroni, H. Qin, M. Rice, J. Schnase, D. Sherman, M. Xu, and M. Yu. Big Data Platform for Storing, Accessing, Mining and Learning Geospatial Data. In proceeding of 2017 American Geophysical Union Fall Meeting, New Orleans, LA, 2017.

ACTINIA: CLOUD BASED GEOPROCESSING

Neteler, M., Gebbert, S., Tawalika, C., Bettge, A., Benelcadi, H., Löw, F., Adams, T., Paulsen, H.

mundialis GmbH & Co. KG
Kölnstraße 99
53111 Bonn, Germany
www.mundialis.de

ABSTRACT

Whether participatory urban planning, digital agriculture or near real-time monitoring of flooded plains – the demand for processing large quantities of Earth Observation (EO) and geodata is constantly increasing. In addition to the amount of data to be processed, the lack of compatibility between different data systems has often been an obstacle.

The cloud based geoprocessing platform *actinia* is able to ingest and analyse large volumes of data already present in the cloud. Through *actinia*'s REST API, following the paradigm of computing next to the data, users can now process and analyse EO- and geodata. Due to the scalability of cloud platforms, insights and tailor made information are delivered in near real-time. Furthermore, methods and algorithms can be easily integrated into own business processes.

Actinia provides an open source REST API for scalable, distributed, and high performance processing of geographical data that mainly uses GRASS GIS for computational tasks.

Index Terms— Earth Observation applications, GIS, cloud based processing, geospatial analysis, open source

1. INTRODUCTION

While open geo- and/or earth observation data is increasingly available, the massive processing of such data still remains a challenge. This causes a mismatch between the enormous information potential of the data on the one side and its actual use on the other side. Organisations, businesses or administrations interested in such information still need special knowledge, appropriate software tools, access to the required data and also adequate processing power.

With *actinia* the company mundialis develops a cloud-based geoprocessing-platform, that aims at enabling users to access and process these kinds of geodata as easily as possible. The access to *actinia* is available through a web-based application programming interface (API, e.g. available at https://actinia.mundialis.de/). In order to use it, the API can be integrated into existing business workflows. For GIS experts and developers, usage of the well documented API and the power of *actinia* is more suitable.

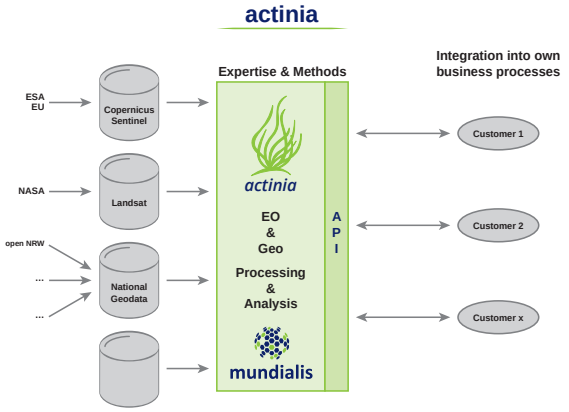


Fig. 1. Workflow of *actinia* geoprocessing engine.

It is also possible to define own processes by chaining the powerful processing-tools provided by *actinia*. Since *actinia* supports space-time cubes through its GRASS GIS backend (https://grass.osgeo.org), data aggregation over time, zonal statistics and more, the users are provided with latest technology being applied to Sentinel time series. Figure 1 shows the general workflow of *actinia* geoprocessing engine. Multiple data sources can be integrated into the different processes for EO and geodata within *actinia* which can be connected to different business processes via API.

The *actinia* engine consists of several components: i) *actinia-core* (available at https://github.com/mundialis/actinia_core), ii) *actinia-gdi* with an interface to Geonetwork Open Source for the access to a metadata catalogue (under development), and iii) *actinia-plugins* for domain-specific applications (under development).

In the remainder of the paper we will illustrate the architecture of *actinia* (user, job and data management, backend), as well as *actinia* process chains and plugins. This is followed by means of deploying *actinia* locally, on the cloud and embedded systems. After the description of integration

with other systems we briefly describe a use case, followed by conclusions.

2. ARCHITECTURE OF ACTINIA

The REST API of *actinia* allows the user to process satellite images, time series of satellite images, arbitrary raster data with geographical relations and vector data. The REST interface accesses, manages and manipulates the GRASS GIS database via HTTP GET, PUT, POST and DELETE requests. Processing of raster, vector and time series is performed on data located in persistent or in ephemeral GRASS GIS databases. *Actinia* supports the processing of cloud based data, for example all available Landsat 4-8 scenes as well as all Sentinel-2 scenes. The API endpoints are realized with the Python framework called Flask-RESTful.

2.1. User management

Actinia provides a sophisticated user, user role and user group management, that allows administrators to specify fine granular access to *actinia* resources as well as time and memory consumption. The following specific concepts are implemented in *actinia*:

- user role hierarchy: superadmin – admin – user – guest, with restricted access to *actinia* endpoints for each user role,
- read-only access to the global *actinia* persistent GRASS GIS spatial database commonly used for base geodata,
- read-write access to the user group specific persistent GRASS GIS spatial database,
- maximum size of the computational region (i.e. amount of allowed pixels), and maximum number of processes for a single process for each user,
- process specific maximum computational and enqueue time for each user.

2.2. Job management

Several *actinia* REST API endpoints are provided to enqueue and manage jobs. The *actinia* REST framework supports asynchronous and synchronous processing and therefore provides a job queue to manage the execution of *actinia* process chains.

Actinia will keep track of the processing time of each computational process as well as the size of the user specific data. Processes can be terminated by their owners or administrator. Each job has a maximum runtime and will be terminated by the queuing system if the runtime was exceeded to avoid broken jobs to block the computational subsystem.

All jobs as well as their logs and resources are written temporarily into a Redis database for fast REST API access and persistently into an elastic search database for analytical and accounting tasks.

2.3. Data management

Actinia uses the GRASS GIS database to store global and user specific geo-data in i) ephemeral or ii) persistent databases.

i) Ephemeral databases exist only for the computational time and will be deleted after processing has finished. However, the computational results of ephemeral processing are available via object storage as GeoTIFF files for raster data or zipped GeoJSON files for vector data.

ii) Two kinds of persistent databases are provided in *actinia*: The read-only global persistent database and the read- and write-able user group specific persistent database. Both databases can be accessed read-only from processes that work on ephemeral databases. The *actinia* REST API allows the import of any online available geo-data source into ephemeral and user group specific persistent databases. Geo-data can be raster images, vector feature collections or raster collections. Import of online resources can be directly defined in an *actinia* process chain.

Actinia provides several endpoints to list all data of user specific accessible persistent databases. These endpoints provide functionality to query raster, vector and raster time series data and to analyze their metadata like data descriptions, valid time and transaction time stamps as well as statistics generated at runtime.

2.4. Backend

The core of *actinia* is GRASS GIS [3], an open source geoinformation system (`https://grass.osgeo.org`). It comes with an integrated temporal framework that provides spatio-temporal capabilities, delivering comprehensive functionality to implement a fully featured temporal geographic information system (GIS) based on a combined field and object-based approach [2]. Through that it manages spatial fields of raster, three-dimensional raster and vector type in time series referred to as space-time data sets (STDS).

2.5. Actinia process chains

GRASS GIS provides over 500 processing modules. The *actinia* REST API that is used to access the GRASS GIS backend supports the definition of process chains, in which any number of GRASS GIS modules can be defined and their data exchanged. The *actinia* process chain approach provides a great flexibility to define very complex analytic processes that involve many GRASS GIS modules as well as external tools like GDAL (Geospatial Data Abstraction Library). However, complex analytic tasks are already available as simple REST

endpoints that require no knowledge about process chain creation. Some of these endpoints require the installation of the *actinia-plugins* (see Sec. 2.6). They include following functionality:

- creation of Landsat 4-8 and Sentinel-2 raster time series,
- NDVI computation for arbitrary Landsat 4-8 and Sentinel-2 scenes,
- zonal statistics on raster time series based on vector polygons,
- spatio-temporal sampling of and algebraic operation on raster time series,
- raster and vector data export.

2.6. Actinia plugins

Actinia can be extended using plugins. Plugins provide endpoints for very specific computational tasks, that can be exposed as single REST endpoints. At the moment there are two plugins available: i) processing plugin for satellite data from Landsat and Sentinel-2 satellites; ii) statistical analysis plugin for raster and raster time series data as well as spatio-temporal sampling.

2.7. Support for GRASS GIS addons

The *actinia* framework supports any addon that is available for GRASS GIS. Moreover, developers and power-users can use the PyGRASS [5] and the GRASS GIS Temporal framework [2, 1] in GRASS GIS to implement high performance spatio-temporal applications that can directly be used as processes in the *actinia* process chain.

2.8. Authentication

The *actinia* REST framework uses basic HTTP authentication using user/password or authentication tokens. REST API endpoints are available to create authentication tokens for specific users.

3. DEPLOYING ACTINIA

Actinia can be deployed on a wide range of hardware and cloud environments.

3.1. Personal computer and local data center

Actinia can be deployed on a workstation computer or local data center as being a Python package together with GRASS GIS and a Redis database server. Moreover, *actinia* and all of its sub-components can also be deployed as docker image at these systems as well.

3.2. Public and private Cloud

Actinia and its sub-components: geospatial backend (GRASS GIS), user and job database (Redis), and logging (fluentd, elasticsearch, kibana) can be deployed either as docker swarm or Kubernetes cluster in scalable cloud environments. It is designed to run in thousands of containers in parallel. *Actinia* provides a granular locking mechanism, so that users that work in parallel on the same persistent database, can not interfere or corrupt each others data. *Actinia* has been successfully deployed in the Google Cloud, Amazon AWS, CloudFerro, the Open Telecom Cloud and many more.

3.3. Embedded systems

Actinia requires a Linux environment, Python3, GRASS GIS and Redis to work. These components are available on many embedded systems, which makes *actinia* fit for running in rough environments that are populated by IoT devices, remote controlled robots and drones as well as satellites to provide real-time and batch processed sensor and image analyses.

4. INTEGRATION WITH OTHER SYSTEMS

4.1. Business logic

The REST API based approach of *actinia* in combination with *ad hoc* processing of literally any online available geo-data using ephemeral databases makes *actinia* suitable to be used in business logic systems which require high performance spatial and spatio-temporal analysis and processing.

4.2. openEO wrapper and User Defined Functions

The OpenEO H2020 project (http://openeo.org/) aims at developing an open API to connect R, Python, Javascript and other clients to big Earth Observation cloud back-ends in a simple and unified way [4]. GRASS GIS is one of the back-ends; the GRASS GIS driver uses the *actinia* REST API to send processing requests based on the *actinia* process chain approach to the GRASS GIS backend (https://github.com/Open-EO/openeo-grassgis-driver). The current public endpoint is available at http://openeo.mundialis.de:5000/capabilities. The full archives of Landsat 4 - 8 and Sentinel-2 are accessible and in near future also Sentinel-1.

The support for openEO approach of User Defined Function (UDF) is currently under development and will run as part of an *actinia* process chain (https://open-eo.github.io/openeo-udf/api_docs/).

4.3. Copernicus DIAS

As part of the European Union's Copernicus program several Data and Information Access Services (DIAS) platforms are currently being implemented. They consist of cloud-based

infrastructures that will make Copernicus data available to its users along with tools and marketplaces to offer third party applications. While *actinia* is cloud vendor agnostic and compatible to Google Cloud or Amazon AWS; it would be possible to deploy *actinia* on the new European DIAS platforms.

4.4. *Actinia* shell in GRASS GIS

GRASS GIS was recently extended to support the creation of valid *actinia* process chain JSON building blocks by simply calling a GRASS GIS module with all of the required parameters. At time the development of a GRASS GIS shell extension is under way, so that module calls that operate on the local GRASS GIS database can be directly sent to an *actinia* server for remote execution. Hence, complex process chains can be tested on a local GRASS GIS installation and then send directly via HTTP request to a remote *actinia* server that resides near the cloud based geo-data (Sentinel-2, Landsat, MODIS, ...) without the requirement of deploying *actinia* locally.

5. POTENTIAL OF ACTINIA IN THE FIELDS: AGRICULTURE, LULC AND FORESTRY

Automated Sentinel-1 and Sentinel-2 time series preprocessing using *actinia* is a prerequisite for numerous Earth Observation applications. For agricultural monitoring, we developed a cutting frequency index besides crop type classification. Furthermore, Land Use/Land Cover (LULC) time series status and change maps are derived in *actinia* using machine learning algorithms. For example, LULC maps are the basis for modeling and decision making in the field of water management and projecting the water consumption in the future in the region of Drâa in Morocco.

Being weather independent and generated by active sensors, SAR images are most adequate to monitoring tropical rain forest disturbances. Sentinel-1 image processing requires large RAM which becomes an issue when working with desktop computers, e.g. using SNAP (Sentinel Tool Box). For massive SAR data processing, SNAPPY (SNAP Python API) as the Python interface to the SNAP processing technology is most promising and allows the integration of SNAPPY based processing chains into the *actinia* cloud environment. As *actinia* itself is mainly based on GRASS GIS, the SNAPPY processing chain can be wrapped as a standard GRASS GIS Python addon. A study area in Guyana of a managed forest concession is selected to evaluate the potential of *actinia* to monitoring the forest logging damages between the years 2015 and 2017 using Sentinel-1 images. The *actinia* based processing shows a considerable processing time reduction due to the automation of the processing steps (example: i) EO analyst – 50 min for data selection, download, manual preprocessing, index creation; *actinia* – ¡ 5 min for data selection and automated processing), allowing for fast decision making and support through improved reporting in the context of sustainable development and climate change commitments.

6. CONCLUSIONS

The cloud based geoprocessing engine *actinia* follows the central paradigm of Big Data to "bring the software to the data". It is cloud vendor agnostic open source software. Its functionality be extended through user plugins written in Python, C, or originating from other software packages. The user management supports group, roles and access rights. Process chains can be written in JSON (developer friendly), an *actinia* shell is available (power user friendly); it supports integration into own business processes through API (business friendly). To comply with the general data protection regulations it can be deployed in European cloud solutions.

REFERENCES

[1] S. Gebbert and E. Pebesma. The GRASS GIS Temporal Framework: Object oriented code design with examples, January 2017. URL https://trac.osgeo.org/grass/browser/grass/trunk/lib/python/docs/src/Temporal-Framework-API-Description.pdf.

[2] S. Gebbert and E. Pebesma. The GRASS GIS temporal framework. *International Journal of Geographical Information Science*, 31(7):1273–1292, 2017. doi: 10.1080/13658816.2017.1306862.

[3] M. Neteler, M.H. Bowman, M. Landa, and M. Metz. GRASS GIS: A multi-purpose open source GIS. *Environmental Modelling & Software*, 31:124–130, 2012. doi: 10.1016/j.envsoft.2011.11.014.

[4] E. Pebesma, W. Wagner, P. Soille, M. Kadunc, N. Gorelick, M. Schramm, J. Verbesselt, J. Reiche, M. Appel, J. Dries, A. Jacob, M. Neteler, S. Gebbert, C. Briese, and P. Kempeneers. openEO: an open API for cloud-based big Earth Observation processing platforms. In *EGU General Assembly Conference Abstracts*, volume 20, page 4957, April 2018. URL https://meetingorganizer.copernicus.org/EGU2018/EGU2018-4957-1.pdf.

[5] P. Zambelli, S. Gebbert, and M. Ciolli. Pygrass: An Object Oriented Python Application Programming Interface (API) for Geographic Resources Analysis Support System (GRASS) Geographic Information System (GIS). *ISPRS Int. J. Geo-Information*, 2(1):201–219, 2013. doi: 10.3390/ijgi2010201.

ADVANCES IN INTERACTIVE PROCESSING AND VISUALISATION WITH JUPYTERLAB ON THE JRC BIG DATA PLATFORM (JEODPP)

Davide De Marchi and Pierre Soille

European Commission, Joint Research Centre (JRC)
Directorate I. Competences, Unit I.3 Text Data Mining, via Fermi 2749, 21027 Ispra (VA), Italy

ABSTRACT

The JRC Big Data Platform (JEODPP) is serving JRC projects and their partners for any big data application with emphasis on geospatial data. It has recently evolved into a multi-petabyte scale platform offering advanced web enabled services for container-based batch processing, remote desktop, and interactive analysis and visualization. This later service, based on Jupyter, has recently unfolded into a complete and powerful prototyping environment. This paper describes the most significant advances in this area.

Index Terms— deferred processing, Sentinel, Copernicus, visualisation, Jupyter, JupyterLab, IPython

1. INTRODUCTION

The results presented in this paper are implemented on the JRC Big Data Platform (JEODPP) [4]. This platform serves the needs of JRC policy support activities requiring big data capabilities with emphasis on geospatial data as well as any data sources associated with a geolocation (news articles, official statistics, pictures, etc.). The platform is implemented on a commodity hardware solution scalable to the multi-petabyte scale. This is achieved through distributed storage (CERN EOS) coupled with a cluster of computing nodes for distributed computing. The JEODPP platform can be viewed as a three layer pyramid with a multi-petabyte scale storage and processing basis. The first layer accommodates massive batch processing. The second layer provides a remote desktop environment with all software needed for further developing legacy applications. The tip of the pyramid (third layer) provides interactive visualization and analysis in a web-based environment integrated in a Jupyter notebook [1].

This paper concentrates on the advances regarding the interactive analysis and visualization layer. The following aspects are detailed: evolution from Jupyter to JupyterLab, availability of new data collections, the possibility to execute arbitrary Python code, and applications for users without programming capabilities exploiting the temporal dimension of geospatial data cubes.

2. FROM JUPYTER TO JUPYTERLAB

JupyterLab[1] is an evolution of Jupyter released to users in February 2018. It provides a series of features improving the user-experience such as the use text editors, terminals, data file viewers, and other custom components side by side with notebooks in a tabbed work area [2]. In particular, it gives the possibility to redirect the map view area in a side map.

3. DATA COLLECTIONS

Since its inception, the JEODPP platform has been characterized by providing users with a wide variety of raster and vector geospatial datasets. Complex raster collections such as those originating from the Sentinel-1 and Sentinel-2 satellites are available alongside continental or global open-source DEMs (EU-DEM, SRTM, GEBCO, MERIT, etc.). They can be interactively visualised together with vector datasets such as NATURA2000, EFFIS forest fires, administrative territorial units, and land use-land cover at European level (Corine Land Cover and Urban Atlas). New datasets are continuously integrated in the interactive mode. This is for instance the case of the new global DEM at a resolution of 30 meters produced by the Japan Aerospace Exploration Agency (JAXA) and called "ALOS World 3D AW3D30" [6]. The deferred mode processing functions of the JEODPP APIs allow to obtain high-impact visualizations such as that presented in Fig. 1 displaying the coloured hill-shading of the ALOS DEM with a custom colour scheme.

Other dataset recently made available in the interactive mode are the Sentinel-1 global mosaic [5], and a cloud-free Sentinel-2 global mosaic calculated from images acquired in 2017 [3]. No less important is the availability of many basemaps that can be used as background of views and ranging from the classic OpenStreetMap, OpenTopoMap, up to MODIS data with daily granularity, high resolution aerial images, maps in neutral colors that better enhance the geographic content superimposed on them. The selection of the basemap to use or of the dataset to be displayed, takes place in a very simple way that allows the user to view the datasets

[1] JupyterLab: `https://jupyterlab.readthedocs.io`

doi:10.2760/848593

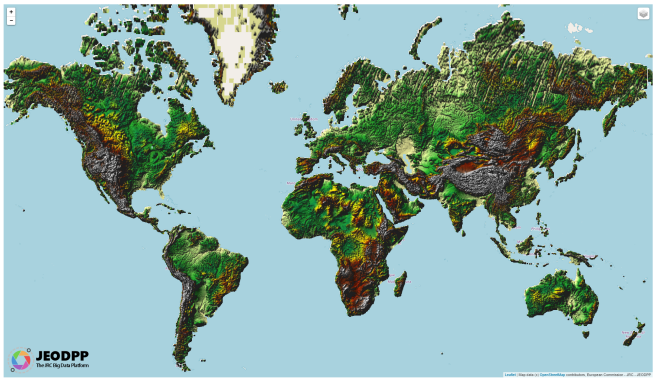


Fig. 1: ALOS Global Digital Surface Model "ALOS World 3D 30m (AW3D30)" rendered on-the-fly on the JEODPP with a colored hill-shading whose parameters are user-defined.

Stubble burning detection for Sentinel2:

```
B04 < 1000 AND
B06 < 1200 AND
B08 < 1200 AND
B11 > 500  AND
B11 < 1600
```

Function definiton that implements the stubble burning algorithm:

```
def stubble(v4, v6, v8, v11min, v11max):
    global img
    img = numpy.ones_like(band0)
    img[band0>=v4] = 0
    img[band1>=v6] = 0
    img[band2>=v8] = 0
    img[numpy.logical_or(band3<=v11min,
                         band3>=v11max)] = 0
```

Multi-band processing chain containing the python function call:

```
bands = ["B04","B06","B08","B11"]
p = coll.processMulti(bands)
        .execute(stubble,1000,1200,1200,500,1600)
        .band(0).scale(0,1).colorCustom(["Lime"])
```

Fig. 2: (Top left) Informal description of the Stubble Burning detection algorithm. (Top right) Python implementation of the algorithm using Numpy methods. (Bottom) Multi-band processing chain containing the execute step calling the custom Python function.

available in a tree structure, easily navigable and searchable using keywords as well as the self-completion functions available in the Python language. Finally, place names originating from CartoDB[2] can be overlaid on the displayed layer.

4. EMBEDDING PYTHON CODE IN THE TILE ENGINE

At the base of the interactive component of JEODPP there is a library developed in C++ language that represents the real heart of the Tile Engine, a highly parallelized component that creates in real-time and in deferred mode, the raster tiles to send to the visualization based on the ipyleaflet widget[3].

After selecting the datasets to be displayed, processing chains transforming and processing the input data in order to extract the desired information are defined. The basic elements of these chains are a series of processing steps that have been implemented within the Tile Engine through the integration of open source processing libraries or libraries developed over the years at the JRC. These libraries (mialib and pktools, among others) provide the main functions necessary for the processing of geographical data and allow for the creation of complex processing chains based on morphological operators, image classification and segmentation functions, band arithmetic, and image filtering in space and time.

Although the list of processing steps is rather extensive and comprehensive, the need has emerged for adding user-defined functions. In a language like Python, dozens of very efficient libraries manipulating images are available (e.g., Numpy, Scipy nd-image, Scikit-image, Python Imaging Library, and OpenCV). By adding a special processing step, called execute, which allows the user to send to the Tile Engine that operates server-side, any function written in Python and that could potentially use such libraries.

[2]CartoDB: `https://carto.com`
[3]Ipyleaflet: `https://github.com/jupyter-widgets/ipyleaflet`

Thanks to the Python inspect module, the source lines of the user function are read and sent to the C++ Tile Engine server, where a Python on-the-fly interpreter is instantiated. The code is then executed within the interpreter at each tile request. The function can freely modify the input image pixels to pass them to the next step of the processing chain.

As an example, in Fig. 2, an application of the execute processing chain to the detection of stubble burning (the deliberate setting fire of the straw stubble that remains after wheat and other grains have been harvested) from Sentinel-2 images, based on a simple algorithm that takes as inputs the bands B04, B06, B08 and B11 (Short Wave Infra Red band).

This new development is opening many new scenarios to the JEODPP users that gain a completely new flexibility in the analysis and processing of geospatial datasets. Moreover, it will allow, in the near future, to also take benefit from the many Machine Learning libraries available in Python that could be injected inside the server-side processing chain to extract valuable information from EO data using artificial intelligence techniques.

4.1. Widget enabled applications

Researchers or scientists who possess programming capabilities can easily get into using python to interact with the viewing and analysis environment within Jupyyer notebooks. Despite this, the need has emerged to create simpler tools that would allow to work with geographic datasets through a graphical user interface. This is even more true for the manager level users and policy officers having no or very little programming knowledge. For these users, an interface exploiting the analysis/viewing capabilities without the need to write code is needed. Within the Jupyter world, this can be effectively and efficiently enabled by using the ipy-

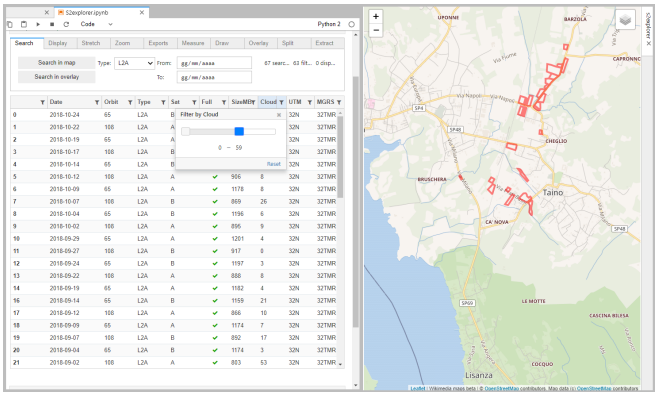


Fig. 3: S2-Explorer notebook in action. Loading of a custom vector shapefile on the map, search of the Sentinel-2 products that cover it and filtering on cloud cover percentage.

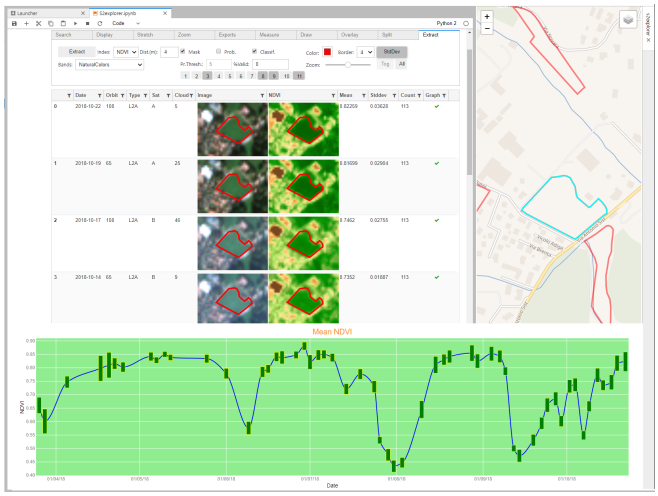


Fig. 4: Extraction of the multi-temporal NDVI profile over the pixels inside a polygon by the S2-Explorer tool.

widgets suite[4] for the creation of user interfaces and on the use of components such as Bqplot[5] for charting functions and Qgrid [6] for displaying alphanumeric data in rows with intuitive scrolling, sorting, and filtering controls.

The first fully fledged application implemented, called the S2-Explorer, is devoted to easy browsing and searching of Sentinel-2 products. It consists of a tabbed interface containing many functions going from searching of products covering the current view extents, displaying one or multiple products on the map, selecting between many possible RGB bands compositions, apply local stretching to the products visualization, and filtering of the searched products on multiple criteria (for instance on cloud cover or product type or acquisition dates, etc.). Input capabilities allow the users of S2-Explorer to easily add a custom vector shapefile to the map, and use it to select the Sentinel products covering its features. Many export functions are available, e.g. export the list of the selected Sentinel-2 products to be used for batch processing operations, export of the map view in high resolution TIFF file, creation of an animation video that displays one after the other all the filtered products. The measure and draw tabs can be used to measure distances and areas on the map and the creation of vector features with the possibility to save them in a vector format. Figure 3 shows a snapshot of the S2-Explorer in action.

5. EXPLOITING THE TEMPORAL DIMENSION

The Copernicus constellation of EO satellites is characterized by the high temporal resolution of its acquisitions. As an example, after the lunch of Sentinel-2B satellite, all emerged lands will be covered by a new optical product at least every five days. This gives space to many applications were the exploitation of the short revisit time allows for important analytical results, whether for the agricultural, the forest, or disaster monitoring applications. The JEODPP S2-Explorer leverages on the temporal dimension of Sentinel products by providing a series of analysis tools based on multi-temporal data.

5.1. Temporal profiles

One of the tabs of the S2-Explorer tools is dedicated to the extraction of multi-temporal information over the pixels covered by polygonal features. After having searched and filtered the Sentinel-2 products, users can ask the system to calculate, for each of the products, the mean value of an index (e.g., NDVI, EVI, or SAVI) or of a band, together with their standard deviation and plot both values on a temporal graph where the line series represent the mean value inside the polygon and the vertical bars the homogeneity measured by the standard deviation (see Fig. 4). This function has many applications and can be activated also on a custom edited polygon. It has to be noted that the extraction is done server-side on the highly parallel cluster and gives the results in few seconds even in case of dozens or hundreds of input products. Once the desired measurement is determined, it can be scaled to datasets of any size via the batch processing service.

5.2. Easy comparison using the split map control

The ability to easily compare two different views of the same area, whether based on two different datasets or on two different processing chains applied to the same data, is a fundamental capability of any geospatial data viewer. It can be used for comparing two complementary datasets (like a DEM and a EO image, or a basemap and a vector dataset, etc.) or two

[4]ipywidgets: `https://ipywidgets.readthedocs.io/en/stable/`
[5]Bqplot: `https://github.com/bloomberg/bqplot`
[6]Qgrid: `https://qgrid.readthedocs.io/en/latest`

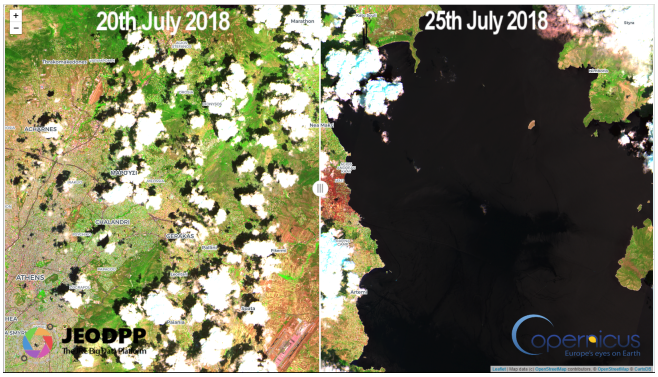


Fig. 5: Temporal comparison of two SWIR RGB compositions, from Sentinel-2 images acquired before and after the 2018 Mati fires in Greece.

Fig. 6: Video overlay of a georeferenced temporal video showing one after the other all the Sentinel-2 images acquired over an area within a selected time period.

acquisition dates from the same sensor. This is implemented thanks to a new function available on ipyleaflet: the split map control. When activated, the tool displays a vertical line at the center of the map and the two datasets on each side of the line. The user can move the line horizontally to quickly compare the left and right map on the same geographic area. This is illustrated in Fig. 5 in the case of forest fires by comparing the last Sentinel-2 product acquired before the event with the first acquired after it.

5.3. Georeferenced temporal video

Recent versions of ipyleaflet allow for the overlay of videos on top of the map. We are using this new function to create georeferenced videos starting from the multi-temporal Sentinel-2 products acquired over the same area. The videos can be played directly over the map (even while zooming and panning) and give an interesting insight on the evolution of the land over time with possible application in agriculture (like the evaluation of harvesting times) or in forest monitoring (e.g. control of illegal deforestation activities).

6. OUTLOOK

Batch processing and remote desktop capabilities are fundamental to any platform aiming at extracting insights from massive datasets. Interactive analysis and visualization are also essential to leverage on the increasing number and diversity of the available datasets. They help users to discover new datasets and combine them for prototyping new information extraction workflows. In this respect, the possibility to execute arbitrary code opens an avenue for countless possibilities and in particular has unfolded the interactive analysis and visualisation layer into a complete and powerful prototyping environment. In addition, the code used in the interactive mode is decoupled from the complexity of the visualisation engine so that it can be directly used for batch processing for which precise analysis can be performed in any desired projection. On the other hand, higher level interfaces based on widgets are increasing the outreach and impact of geospatial data by attracting users with no programming skills. Future developments of the JRC Big Data Platform include the integration of machine learning in all layers of the platform while expanding the variety of the available datasets including those not related directly to geospatial data in support to decision and policy making.

7. REFERENCES

[1] De Marchi, D. et al. “Interactive visualisation and analysis of geospatial data with Jupyter”. In: *Proc. of the BiDS’17*. 2017, pp. 71–74. DOI: 10.2760/383579.

[2] Granger, B. and Grout, J. “JupyterLab: Building Blocks for Interactive Computing”. Slides of presentation made at SciPy’2016. 2016. URL: http://archive.ipython.org/media/SciPy2016JupyterLab.pdf.

[3] Kempeneers, P. and Soille, P. “Optimizing Sentinel-2 image selection in a Big Data context”. *Big Earth Data* 1.1–2 (2017), pp. 145–158. DOI: 10.1080/20964471.2017.1407489.

[4] Soille, P. et al. “A Versatile Data-Intensive Computing Platform for Information Retrieval from Big Geospatial Data”. *Future Generation Computer Systems* 81.4 (Apr. 2018), pp. 30–40. DOI: 10.1016/j.future.2017.11.007.

[5] Syrris, V. et al. “Mosaicking Copernicus Sentinel-1 data at global scale”. *IEEE Transactions on Big Data* (2018). DOI: 10.1109/TBDATA.2018.2846265.

[6] Takaku, J. and Tadono, T. “Quality updates of AW3D global DSM generated from ALOS PRISM”. In: *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. July 2017, pp. 5666–5669. DOI: 10.1109/IGARSS.2017.8128293.

THE PANGEO BIG DATA ECOSYSTEM AND ITS USE AT CNES

Guillaume Eynard-Bontemps[1], *Ryan Abernathey*[2], *Joseph Hamman*[3], *Aurelien Ponte*[4], *Willi Rath*[5]

[1]Centre National d'Etudes Spatiales (CNES), Toulouse, France,
[2]Columbia University / Lamont Doherty Earth Observatory, New-York, USA,
[3]National Center for Atmospheric Research (NCAR), Boulder, USA,
[4]Ifremer, Univ. Brest, CNRS, IRD, Laboratoire dOcanographie Physique et Spatiale (LOPS), IUEM, Brest 29280, France,
[5]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany.

ABSTRACT

Pangeo[1] is a community-driven effort for open-source big data initially focused on the Earth System Sciences. One of its primary goals is to enable scientists in analyzing peta-scale datasets both on classical high-performance computing (HPC) and on public cloud infrastructure. In only a few years, Pangeo has grown into a very productive community collaborating on the development of open-source analysis tools for science. It provides a set of example deployments based on open-source Scientific Python packages like Jupyter[2], Dask[3], and Xarray[4] that bring together scientists and developer with their actual use-cases.

In this paper, we first describe Pangeo ecosystem and community. We then present its impact on the work of scientists from CNES on the HPC deployment there. We conclude with a future outlook for Pangeo in this agency and beyond.

Index Terms— Pangeo, Dask, Jupyter, HPC, Cloud, Big Data, Analysis, Open Source

1. PANGEO

1.1. Motivations

The science community is facing several building crises: datasets are growing exponentially (see 1) and legacy software tools for scientific analysis cannot handle them; a growing technology gap between the technological sophistication of industry solutions (high) and scientific software (low); the fragmentation of software tools and environments renders most science research effectively unreproducible and prone to failure.

Pangeo's core mission is to cultivate a collaborative environment in which the next generation of open-source analysis tools for ocean, atmosphere, climate and eventually other sciences can be developed, distributed, and sustained. These tools must be scalable in order to meet the current and future challenges of big data, and these solutions should leverage the existing expertise both inside and outside of the geoscience community.

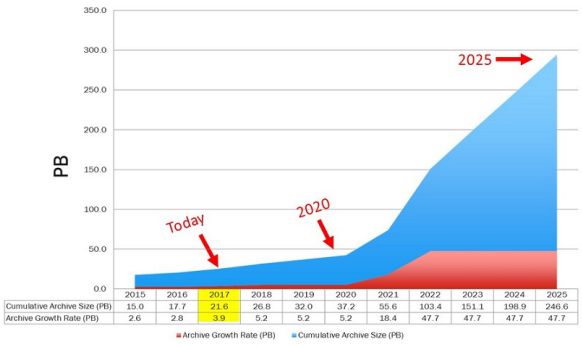


Fig. 1. Projected NASA EOS Cloud storage[5].

1.2. Community

Rather than being controlled by a single organization, Pangeo is a community-driven project, in the model of successful open-source software like Linux, Python, and Jupyter. In this spirit, all Pangeo discussions and products occur on GitHub[6], where anyone can easily get involved in the community. As of today, Pangeo spans a list of people from different governmental agencies, universities, or private companies, and from different countries (the USA, the UK, France, Austalia to name a few). By making the collaboration as broad as possible, Pangeo successfully leverages shared expertise to accomplish things no institution could alone.

1.3. Technology contributions to the Scientific Python stack

Pangeo's software ecosystem fits directly into the Scientific Python stack, involving packages such as Numpy, Pandas, or Sickit-learn. Three Python software projects are at the core of Pangeo:

- Dask is a library for parallel and distributed computing that coordinates with Pythons existing scientific soft-

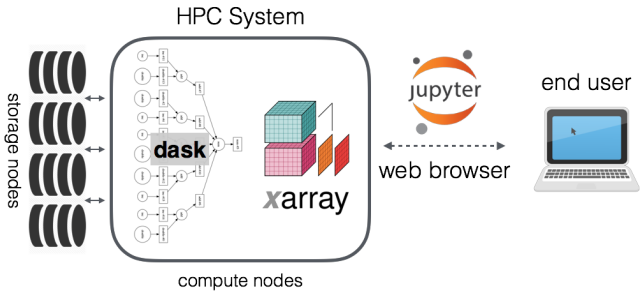


Fig. 2. Pangeo platform main components.

ware ecosystem. In many cases, it offers users the ability to take existing workflows and quickly scale them to much larger applications.

- Xarray is the interface for working with big datasets: it provides a Pandas-like API for labelled n-dimensional arrays and has backends for established and upcoming self-describing community data file formats and access protocols like netCDF, GeoTIFF, OPeNDAP, and Zarr. Xarray transparently integrates Dask arrays and hence enables users to easily scale their work to massively parallel computations.
- Jupyter: Jupyter notebooks and Jupyter Lab enable interactive computing and analysis from a web browser, and JupyterHub adds multi-user support. Jupyter notebooks are quickly becoming the standard open-source format for interactive computing not only in Python, but also in languages such as Julia and R.

There are several developments that either started in or are fueled by the Pangeo community. Modules to automatically deploy Dask distributed clusters in various infrastructures are being developed: dask-kubernetes for Kubernetes clusters and thus for the public cloud, dask-jobqueue[7] for HPC systems using scheduler such as Slurm, PBS Pro or LSF, and dask-yarn for YARN clusters (i.e. Hadoop). The integration of a Zarr backend in Xarray paved the way to directly access cloud-based object storage in parallel computations.

1.4. How Pangeo compares to other solutions

Pangeo primarily offers tools and environments for interactive or batch analysis of scientific datasets at scale. At the heart of this sentence are the three packages mentioned above: Jupyter for interactive, Dask for scale, Xarray (and Dask) for scientific datasets. As such it can be compared to a lot of existing tools. We can think of Apache Spark, Rasdaman, SciDB, Myria, or even more specific libraries like TensorFlow or Orfeo ToolBox (OTB). First ones are already mentioned in a paper about imagery analysis [8] (where they are compared to Dask only), which gives a good first overview.

Pangeo is not another Datacube or scientific database, but you can build one using its core packages, as the Open Data Cube team is doing. Xarray backed by Dask allows powerful and at scale complex data manipulation such as regridding, datasets fusion and so one. Another demonstrated use is to perform some Google Earth Engine like analysis using Pangeo [9].

Dask can be compared to Spark, but is more versatile: it does not only handles big collections or SQL like queries, it can also perform distributed nd-arrays operations, or any kind of inter node multiprocessing workload using Delayed or Future API. There has been some experiment to use Spark for Dask/Xarray like analysis with NASA SciSpark, but it was not really conclusive.

Dask and Xarray can be leveraged interactively from a Jupyter notebook plugged to a compute infrastructure (e.g. HPC or Kubernetes cluster in the Cloud). A user can make use of the computing power with a few lines of code and perform manipulation on tens of Terabytes of data as if handling some Megabytes on its own laptop.

Pangeo is not a focused scientific library such as OTB or TensorFlow. It provides more high level means to use these domain specific modules: it could be used to provide some ways to orchestrate complex OTB applications, or launch them at scale over thousand of images. It could also be used to prepare data before feeding it to TensorFlow or other Deep Learning libraries.

1.5. Deployment

The GitHub community offers online documentation, scripts and other tools to link them together in order to deploy a Pangeo platform (see 2) and put the software stack on HPC systems or in the public cloud. The main elements allowing to build and use the platform in the cloud are a set of scripts and documentation that allows automatically creating the necessary cloud infrastructure (a Kubernetes autoscaling cluster), a Helm-chart and associated Docker image, which heavily relies on Jupyterhub solutions. This is currently available for Google Cloud Engine (GCE), and work is underway for documenting Amazon Web Services (AWS) use. Continuous Deployment tooling is also used for automating new Pangeo versions deployment with Hubploy and CircleCI.

To facilitate learning and using the Pangeo software stack at scale, there is a Binder deployment[10] enabling the live execution of any compatible github repo (providing environment description and some example notebooks) in a single click, on Pangeo existing Google cloud infrastructure.

1.6. Applications and data

There are already several applications documented in earth-system sciences that can be found online[11]. These examples can be directly executed on cloud infrastructure from a

web browser, associated datasets are also published on Pangeo public bucket. Other scientific domains have developed an interest in the Pangeo approach, including Satellite imagery analysis [9], Astronomy or Neuroscience.

As the computational resources necessary for most of these applications can only be met with distributed computing, the data must be in a cloud or distributed storage friendly format, like Zarr for multi-dimensional data, Cloud optimized Geotiff for satellite imagery, or Parquet for tabular data. See [12] and [13] for more details on this crucial point.

2. CNES DEPLOYMENT AND USE CASES

2.1. Context and projects

The Centre National d'Etudes Spatiales (CNES) is the government agency responsible for shaping and implementing France's space policy in Europe. As such it covers a wide range of subjects: Ariane launcher, Sciences, Earth observation, telecommunication and defense.

On the ground segment processing side, it is involved in several Big Data projects, most of them being hosted in CNES Data Center: one of the Gaia data processing center; Sentinel product Exploitation Platform, for sharing and online processing of Copernicus products; Surface Water and Ocean Topography (SWOT) mission...

CNES HPC platform is also hosting a lot of other processing involving remote sensing data, flight dynamics, altimetry or other domains.

2.2. Pangeo on our HPC System

CNES main processing platform is a modestly sized High Performance Computer named HAL. Its computational resources are about 8000 Intel cores and 6PB high bandwidth storage. It uses PBS Pro jobqueue system to schedule the load on compute nodes and handle user and project resource sharing.

Pangeo platform has recently been deployed on the cluster, which basically means the configuration of two main components: a JupyterHub, and Dask through dask-jobqueue (see 3 for a graphical overview). JupyterHub has been deployed on a Virtual Machine, which has direct access to HAL cluster through PBS commands. This allows configuring JupyterHub Batchspawner which launches user notebooks using PBS *qsub* command, alongside Wrapspwaner to be able to select adequate system resources for launched notebook.

CNES directly contributes to dask-jobqueue in order to improve its usability on HAL. This Python module deployment (alike Xarray or other domain scientific library) is quite simple as it can be done through conda or pip packaging system. All of this is documented, and some demonstration notebooks have been shared internally.

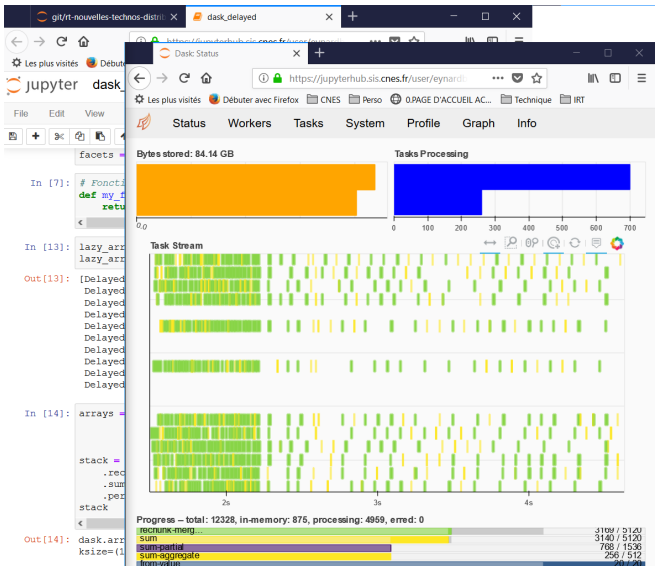


Fig. 3. Computation in Jupyter and Dask dashboard.

2.3. From embarrassingly parallel to more complex workflows with Dask

One important use of our cluster is to do repetitive jobs: apply the same computation or process to several inputs, which can for example be a list of files or a list of parameters. This is usually done with job arrays PBS mechanism. Results are then written into our central storage facility, and a final job gathers and consolidates them if needed. There are three main drawbacks to this approach:

- When scaling this mechanism to hundreds of thousands and short (under one minute) jobs, this can lead to PBS scheduler contention and slow responsiveness.
- This often means a lot of bash script and machinery to chain several analysis together, leading to workflows that are hardly readable and difficult to maintain.
- As results can be really small and are exchanged through a centralized storage system, this also means a lot of I/O load onto the File System and side effects for other users.

Pangeo, mainly through Dask, gives an adequate solution to all these problems (see 4 for an example). All the workflow, cluster creation, data management parts are handled from Python code. Reservation to PBS are done using dask-mpi or dask-jobqueue. No need to write or exchange data through disk, all data management is done through memory or high-speed network. This allows users to analyze the result of a simulation in the same piece of code where it has been launched, and eventually gather only the reduced valuable part for later use. The result of all this is elegant and

```
# Create cluster and scale to 8 nodes
from dask_jobqueue import PBSCluster
cluster = PBSCluster(cores=24, memory="120GB",
                     interface='ib0')
cluster.scale(8)

# Connect to cluster
from dask.distributed import Client
client = Client(cluster)

# Submit a function on a list of inputs
futures = client.map(my_costly_simulation,
                     input_params)
results = client.gather(futures)
```

Fig. 4. Embarrassingly parallel workload using Dask.

simple Python code, which can scale easily to thousands of cores and does not stress our HPC cluster.

2.4. Surface ocean currents analysis at scale

Ifremer (Institut franais de recherche pour l'exploitation de la mer) recently used Pangeo tools and the CNES HPC cluster to develop and run one of their workflow. Ifremer had already used Dask before for scientific applications[14].

Surface ocean currents from a global high resolution numerical simulation are analyzed in order to compute time-frequency kinetic energy spectra. Spectra are averaged zonally in order to emphasize meridional variations of these spectra. These spectra may be compared to observed estimates in order to validate the numerical model for example. Such validation are critical because these numerical simulations are used in order to perform Observing System Simulation Experiments (OSSEs) for missions that are either under development (e.g. SWOT) or proposed (e.g. SKIM). The input data consists of a collection of snapshots which is a layout that does not allow computations that are global in time such as an *fft*. The analysis thus starts with a rechunking of the data into larger temporal chunks and smaller spatial ones. Spectra are then computed and averaged in latitude bins. Each of the latter stage leverage the distribution of existing sequential *fft* code to Xarray objects via the *apply_ufunc* method.

3. CONCLUSION

Pangeo is a powerful ecosystem which enables science at scale on Cloud or on premise infrastructure. We encourage every lab, government agency, or even industry players to take a look at what it provides. The community is open and eager to welcome new users and collaborators. At CNES, we decided to focus on this tooling for our shared computing infrastructure, and it is already showing its power and its benefit. Ongoing work is focusing on developing more and more use cases with Pangeo to identify where it is most useful, and possible limits to the software stack. We are also trying to participate actively in the community and share our vision and needs, helping to steer the common effort in a direction beneficial to CNES researchers.

REFERENCES

[1] Ryan Abernathey, Kevin Paul, Joseph Hamman, Matthew Rocklin, Chiara Lepore, Michaem Tippett, et al. (2017): Pangeo NSF Earthcube Proposal.

[2] Thomas Kluyver et al: Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016.

[3] Dask Development Team: Dask: Library for dynamic task scheduling, 2016.

[4] Stephan Hoyer and Joseph J. Hamman: xarray: N-d labeled arrays and datasets in python. *Journal of Open Research Software*, 5, apr 2017. doi: 10.5334/jors.148.

[5] Mark McInerney, ESDIS Project Deputy Project Manager/Technical: EOSDIS Cloud EvolutionNASA EOSDIS web site.

[6] Ryan Abernathey et al: Pangeo github project issue tracker.

[7] Joseph Hamman, Matthew Rocklin, Jim Edwards, Guillaume Eynard-Bontemps, Loic Esteve (2018): Scalable interactive analysis workflows using dask on HPC Systems.

[8] Parmita Mehta et al (2016): Comparative Evaluation of Big-Data Systems on Scientific Image Analytics Workloads.

[9] Scott Henderson, Daniel Rothenberg, Matthew Rocklin, Ryan Abernathey, Joseph Hamman, Rich Signell, and Rob Fatland: Cloud Native Geoprocessing of Earth Observation Satellite Data with Pangeo.

[10] Joseph Hamman, Ryan Abernathey: Pangeo meets Binder.

[11] Pangeo community: Pangeo use cases.

[12] Ryan Abernathey: Step-by-Step Guide to Building a Big Data Portal.

[13] Matthew Rocklin: HDF in the Cloud — Challenges and solutions for scientific data. doi: 10.3233/978-1-61499-649-1-87.

[14] S. Fresnay, A. L. Ponte, S. Le Gentil, and J. Le Sommer: Reconstruction of the 3-D Dynamics From Surface Variables in a High-Resolution Simulation of North Atlantic. DOI: 10.1002/2017JC013400.

[15] Joe Hamman: Pangeo applications for NASA Earth Observing Data.

QUOTING AND BILLING: COMMERCIALIZATION OF BIG DATA ANALYTICS

Ingo Simonis

Open Geospatial Consortium (OGC), London, UK

ABSTRACT

Big Data processing chains working on geospatial data such satellite imagery have been subject of research for many years and are now being operationalized in standardized ways. Well-defined interfaces and standardized data exchange formats have been developed that make working with satellite imagery end-user friendly. Recent advances in the standardization domain even support the 'applications to the data' paradigm that allows executing arbitrary applications close to the physical location of the data. What was missing so far was a smooth integration of commercial principles, such as solid authentication and authorization, as well as quotation and billing mechanisms. This paper outlines how the integration of those commercial key elements can happen.

Index Terms— spatial data infrastructures, markets, standards, big data, billing, security

1. INTRODUCTION

Historically, space and more precisely earth observing satellites has been an industry only space administrations could handle. The picture is changing with more private industry now operating their own satellites, though most of earth-observing data is still produced by space administrations. Nevertheless, given that prices to build and launch a satellite into orbit have dropped from USD 200M a decade ago to just around 200k, space becomes more accessible and with the growing amount of data, new exploitation avenues are opening. The data itself is becoming much more useful for businesses and governments. In particular the recent advances in machine learning and artificial intelligence allow developing applications that provide previously inaccessible insights on global-scale economic, ecological, social and industrial processes.

At the same time, the availability of cloud based storage and processing environments and the widespread acceptance of container technologies open new exploitation pathways. Data does not need to be sold anymore. Computing cycles, input/output transactions, and on-demand storage capacities form new markets. An essential requirement for these emerging markets are robust billing and quotation, authentication and authorization,and auditing services. These services combined with new data usage principles are the cornerstones towards a new era of Big Earth Observation data commercialization. Data is not purchased anymore, but rented just for the time of analysis. This approach avoids high overhead costs for data that might be used only once before being replaced by newer data.

The private sector with companies such as Digital Globe has made their petabyte sized satellite data archives available on a "pay-per-use basis" for quite some time now. Lately, space agencies such as ESA follow these developments by outsourcing satellite product storage, access, and processing to commercial cloud operators. Data that has been formally available for purchase is now offered through a number of commercial partners for rent. End-users now enter in contracts and service level agreements with cloud operators, but instead of downloading the data, it is rented for the time of processing. Rental fees depend on provider and nature of data, with more and more data offered royalty-free. In such a case, payments are being made afterwards based on used computing cycles or other parameters typical to cloud environment, such as input/output transactions or temporary storage volume, but less often on a "per-product" basis. In such an environment where data is offered by a multitude of providers, where applications can be developed and offered by third parties that barely get in contact with the end-users, and consumers that are interested in the results produced by the applications rather than the original satellite data, standards play an essential role. In the following, this paper will layout a standards-based software architecture that allows such a market to grow and succeed. It will identify the various decoupled players and their respective roles, and identify important standards and design principles that allow a geospatial Big Data market to be successfully established. It introduces a billing and quoting model that has been implemented and tested in the recent OGC Innovation Program initiative Testbed-14.

2. STANDARDIZED BIG DATA PROCESSING

Standardization efforts are under way that allow the ad-hoc deployment and execution of arbitrary applications close to the physical storage location of the data. The Open Geospatial Consortium (OGC) has worked for the last two years on a set of standards and software design principles that allow a vendor and platform neutral secure Big Data processing ar-

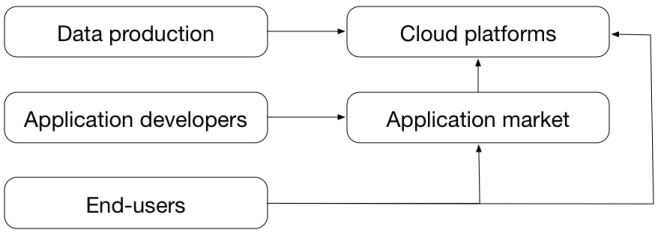


Fig. 1. High level software architecture

chitecture. Driven by requirements set by European Space Agency and Natural Resources Canada, OGC has developed a software architecture that decouples the data and cloud operators from earth observation data application developers and end-consumers.

2.1. Decoupled Roles for Successful Markets

As shown in figure 1, data generated by a data production organization is hosted and made available for access by a second organization that operates a cloud platform. Application developers make their products available on application markets that are used by end-users for application discovery. An application market can be organized as an independent entity, though in its current setup, it is rather tightly coupled to the cloud platforms. A connection that needs to be relaxed in future. To ensure maximal interoperability between all components, the envisioned architecture illustrated in figure 2 is leveraging OGC standards and has been complemented lately with security and billing and quotation elements to further qualify for competitive market situations.

Data producers nowadays often host their data on cloud platforms that are physically and organizationally decoupled, i.e. the actual hosting of the data and the provisioning of additional cloud features is outsourced to third parties. In our working example, ESA data is provided on cloud platforms operated by Cloudferro, Vito, or Amazon. These cloud operators generate revenue by selling cloud services on top of the actual data. Often, download capabilities are limited and cloud services such as deployment and execution of containerized applications are offered instead. This approach allows independent application developers to generate applications that operate on the data. These applications follow emerging OGC standards to ensure interoperability across the various cloud platforms. In future, end-users will discover the applications on application markets similar to the Apple's App store or Google's Play market. Thanks to standards, these applications can be deployed and executed on the various cloud platforms seamlessly. Results are made available through standardized data access services.

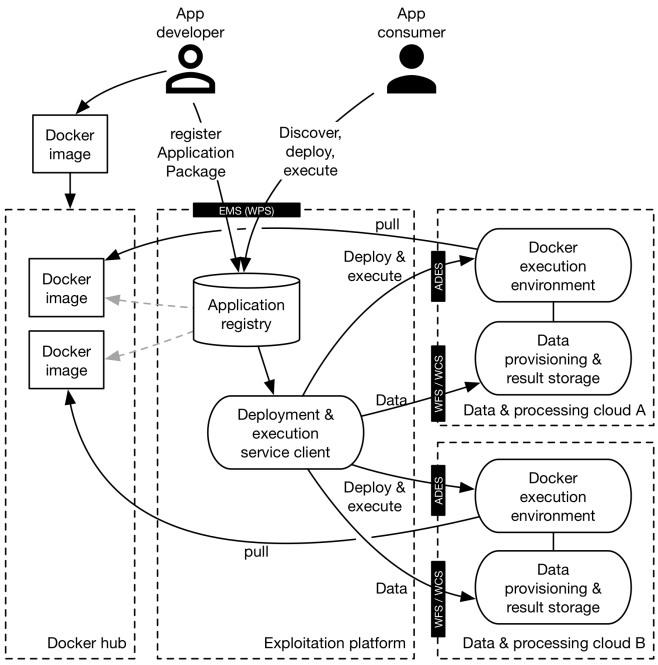


Fig. 2. Detailed Architecture with ADES, EMS, and AP

2.2. Big Data Processing Architecture

The architecture, described in full detail in [1] and outlined in figure 2, builds primarily on three emerging standards: The application deployment and execution service (ADES), the execution management service (EMS), and the application package (AP) standard. All three standards have been initially developed in OGC Innovation Program initiative Testbed-13 and are available online [2, 3]. They are currently revised in Testbed-14, with final results expected in early 2019.

As illustrated in figure 2, the Execution Management Service (EMS) represents the front-end to both application developers and consumers. It makes available an OGC Web Processing Service interface that implements the new resource oriented paradigm, i.e. provides a Web API that follows REST principles (WPS v3.0). The API supports the registration of new applications. The applications themselves are made available by reference in the form of containerized Docker images that are uploaded to Docker Hubs. These hubs may be operated centrally by Docker itself, by the cloud providers, or as private instances that only serve a very limited set of applications. Initially developed to deploy applications only, the EMS is now emerging into a workflow environment that allows application developers to re-use existing applications and orchestrate them into sequential work-flows that can be made available as new applications again. This process is transparent to the application consumer.

The Application Package (AP) serves as the application meta data container that describes all essential elements of an application, such as its functionality, required satellite data,

other auxiliary data, and input parameters to be set at execution time. The application package describes the output data and defines mount points to allow the execution environment to serve data to an application that is actually executed in a secure memory space; and to allow for persistent storage of results before a container is terminated.

The execution platform, which offers EMS functionality to application developers and consumers, acts itself as a client to the Application Deployment and Execution Services (ADES) offered by the data storing cloud platforms. The cloud platforms support the ad-hoc deployment and execution of Docker images that are pulled from the Docker hubs using the references made available in the deployment request.

Once application consumers request the execution of an app, the exploitation platform forwards the execution request to the processing clouds and makes final results available at standardized interfaces again, e.g. at Web Feature Service (WFS) or Web Coverage Service (WCS) instances. In the case of workflows that execute a number of applications sequentially, the exploitation platform realizes the transport of data from one process to the other. Upon completion, the application consumer is provided a data access service endpoint to retrieve the final results. All communication is established in a web-friendly way implementing the emerging next generation of OGC services known as WPS, WFS, and WCS 3.0.

3. BILLING AND QUOTING

Currently, satellite image processing still happens to a good extent on the physical machine of the end-user. This approach allows the end-user to understand all processing costs upfront. The hardware is purchased, prices per satellite product are known in advance, and actual processing costs are defined by the user's time required to supervise the process. The approach is even reflected in procurement rules and policies at most organizations that often require a number of quotes before an actual procurement is authorized.

The new approach outlined in this paper requires a complete change of thinking. No hardware other than any machine with a browser, which could even be a cell phone) needs to be purchased. Satellite imagery is not purchased or downloaded anymore, but rented just for the time of processing, and the final processing costs are set by the computational resource requirements of the process. Thus, most of the cost factors are hidden from the end-user, who does not necessarily know if his request results in a single satellite image process that can run on a tiny virtual machine, or a massive amount of satellite images that are processed in parallel on a 100+ machines cluster. The currently ongoing efforts to store Earth Observation data in Datacubes adds to the complexity to estimate the actual data consumption, because the old unit "satellite image" is blurred with data is stored in multi-dimensional structures not made transparent to the user. Often, it is even difficult for the cloud operator to calculate exact costs prior to the completed execution of a process. This leads to the difficult situation for both cloud operators that have to calculate costs upfront, and end-users that do not want to be negatively surprised by the final invoice for their processing request.

3.1. Billing and Quoting Standardization Efforts

The OGC has started the integration of quoting and billing services into the cloud processing architecture illustrated in figure 2. The goal is to complement service interfaces and defined resources with billing and quoting information. These allow a user to understand upfront what costs may occur for a given service call, and they allow execution platforms to identify the most cost-effective cloud platform for any given application execution request.

Quoting and Billing information has been added to the Execution Management Service (EMS) and the Application Deployment and Execution Service (ADES). Both services are implemented in a web-friendly way as a Resource Oriented Architecture (ROA) Web API that resembles the behavior of the current transactional OGC Web Processing Service v2.0 (this version, v.3.0, is not published yet by OGC). The API is described as an OpenAPI v3.0.0 that allows deploying and executing new processes by sending HTTP POST request to the "DeployProcess" operation or "Execute" operation endpoints. Following the same pattern, it allows posting similar requests against the Quota endpoint, which causes a JSON response with all quote related data. The sequence diagram in figure 3 illustrates the workflow. A user sends an HTTP POST request to provide an quasi-execution request to the EMS /quotation endpoint. The EMS now uses the same mechanism to obtain quotes from all cloud platforms that offer deployment and execution for the requested application. In case of a single application that is deployed and executed on a single cloud only, the EMS uses the approach to identify the most cost-efficient platform. In case of a workflow that includes multiple applications being executed in sequence, the EMS aggregates involved cloud platforms to generate a quote for the full request. Identification of the most cost-efficient execution is not straight forward in this case, as cost-efficiency can be considered a function of processing time and monetary costs involved. In all cases, a quote is returned to the user. The quote model is intentionally simple. In addition to some identification and description details, it only contains information about its creation and expiration date, currency and pricetag, and an optional processing time element. It further repeats all user-defined parameters for reference and optionally includes quotations for alternatives, e.g. at higher costs but reduced processing time or vice versa. These can for example include longer estimated processing times at reduced costs.

Quotation requests resemble execution requests, i.e. contain the same elements and values as if an execution would

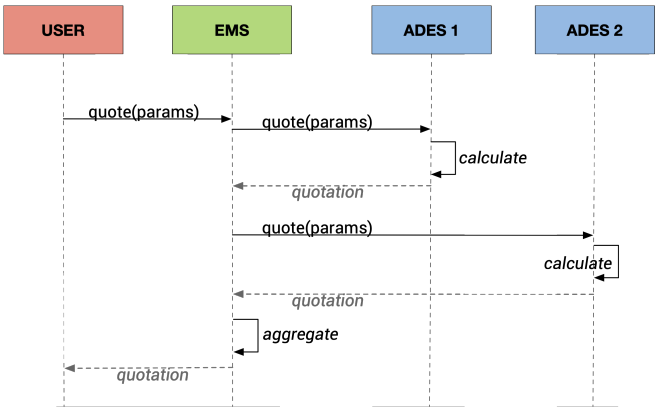


Fig. 3. Sequence diagram for quotes

have been requested. It is then up to the execution platform to obtain realistic quotes. This process is platform specific and can be implemented in multiple ways. We expect machine learning approaches to play an important role in this context. Platforms learn over time what costs are caused by specific requests and can generate better quotes for future requests by learning. The generation of a quote may not be based on calculations and experiences from prior requests exclusively, but may take business considerations into account. As cloud platforms are competing with each other, a platform might be motivated to advertise its performance by providing specifically low quotes for a limited period of time.

4. SECURITY

Reliable communication within business environments requires some level of security. This includes all public interfaces as well as data being secured during transport. As shown in 4, the system uses identity providers to retrieve access tokens that can be used in all future communication between the application consumer, EMS, and ADES. The authentication loop is required to handle multiple protocols to support existing, e.g. eduGAIN [4], as well as emerging identity federations. Once an authentication token has been received, all future communication is handled over HTTPS and handles authorization based on the provided access token. Full details on the security solution are provided in [5].

5. OUTLOOK

The current setup allows requesting quotations for service requests and supports sequential executed processing chains. The quotation and billing model are both intentionally simple to be applicable to a wide range of domains. Both models feature simple extension mechanisms to address specific needs in some communities. Further research is necessary to develop the described approach into a core and profile model.

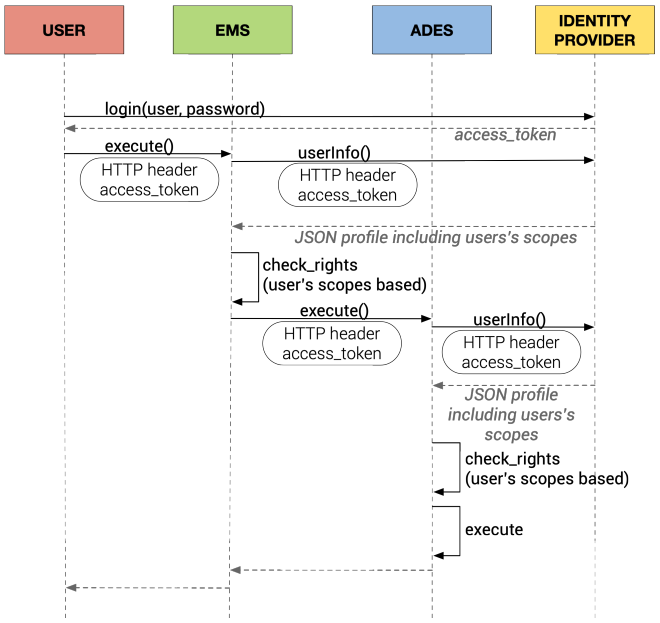


Fig. 4. Security overview

This model would then allow negotiating specific quote and billing profiles at runtime. Similar mechanisms can be used to define Service Level Agreements on the fly. Further research is required to add Linked Data principles that would allow making the billing and quotation model to leverage Semantic Web principles and capabilties. Processing results could receive identity, be linked to the original data and input parameter, and may serve as a valuable resource to others.

REFERENCES

[1] Simonis, I. (2018). Standardized Big Data Processing in Hybrid Clouds. In Proceedings of the 4th International Conference on Geographical Information Systems Theory, Applications and Management - Volume 1: GISTAM (pp. 20510). SciTePress. doi: 10.5220/0006681102050210

[2] OGC (2018). OGC Testbed-13: Application Deployment and Execution Service Engineering Report. OGC Document OGC 17-024. 17-024.html

[3] OGC (2018). OGC Testbed-13: Enterprise Platform Application Package Engineering Report. OGC Document OGC 17-023. 17-023.html

[4] GEANT (2018). eduGAIN Home. Website eduGAIN

[5] OGC (2019). OGC Testbed-14: Authorisation, Authentication, Billing Engineering Report. OGC Document OGC 18-057.

NEW LEGAL CHALLENGES FOR EARTH OBSERVATION DATA AND SERVICES?

Ingo Baumann & Gerhard Deiters

BHO Legal

ABSTRACT

The market for Earth Observation data and services is changing dramatically. Around the world, government agencies and commercial companies are investing in new Earth Observation satellites, sometimes including large constellations of small satellites. The rise of Earth Observation satellites goes along with a massive increase in available data. We see a shift from traditional data delivery to online, cloud-driven exploitation platforms as well as a shift from pure data delivery to more complex services. From the legal perspective, the current technology and market developments do mostly not raise previously unknown issues. However, the transposition to and use within the Earth Observation sector raises challenges for institutional and commercial stakeholders. In addition, some well-known legal issues in Earth Observation come into new perspectives.

***Index Terms*—** Earth observation, online platforms, cloud computing, legal issues

1. INTRODUCTION

The market for Earth Observation (EO) data and services is changing dramatically. According to Euroconsult, more than 400 EO satellites are expected to be launched during the next decade, generating $35.5 billion in manufacturing revenues.[1] The rise of EO satellites goes along with a massive increase in available data. The huge amount of data requires novel approaches for storage, archiving, and distribution.

The market for EO services is also showing positive developments, with average yearly growth rates of 10% or more. Value adding services traditionally are bespoke, delivered upon individual customer request in dedicated projects. However, the market shifts to online services with new business models, such as automated delivery of regular updates.[2] EO data and services are increasingly provided via the internet, through platforms with E-commerce type elements. To handle the massive data volumes, such platforms are backed with Cloud computing services. Such Cloud computing services are used for data storage and long-term archiving. More and more, for handling data access and distribution as well as for data analytics, processing, visualisation and value adding. Some elaborate platforms comprehensively offer all such functionalities. Users do not have to download the data to subsequently process them within their own laboratories and with their own value-adding tools; all these processes of the value chain can now be handled online.

The strong need for cloud computing services adds new stakeholders to the EO market. Over the last years, several large IT players such as Google, Amazon, T-Systems, ATOS, or SAP have developed special solutions for EO (or, more broadly, geospatial) data and services. Similar solutions are also being developed by major players in the geospatial market such as Hexagon or ESRI. These new players indicate that the EO market starts to move out of its relative isolation and becomes part of the broader geospatial, data, and information markets. Technology convergence goes hand in hand with market convergence.

From the legal perspective, the above-described developments generally do not raise previously unknown issues. Commercial E-commerce platforms are widespread and show similar approaches and functionalities as the specialized EO platforms. Specific legal issues of such platforms include E-commerce, consumer rights, and e-privacy (section 2), cloud computing (section 3), open source software (section 4), or liability for third-party content and hyperlinks (section 5). In addition, other well-known legal issues in EO come into new perspectives. This includes data policies, copyright, data licensing (section 6), personal data protection (section 7), standardization and interoperability (section 8), as well as warranty and liability for EO data (section 9). The following analyses of these legal issues will focus on EU and national laws of EU member states.

2. E-COMMERCE, CONSUMER RIGHTS, AND E-PRIVACY

EO online platforms are websites accessible on the Internet via web-browsers. Accordingly, they must include various information as prescribed by relevant EU and national laws. Where platforms enable the purchase of data, tools and services, the operators of such platforms also have to comply with applicable E-commerce laws.[3] Where EO products and services are also offered to consumers, distance-selling laws in EU member states by virtue of the Consumer Rights Directive may apply. The obligations imposed by the previously mentioned legal frameworks apply both to the operators of the EO platforms and to any third party offering EO data, tools, and services on such platform. This is especially relevant for EO platforms with marketplace functionalities, where commercial providers can set up their own web-stores on the platform for are value-adding products and services directly to their customers.

doi:10.2760/848593

3. CLOUD COMPUTING

Cloud computing supports the shift from simple data download towards fully-fledged E-commerce type platforms with virtual workspaces, where data can be uploaded, analysed, merged, and processed with the help of software tools. The applicable EU and national legal frameworks for Cloud computing are complex. Cloud computing raises numerous legal challenges, including data ownership, data security, data protection, storage location, portability, performance obligations, warranty, and risk of provider lock-in. Concerns regarding these issues are still hampering the wide use of cloud computing in Europe, especially by public authorities. Contracts for cloud computing services need to cover these issues. While best practices exist, and certain harmonization work is undertaken, terms and conditions of Cloud computing services contracts strongly differ.[4] Generally, contractual obligations must be flexible enough to cope with technology advances, emerging threats, and changing requirements.

4. OPEN SOURCE SOFTWARE

Open source software (OSS) is increasingly employed in EO for processing and value adding of EO data, in order to save development and maintenance costs for the individual users, avoids lock-in situations with the original creator(s), facilitates rapid evolution, and encourages reuse.[5] However, the use of OSS may also involve certain drawbacks, mostly caused by a lack of awareness by both developers and users of the characteristics of open source, applicable license conditions, and resulting legal implications.

Many users of OSS assume that such software can be used without restrictions. However, OSS is copyright protected and the term "open" itself does not have the meaning of "unconditional." OSS is subject to license terms, which have to be accepted by the user. Without such acceptance, the use of OSS is prohibited. License conditions typically impose so-called copyleft or share-alike regimes and stipulate exclusions of warranty and/or limitations of liability. The various open source licenses however differ from each other in many ways. Some open source licenses limit free use and, for example, exclude the right to amend the software for commercial purposes, while other open source licenses do not limit the use of the software at all. Some licenses have a strict copyleft regime (allowing the user only to distribute derivative works under the very same license), while other licenses use the "share-alike" approach (allowing the distribution also under a "compatible" license).

5. LIABILITY FOR THIRD PARTY CONTENT AND HYPERLINKS

EO online platforms with marketplace functionalities provide content of the different service providers present on the platforms. The question thus arises whether and to what extent the operators of such platforms – so called intermediaries – can be held liable for the third-party content. As a rule, intermediaries are exempt from liability for third party content, have no general obligation to monitor such content, and have no general obligation to seek circumstances indicating illegal activities. In practice, so-called "notice and take-down" procedures are applied, meaning that operators of an online marketplace have to remove infringing offers upon receipt of corresponding notifications by right owners.

As with every website, EO online platforms may provide hyperlinks, some of which are to third party online services and some of which are for advertisements. The responsibility of the platform operator for such hyperlinks is of particular concern for matters such as third-party intellectual property rights (IPR) and personal data protection. According to recent jurisprudence of the ECJ,[6] the operator of a website who does not pursue a profit can generally not be held responsible if the linked website contains works published without the consent of the rightsholder, except when it is established that the operator knew or ought to have known that the hyperlink he posted provides access to such illegal content. When a website is operated for profit, the operator that posts a hyperlink is, according to the ECJ, under the obligation to carry out the necessary checks to ensure that the hyperlinks do not provide access to "illegal" content. Up to now, the scope and content of this obligation is rather ambiguous.

6. DATA POLICIES, COPYRIGHT AND LICENSES

There is a general trend for public EO data to be provided under open data policies. These policies, however, differ significantly in approach, definitions, scope, and content. This does not cause that many difficulties, as long as the user is accessing one particular data set from one supplier. On EO online platforms, users however combine datasets from multiple suppliers – each of which may apply its own data policies.

Data policies generally establish that the mission operator (or data owner, if separate) retains ownership of the relevant data. Furthermore, data policies usually state that the mission owner holds the relevant intellectual property rights, namely copyright. Copyright laws generally require intellectual creation with a minimum level of originality. Accordingly, automatically generated data such as raw data from EO satellites are generally not protected under copyright laws. In addition, processing and value adding steps are now increasingly run by automated software applications. The trend of automatization will accelerate with cloud-based platforms providing virtual workspaces. Whether or not processed data and final products are still copyright protected under the applicable law therefore requires individual evaluation, leaving significant uncertainties for the data owners.[7]

Licenses are the instruments for implementing applicable data policies. As copyright protection is increasingly questioned also for processed data and final products,

doi:10.2760/848593

effective license management through the whole distribution chain becomes even more important to protect the ownership rights of the mission operator.

From a user perspective, understanding, accepting, and observing multiple EO data licenses with divergent provisions is generally challenging. The existence of multiple data licenses then becomes a critical issue, when data sets from numerous sources are to be merged and processed towards final products or services.

7. PERSONAL DATA PROTECTION

Objectively, EO data per se are not very sensitive regarding personal data. The most advanced commercial imaging sensors provide a spatial resolution of around 30 cm (WorldView-3 image data from DigitalGlobe). With such resolution, it is not possible to directly identify an individual person from space. Due to the high speed of low-orbiting satellites, a specific scene can only be captured for a short time – it takes then at least one orbit until revisit. Therefore, it is not (yet) possible to "track" movements of individuals in real-time. However, high-resolution optical data, depending on their spatial resolution, may have the same quality as aerial photography and therefore may raise respective privacy issues. In addition, EO data may be combined with other data sets and then may raise privacy concerns, even if the raw or pre-processed data itself do not.[8]

Even if EO data can be considered personal data, the lawfulness of its processing in general depends on the purpose of such processing. For example, while the processing of personal data for the purpose of contract management can be considered lawful, the processing of the very same data for the purposes of personalised advertising may be unlawful (without the data subject's prior consent). Accordingly, the lawfulness of the processing of personal data in the context of EO services has to be reviewed on a case-by-case basis. For EO online platforms, platform operators need to comply with the applicable personal data protection laws, when processing user data for registration, identity checks, and user account management.

8. STANDARDIZATION AND INTEROPERABILITY

Data do not exist in a vacuum. To be useful, data must be accompanied by context on how they are generated, captured, calibrated, processed, and validated along with other information that enables their proper interpretation and use. Users require a solid foundation of such contextual information in order to verify data validity, accuracy, and reliability. This is vital for the creation of accurate value adding products and services. Appropriate technical and normative approaches are required to identify, capture, and track all necessary details to this end. In practical terms, this means that the metadata and further contextual information describing all the steps in the chain have to be made available in transparent way, so that users can easily check them. Compliance with recognized standards and interoperability across the different data sources play important roles in this respect. This will become even more important for EO products and services merged from several data sources. Today, there is however only partial convergence on standard ways of holding and transferring EO and other geospatial data and information.

The variety of standards for EO data provide a technical challenge to users. But there are also legal implications, namely regarding warranty and liability of the data provider. In case users do not have access to metadata and necessary contextual information, they may come to wrong expectations or interpretations regarding the validity or accuracy of the data, leading to wrongful conclusions and resulting actions. The same applies, when the metadata and information is incomplete, outdated, or inaccurate. Users (or third parties) may base claims on the data provider´s breach of professional duty of care.

9. WARRANTY AND LIABILITY

The provision of EO data, as with geospatial data more generally, is faced with uncertainties regarding warranty and liability risks. So far, liability claims by users or third parties have been rare. The few known cases around the world mostly relate to aeronautical or maritime charts or to traditional maps and do not create sufficient precedence. However, a broader perspective into the geospatial sector shows the relevance of the matter; advances in technologies for data dissemination and new business models indicate increasing risks for the future.

Far-reaching limitations or even exclusions of warranty and liability for EO data are still standard, both for public and commercial data. At best, the data provider commits to repair or replace defective data sets upon notice by the user. Further, license conditions aim to protect from liability claims for damage arising in relation to the use of the data, sometimes including infringement of third party IPR.[9]

As to now, users seem to accept the exclusions/limitations of warranty and liability in data licenses. However, three considerations justify prudence regarding future developments. First, there is a general trend in the EO sector to move from licenses for the provision of individual data towards more comprehensive EO services contracts. While limitations of warranty and liability may be regarded as appropriate for individual data (especially where provided free of charge), this is not the case for the provision of EO services. Second, the EO sector becomes more and more part of the broader geospatial and ICT sectors. As convergence continues, the attitude and practice regarding warranty and liability may change. Finally, far reaching exclusions of warranty and liability may hinder the growth of the commercial EO market. Customers will more and more expect warranty for products delivered, as well as performance obligations for services provided. Value adding providers however have often no possibility of taking

doi:10.2760/848593

recourse against the input data providers due to respective license terms.

10. CONCLUSIONS

The evolution of the EO market changes the legal framework considerably. Following technical convergence and the growing advent of online services, EO becomes part of the larger data and digital economy. Consequently, relevant legal issues fall under IT law much more than under international or national space law. The issues are not necessarily new; the challenge is how to apply existing regulations and best practices from the IT to the EO market.
In addition, it might be required to rethink some of the heritage legal issues in EO. As described, the growing automatization in data acquisition, processing and value adding reduces the availability of copyright protection under the applicable law. However, individual licenses hinder the use and merger of data from multiple sources. Data interoperability requires more harmonized licensing conditions. The growing combination of EO data with other data sets may raise privacy concerns. The shift from data supply to service contracts will likely change the approach regarding limitations of warranty and liability. Public as well as commercial satellite operators, data distributors, platform providers, and value adding service companies all have to find reliable answers to these issues.

11. REFERENCES

[1] Euroconsult, Earth Observation Manufacturing, Data Markets Continue Expansion, 2016, under http://www.euroconsult-ec.com/15_September_2016.

[2] EARSC, Establishing A European Marketplace for EO Services, 2017, under https://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0ahUKEwi97uOqrc3XAhVDJFAKHblbAM8QFgg5MAI&url=http%3A%2F%2Fearsc.org%2Ffile_download%2F418%2FMAEOS%2Bfinal%2Bpresentation%2B20170125.pdf&usg=AOvVaw1nyd0Q4Jo66xgZsPbuoGfg.

[3] For a high level overview on the legal regulations for e-commerce under EU law, see: European Commission, Legal regulations for e-commerce, under https://ec.europa.eu/growth/sectors/tourism/business-portal/understanding-legislation/legal-regulations-e-commerce_en#mandatory

[4] For a comparative analyses on cloud computing contracts in Europe, see: European Commission, Comparative Study on cloud computing contracts, 2015, under https://publications.europa.eu/en/publication-detail/-/publication/40148ba1-1784-4d1a-bb64-334ac3df22c7

[5] The ESA Open Source Policy summarizing benefits and challenges is available under, https://essr.esa.int/esa-open-source-policy

[6] European Court of Justice, Case C-160/15 *GS Media BV v Sanoma Media Netherlands BV and Others* [2016], ECLI:EU:C:2016:644.

[7] For an overview on intellectual property rights associated with satellite images, see: Ito, Legal Aspects of Satellite Remote Sensing, 2011, 213 et seq.

[8] ULD, Datenschutzrechtliche Rahmenbedingungen für die Bereitstellung von Geodaten für die Wirtschaft, Gutachten im Auftrag der GIW Kommission, 22 September 2008.

[9] For an overview on the current licensing practice, see: United Nations Committee of Experts on Global Geospatial Information Management, Compendium on Licensing of Geospatial Information, 2018, under http://ggim.un.org/meetings/GGIM-committee/8th-Session/documents/E-C20-2018-9-Add_2%20Legal_and_Policy_Frameworks_rev.pdf

TOWARDS ECOLOGICAL STEWARDSHIP BASED ON SPATIALLY EXPLICIT ECOSYSTEM ACCOUNTS

Jean-Louis Weber

Research Associate, Ecole Normale Supérieure de Lyon, IXXI, Institute of Complex Systems
Former Special Adviser to the European Environment Agency on Environmental Accounting

ABSTRACT

The purpose of Ecosystem Natural Capital Accounts (ENCA) is to measure the impact of economic activities on ecosystems structures and functions and the sustainability of resource/services that they provide with the purpose of completing current accounting standards where ecosystem degradation recording is absent. Ecosystem accounts supplement "carbon" accounting with land, biodiversity and water accounts. They are tools needed for implementing ecosystem stewardship policies and practices. The spectacular progress of Earth observation by satellites, of in situ monitoring, of GIS and data analysis tools, of cloud computing and the access to Big Data, make it possible to implement swiftly the multi-scale and multi-actors information system needed. Whereas implementation by governments and business will follow various processes, ENCA can be undertaken NOW at the global scale with existing capacities. This first step would put biodiversity/ecosystem policies on par with climate change.

Index Terms — Ecosystem Accounting, Biodiversity, Ecological Value, Intermediation Platforms

1. INTRODUCTION

Our neglect of the natural systems that directly and indirectly provide for our livelihoods results in the depletion of material resources and in the alteration of the functions that permit their renewal, and consequently, our life on Earth. This finding is renewed year after year, through warnings from the IPCC on climate, from the Global Footprint Network on the inexorable use of natural resources or from the WWF and IUCN on the collapse of biodiversity. But few solutions are drawn from these observations that match the magnitude of the ecosystem stewardship challenge that we face.

The problem results more from the absence of an agreed metric for ecosystem and biodiversity than from a misunderstanding of the issue itself. Whereas economic capital depreciation is recorded as a cost and included in the value of commodities, no such recording is done for the ecosystem capital, which allows short-term benefits for producers and consumers. Indeed, ecosystem degradation results from unpaid externalities by economic agents behaving as free riders of the common good. There are growing concerns that these benefits are offset in the longer term by ecological losses as well as by financial losses. Tackling with this matter requires a standard metric for measuring ecosystem degradation, fostering integration of measurements into accounting reports of all actors and verifying the compliance, from governments and companies. To support this purpose, the Convention on Biological Diversity has published in 2014 a technical report on Ecosystem Natural Capital Accounting – A Quick Stat Package (ENCA-QSP) [1] which includes an integrated framework and a measurement unit for representing intrinsic ecological values. ENCA builds on decades of reflections but tests were constrained by gaps in data or technological shortcomings, a situation which has radically changed in the last decade.

2. ACCOUNTING FOR ECOLOGICAL VALUES

Book-keeping is the way to keep track of the multitude of facts with the joint purpose of control and performance assessment. Accounts record transactions and assets in an exhaustive way in order to calculate meaningful balancing items such as Operating Surplus, Profits and Losses, Assets Net Worth, or Value Added and Gross Domestic Product. Accounts are used for internal planning, but also as communication tools (with, in particular, shareholders, tax department, banks, auditors and, regarding the National Accounts, with the Parliament and international organizations such as the IMF, the World Bank and, for Europe, the European Commission…). Particulaly for the latter purposes, accounting standards are needed. For businesses, there are internationally agreed financial accounting standards. For countries, there is the UN System of National Accounts.

Important to note at that stage is that the meaningfulness and fairness of accounts relies on the completeness of accounting books records. Any missing (positive or negative) entries can result in misleading or even fake results, which is frequently a matter for arbitration by courts. A particularly important issue regarding completeness of accounts relates to the deferred consumption of capital goods which is spread over time. This is recorded in financial accounts as capital depreciation and in National Accounts as its equivalent of Consumption of Fixed Capital. It is deducted from companies' revenues and from countries' GDP to calculate income.

However, no such recording is done for the consumption of ecosystem capital by economic activities when they degrade the condition of ecosystems. This incompleteness of accounting standards generates doubts regarding the relevance of conventional accounts for decision making, starting with the recurrent criticism of the Gross Domestic Product aggregate. Implementing economic-environmental accounts therefore has been undertaken as a request of the 1992 Rio Earth Summit, firstly as a statistical project aimed at completing the national accounts. More recently, initiatives have multiplied from the business side considering the financial risks taken by companies from activities that impact the Natural Capital, and by the financial institutions which are funding them.

3. MULTIPLE AND DIVERSIFIED DATA AND KNOWLEDGE FOR A COMMON PURPOSE

Because of the importance of global markets and their interaction with global issues like climate change, biodiversity collapse, and food security, ecosystem capital degradation cannot be addressed only by public policies, but should involve business (from multinational companies to farmers) and the citizen as well. Integrating the environmental factor in decision making requires conveying knowledge to decision makers in a format which matches their usual tools, considering the appropriate scales. Ecosystem natural capital accounts aim at being implemented at different scales from the global to the national, local and business levels.

Ecosystem accounts integration must face the issue of complexity of the natural world, which means that aggregated accounts cannot be simply produced by adding up elementary accounts in the same way as is done in the economic sphere. The production and update of the information system should go beyond the usual vision of top-down and bottom-up processes and encompass the many bottom-bottom data and knowledge flows generated by the many practices. This is made possible by the development of intermediation platforms (Figure 1) which give access to large and frequently updated datasets (from Earth observation by satellite and in situ monitoring), high resolution data specific to particular contexts, geo-referenced statistics and advanced processing tools for data extraction and modelling.

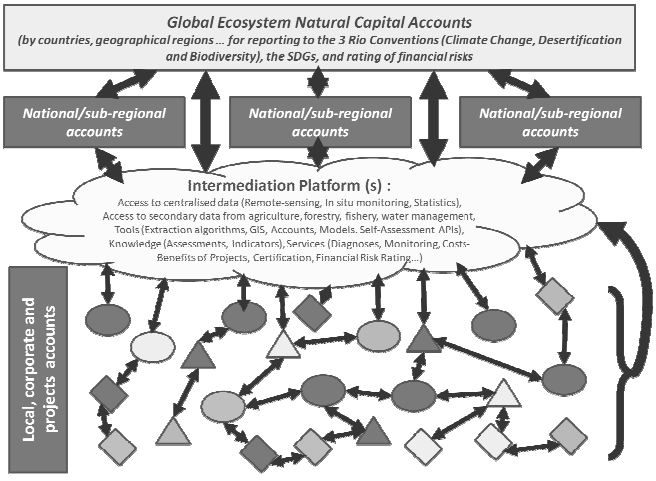


Fig. 1 Intermediation platform(s) for ecosystem accounting

Data platforms and "ecosystems" involving people, enterprises and institutions can contribute to the information system needed for natural ecosystems conservation. "*Platforms create ecosystems in which both users and economic players take a role*" (Grumbach, 2014) [2]. "*…platforms also hold fantastic potential for meeting one of society's most important challenges, the more frugal use of resources*." [2]

At the Global scale, accounts are necessary to assess development's sustainability, the regions and countries at risk, and the needs for international action. Accounts can be used to monitor the efficiency of existing policies and help to avoid distorting effects or "leakages" due to predatory behaviors. One important aspect of accounts monitoring relates to the risk associated with sovereign and private debts, which are increased by unsustainable trends in environment and natural resource management. These concerns have been expressed by the UNEP Finance Initiative and the OECD, as well as by rating agencies and financial institutions which have also started integrating "carbon" credits and debits in their assessments. There are initiatives in both the international development and financial sector realms in search of an equivalent metric for assessing risks associated with ecosystems and biodiversity.

Similar concerns are also present for land and urban planning and for nature conservation at the national and subnational levels and with the drive towards more sustainable practices in energy use, transport systems and agriculture.

The micro-level of ecological management (local governments, enterprises, projects assessments) requires very specific detailed knowledge for action and is, at the same time, embedded in the national and international macro-levels considering both the economy and the ecosystem. In a symmetric way, the micro-level generates rich knowledge and wealth of data which can be now accessed via the algorithmic extraction facilities provided by the Big Data and the use of AI tools. Ecosystem natural capital accounting protocols can be adapted to each level, delivering outputs founded on the same conceptual basis and with a consistent approach to measurement of ecosystem degradation (or enhancement).

4. THE DATA SERVICES NEEDED FOR ECOSYSTEM ACCOUNTING

Accounts, corporate or national, established in money or in physical units, are summaries aimed at measuring performances, communicating results and supporting decision making. Accounts regularity and fairness depends to a large extent on records' completeness. Accounting standards frame the rules needed for communication between the many private and public actors involved, and guarantee comparability of results.

Ecosystem natural capital accounts (ENCA) do not summarize preexisting records done by individual actors but are a model for integrating information in order to calculate ecosystem degradation or enhancement. They are firstly established by ecosystem units for which the assessment is the most meaningful. In a second step they are disaggregated by economic sectors and downscaled by economic units in order to measure liability and allow for estimating the cost of ecosystem capital depreciation. We can note at this stage that direct measurements by sectors or units are possible but only regarding the consumption of materials (extraction and returns after use, including after combustion) and regarding the depletion of stocks. Degradation of ecosystem functions and resilience requires, in addition, scientific and contextual information which goes beyond what can be observed by an isolated economic unit. For this reason, whereas climate change reporting has moved from a sector-based to a land-based approach, ecosystem accounts are "land-based" from the beginning. "Land-based" should be understood here as based on spatial statistical units, which are mapped as land cover units in the case of terrestrial and coastal ecosystems and sea zones for the oceans (the atmosphere being connected to land by precipitations, evapotranspiration and emissions).

doi:10.2760/848593

It is therefore not a surprise that progress in ecosystem accounting is being made in parallel to that of earth observation by satellite and of geo-referenced in situ monitoring systems. Internet and the development of Big Data bring now a new wealth of opportunities to access micro-data and use AI algorithms to extract data needed for ecosystem accounting at the various scales. These technological advancement are particularly relevant for ecosystem capital accounting, which combines measurements of quantities of natural resources (in hectares, tons of carbon, m^3 of water) with diagnosis of ecosystem health. Knowledge transfers are accelerating from the fast development of Big Data and AI applications in other realms, such as human health diagnoses.

This progress should be seen as an enhancement, not as a replacement to other, more traditional, data sources. It is beyond the scope of this paper to address cautions and critics of big data that are commonly formulated by statisticians and other scientists. They relate to incompleteness, instability, and insufficient validation, uncertainty of access in particular in the case of private databases and to the dependency of algorithmic extraction results from the contextual elements of the observed variables. The rapid progress of development of big data methods is quickly making such discussion obsolete. The question that remains, however, is what are the crucial policy questions and outcomes that use of big data will influence?

Big data does not provide specific social objectives for policy, big data is a tool for providing evidence to help to monitor and achieve common objectives. For Ecosystem Natural Capital Accounting, the expected contribution is the measurement of ecosystem degradation (or enhancement) with an agreed standard metrics and with a view for integrating this factor into accounts and accounts-based decision making at all relevant scales. Big data does not help to determine the structure of accounts, which instead must be based on social consensus on targets and on the meaning of ecological value. Another contribution of ENCA is the systematic definition and mapping of data of spatial units for which assessments should be done.

Beyond the difficulties of assessing one's environment from the micro perspective already mentioned, another important point to consider is cost-efficiency. Not every situation demands high resolution real-time assessment. For a very large part of the planet, remote sensing is fit for delivering the broad picture, detecting hot-spots and, jointly through regular monitoring (e.g. meteo, hydrology) and statistics (e.g. population, agriculture), for describing the context for appropriate understanding and action.

As with global warming mitigation, ecosystem sustainable management requires a commitment of the whole international community. In the process of reaching this goal, equity and fairness are important factors, which means that accounts are needed for verification and for assessing the liabilities of others, with their consequences in terms of production costs and financial risks.

Models are based on assumptions that reflect, to some extent, the vision of the modeler. Statistics can be biased by a surveyors' and a respondents' understanding of the questions considering the purpose of the survey. Sampling patterns themselves can be biased due to short term or local priorities. Algorithmic extraction generates secondary data and services from primary data sets which are often, in principle, targeted to different purposes, noting the multifaceted social demands for new services required to run new business models. In all these cases, the combination of diverse data sources is the best response, The neutrality of pixelated images collected from satellites, the exhaustiveness and repetitiveness of products and the capacity to re-process long series of data, all contribute to key and growing role of earth observation by satellite to the development of a measurement frame with consistency across systems.

5. ECONOMIC ACTORS, PEOPLE, INSTITUTIONS, PLATFORMS

Sustainable ecological management of the planet is not the mere policy target of public institutions, but a goal which requires the involvement of everyone. It means that behaviors of free-riders of the common good must change, and social systems must migrate from short term views and ignorance of consequences to ecolocial systems. We should also consider the extreme vulnerability of populations exposed to natural disasters and who survive precariously at the limit of ecosystem degradation. Public policies are essential, but cannot face the task alone. Public policies can promote a paradigm change and design regulations and take measures, but in every case they need active participation of the society for the transition to happen.

In particular, private enterprises need to adapt their business models to integrating ecosystem depreciation in their costs [3]. For the enterprises, this is within their own self-interest so far as it can help reduce financial risks and associated costs in the longer term and so far as it offers new business opportunities. Without engaging into fragile forward looking, it seems obvious that the implementation of ecosystem accounting standards will require important development of expertise regarding sustainable practices for scientific monitoring and assessment as well as ecological balance-sheet bookkeeping and the derived financial rating activities. Beyond scientific consulting, providing access to relevant information on ecosystem degradation embedded in commodities is a business which is critical and will certainly develop alongside the many applications informing on healthy products. Last but not least, ecosystem protection and restoration will generate investments and induce technological change which will stimulate business development.

Regarding the public institutions, statistical offices, environmental agencies, meteorological offices, and, more generally, research and in particular space agencies are interested and in a position to play a crucial role for ecosystem accounting.

The national statistical offices have played an important role in early developments of environmental accounting in the 1970s-80s, with the purpose of broadening the scope of national accounts and macro-economic policy tools to take the environment into account (Weber, 2018) [4] . The main outcome of the process is the UN System is the System of environmental-economic accounting 2012: Central framework (SEEA CF, 2014) [5] and the SEEA Experimental Ecosystem Accounting (SEEA EEA, 2014) [6]. Offical statistical offices are important stakeholders but few of them are presently in a position to take on ecosystem accounting, with exceptions, including Brazil and Mexico where statistics and geography are merged in a single institute. International statistical

doi:10.2760/848593

institutions such as UN Statistical Division, UN Environment and Eurostat broadly have limited data processing capacity for ecosystem issues, an exception being FAO regarding agriculture, fishery, forestry and water statistics which broadly cover ecosystem issues.

The environmental agencies have considerably developed monitoring networks at the national scale. At the international and European levels, their activity is mainly to collect data from countries on the one hand and to carry out analysis and assessments from these.

Meteorological offices have played a key role in the Climate change process and have tremendously developed their observation tools and models, which makes them essential in the ecosystem accounting project regarding the variables that they monitor.

The research sector is of course a key player as ecosystem accounts include diagnoses of ecosystem health and resilience based on expert judgments which go beyond the datasets used for accounting.

Space agencies have a unique position where scientific knowledge meets industrial culture. They have a key role to play in the operational development of ecosystem accounting as they master new earth observation data sources and now have robust experience with managing large datasets and delivering services with space and time consistency. In Europe, major initiatives are the ESA Climate Change Initiative and the experimentation of Mission and Thematic Exploitation Platforms (Coastal, Forestry, Food Security, Geohazards, Hydrology, Polar and Urban) linked to the development of Copernicus thematic (Europe and Global Land services, Atmosphere, Climate Change, Marine and Emergency) and Data and Information Access Services (DIAS) platforms. This core system is part of a broader (data) ecosystem within Europe and interacts with research projects, either institutional (e.g. at the EEA and the JRC) or academic (for example, the SMURBS research, SMart URBan Solutions), and international (through GEO/GEOSS and bilateral cooperation).

6. TOWARDS A GLOBAL ENCA (GLOBENCA)

Ecosystem Natural Capital Accounting (ENCA) requires the broad and fast development of platforms and services. Regarding the national to local and enterprise levels, these platforms will play important role to connect scientific to operational knowledge. The integration of the ecosystem degradation variable into accounts will have effects on the economic activity in terms of resource use, estimation of ecosystem depreciation, integration of financial risks, and hence on quantities and prices of traded goods and services. Hence, issues of comparability of the measurements, methodological stability over time and verification are key for social acceptance of the new standard. This is a domain where earth observation by satellite has a key role to play.

The role of Europe in this progress is considerable, on the policy ground as well as through ambitious programmes such as ENVISAT, the Sentinel satellites or the CCI at ESA and the related COPERNICUS services.

Concerns on biodiversity collapse and ecosystem degradation are increasing. They cannot be addressed using only global models and economic sector data. Instead, biodiversity/ecosystem assessment requires spatially distributed data from the very beginning. The experience, processing and modeling capacity and the data accumulated makes it possible to produce a first generation of ecosystem natural capital accounts for the Planet at a scale meaningful to set the frame and propose a new paradigm for discussion at the next CBD COP 15 in Beijing, in the last quarter of 2020. The Beijing CBD COP 15 will take stock of the progress towards the Aichi Targets of 2010, including Target 2 of incorporation of biodiversity values into management systems including the national accounts. A first generation of simplified accounts could be established for 2 or 3 dates by countries, socio-ecological landscape units, and river catchments. They would be computed using data assimilated into the $1km^2$ grid according to the ENCA Quick Start Package [1] methodology.

This application would not be simply one more ecological assesment case study, but as the first step in a process which will progressively involve countries and economic actors to multi-scale global ecological accounting and progress towards ecosystem sustainable stewardship. It could put biodiversity/ecosystem challenges on par with climate change.

REFERENCES

[1] J.-L. Weber (2014), “Ecosystem Natural Capital Accounts: A Quick Start Package”, Technical Series No. 77, Secretariat of the Convention on Biological Diversity, Montreal, 248 pages. ISBN 92-9225-538-X. www.cbd.int/doc/publications/cbd-ts-77-en.pdf

[2] S. Grumbach (2014), “Intermediation Platforms, an Economic Revolution”, “ERCIM News 99”, https://ercim-news.ercim.eu/en99/challenges-for-icst/intermediation-platforms-an-economic-revolution

[3] J. Rambaud and J. Richard (2015), “The ‘‘Triple Depreciation Line’’ instead of the ‘‘Triple Bottom Line’’: Towards a genuine integrated reporting”, Critical Perspectives on Accounting, 33(2015) 92-116, Elsevier, https://doi.org/10.1016/j.cpa.2015.01.012

[4] J.-L. Weber (2018), “Environmental Accounting”, Oxford Research Encyclopedia of Environmental Science, http://environmentalscience.oxfordre.com/view/10.1093/acrefore/9780199389414.001.0001/acrefore-9780199389414-e-105

[5] SEEA 2012 - CF (2014) “System of Environmental-Economic Accounting 2012: Central Framework - final, official publication”, UNSTAT, https://unstats.un.org/unsd/envaccounting/pubs.asp

[6] SEEA 2012 - EEA (2013) “System of Environmental-Economic Accounting 2012: Experimental Ecosystem Accounting - white cover publication” , UNSTAT https://unstats.un.org/unsd/envaccounting/pubs.asp

SENTINEL-2 SEMANTIC DATA & INFORMATION CUBE AUSTRIA

Dirk Tiede[1], Martin Sudmanns[1], Hannah Augustin[1], Stefan Lang[1] and Andrea Baraldi[1,2]

[1]Department of Geoinformatics – Z_GIS, University of Salzburg, Schillerstr. 30, 5020 Salzburg, Austria; dirk.tiede@sbg.ac.at
[2] Italian Space Agency (ASI), Via del Politecnico, 00133 Rome RM, Italy

ABSTRACT

Data cubes seem to be a promising methodological development to deal with the big EO data challenge. The *Sentinel-2 Semantic Data Cube Austria* (Sen2Cube.at) goes beyond the current state-of-the-art and aims to build an Austrian data & information cube. The main goal is to show that semantic content-based image and information retrieval is possible in big EO image databases, allowing users to query and analyse EO data on a higher semantic level (i.e. based on at least basic land cover units and encoded ontologies). This includes: (1) fully automatic semantic enrichment of Sentinel-2 images up to land cover types ready for semantic content-based analysis; (2) the use of suitable database technologies to develop spatio-temporal modelling and querying techniques using encoded ontologies to decrease the complexity of queries for user interaction; (3) a Web interface for human-like queries based on semantic models of the spatio-temporal 4D physical-world domain; and (4) the demonstration of the potential of the generic data & information cube in future service developments based on different service types.

Index Terms— Earth observation (EO) images, big Earth data, semantic querying, information cube, semantic enrichment, content-based image retrieval, analysis-ready-data, remote sensing, Sentinel-2

1. INTRODUCTION

The Sentinel-2 satellites are specifically designed to deliver high-resolution (HR) imagery as a key information source for the European Copernicus programme. The Sentinel constellation collects significantly more data than any comparable existing or past initiative. Each of the two Copernicus Sentinel-2 satellites produces more than 1.7 Terabyte (Tb) of data per day, depending on the processing level. This translates to the acquisition of several hundred scenes every day. At the time of writing, nearly 6 million products are available for download, cumulating in a total volume of 3.1 Petabytes [1]. The general information potential of Sentinel-2 data is enormous. However, the greatest challenge of Earth observation (EO) *big data* analytics is to produce timely, operational and comprehensive EO value-adding information products and services from available EO *big data* systematically and automatically (without human machine-interaction).

While the sheer amount of data is the main challenge for most big data domains, satellite data requires additional interpretation or conversion into information layers in order to unlock its potential as a source of relevant multi-temporal, geographic information. Addressing this challenge requires intelligent solutions for machine-based data interpretation and efficient cloud infrastructure for fast access and processing of data.

Several big data initiatives in the EO domain have been established in Europe and internationally. These include funded and institutional projects (e.g. Copernicus Data and Information Access Services (DIAS)), private (e.g. Google Earth Engine (GEE), Amazon Web-service), and national initiatives (e.g. Code:DE, PEPS). Next to analytically oriented solutions such as GEE [2], data cube-based systems currently seem to be the most prevalent way to provide processing capabilities together with analysis-ready-data (ARD). Particularly relevant examples are EarthServer-2 and Digital Earth Australia (DEA), which includes the emerging Open Data Cube (ODC) initiative [3], [4].

EO data cube technology is closely linked to the term ARD, which usually defines a minimal set of steps for pre-processing to surface reflectance as a pre-requisite for automated EO data processing through time (see e.g. ARD description from CEOS CARD4L [5]). The aim of ARD is the provision of data in a form that can be directly used as input for analysis without requiring the downstream sector to accomplish any EO format transformation or data pre-processing step.

Radiometric calibration (Cal) of dimensionless digital numbers (DNs) into surface reflectance values, corrected for atmospheric, topographic and adjacency effects, is inherently ill-posed (data alone are not sufficient to uniquely solve the problem). A generic, application-independent scene classification map (SCM) better poses calibration on a class-conditional (i.e. masked, stratified) basis. Applications such as compositing/mosaicking or time-series analysis benefit from radiometrically calibrated EO image stacked with its data-derived SCM by using information contained in the images rather than using reiterative algorithms for every new analysis task. For example, systematic SCM generation enables semantic content-based image retrieval (SCBIR), where high-level (symbolic, semantic) information contained in each EO image can guide an intuitive image retrieval process. In practice, SCBIR is complementary to traditional text-based image queries limited to pre-defined metadata

doi:10.2760/848593

tags, such as scene-wide statistics (quality layers), like in current EO data access portals.

Considering these points of reference, our approach goes beyond state-of-the-art ARD and the use of data cubes solely as data storage by incorporating semantic enrichment (i.e. initial, data-driven information extraction). We aim for a generic, semantic EO data cube concept driven by automated integration of optical EO data and automatic semantic enrichment in contrast to application-driven enrichment (e.g. forest application, crop cycles, specific composites etc.). This generic concept enables diverse queries and analysis possibilities directly within data cubes, including semantic queries for replicable extraction of EO-based indicators from big EO data. Table 1 shows a comparison of current solutions and our approach.

2. METHODS

The major challenge for big EO data is systematically producing relevant information from EO scenes available in archives, which remains unsolved by the EO community. Current systems are limited to specific user interactions (e.g. download), characterised by low re-usability of tools and algorithms [6] or provide only limited image understanding capabilities [7]. We have characterised three means to approach this challenge: (1) available and accessible ARD and information in compliance with the foundational principles of Findability, Accessibility, Interoperability, and Reusability (FAIR) applied to data, information products and information processes [8]; (2) suitable and reliable (accurate, efficient, robust) tools and services; and (3) comprehensive user interaction and portrayal of user-defined results.

Based on existing proof-of-concept implementations [7], [9], [10], Sen2Cube.at will evaluate and scale automated semantic enrichment of free and open Sentinel-2 data up to a big EO image database covering Austria, building an Austrian data & information cube. This project follows a novel and entirely different approach to accessing big Earth observation (EO) image databases, allowing SCBIR and information retrieval through time. The spatial-temporal query capabilities facilitate searches directly related to the scene content or content dynamics, such as changes to any primary land cover category of interest (e.g. water bodies) in a user-specified area-of-interest (AOI) through time. The overarching methodological steps encompass:

- automatic semantic enrichment
- selection of data cube technology
- development of a Web-based inference engine
- demo services and evaluation

Perhaps the most crucial step is employing reliable, repeatable, **automatic semantic enrichment** of Sentinel-2 images. This is a pre-requisite for any SCBIR system, and is the main innovative difference to any other EO data cube to date. A combination of pre-classifications using SIAM (Satellite Image Automatic Mapper; [11], [12]) together with object delineation and texture information aims to reach the goal of a basic, fully automatic classification for Sentinel-2 images inferior or up to basic land cover classes (e.g. up to FAO LCCS Phase 1 - Dichotomous Phase). Interoperable SCMs whose thematic map legend is consistent across space-time and EO imaging sensors is the basis for human-like query development in big image databases. Other free and open ancillary datasets, such as a digital elevation model (DEM), can provide additional evidence if incorporated. For example, a DEM allows explicitly querying the phenology of alpine flora or distinguishing between mountain lakes and non-mountain lakes.

Table 1: Feature-matrix for different approaches of storing and processing EO images

Feature	File-based EO image hubs (e.g. Copernicus open access hub)	State-of-the-art data cubes	*Sen2Cube.at* data & information cube approach
Image download	✓	✓	✓
Metadata-based search	✓	✓	✓
Image-wide processing	✓	✓	✓
AOI-based processing	⊘	✓	✓
Time series analysis (statistical)	⊘	✓	✓
AOI-based cloud-free image search & mosaicking	⊘	⊘	✓
Time series analysis (semantic)	⊘	⊘	✓
Semantic content-based image retrieval (SCBIR)	⊘	⊘	✓
Content-based best pixel selection for cloud-free composites	⊘	⊘	✓
No expert-knowledge required to produce information on a higher level	⊘	⊘	✓
Generic approach with re-usable and sharable tools	⊘	⊘	✓
Additional data (e.g. DEM) can be used in the high-level queries	⊘	⊘	✓

Multiple **data cube technologies** exist, and evaluating and selecting one that is suitable and adaptable to specific requirements and semantic queries of EO data is critical. This concept extends state-of-the-art technology to provide a single, integrated information space to store data together with information (see Figure 1). Users are able to conduct semantic queries directly in the data and information cube, rather than simply searching for EO imagery based on metadata information exclusively. Performance and scaling tests are being conducted concurrently to the data cube design and will demonstrate that the user requirements can be met.

Designing a **Web-based inference engine** for different user scenarios is necessary because the data and information cube contains generic semantic enrichment suitable for multiple domains. This graphical inference engine is intended to produce new information without imposing unnecessary restrictions on what is possible using a user-friendly graphical

user interface (GUI). Users will be able to utilise the GUI to augment a knowledge base by means of designing semantic, domain-specific models. Initial models for different usage scenarios will be designed and tested with Sentinel-2 images of Austria plus additional available free and open geo-data (e.g. a DEM), but are in general sensor- and geography-independent, which fosters reproducibility and potential transferability to other initiatives outside Austria and other optical sensors.

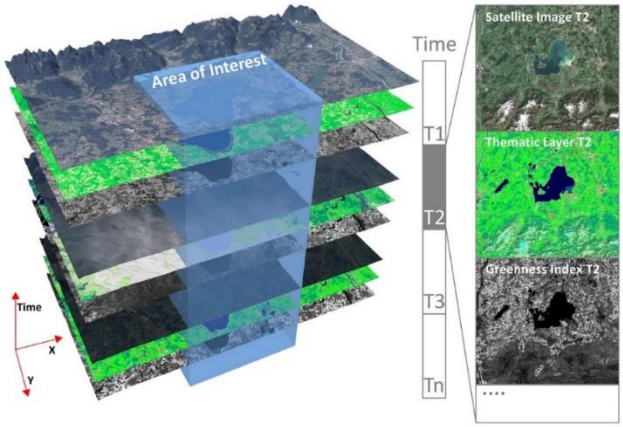


Figure 1: Automatically generated information layers will be linked with the Sentinel-2 data ready for spatio-temporal semantic queries in user-defined AOIs (from [7])

The generic data & information cube concept presented here allows for the development of completely different **demo services**, addressing a multitude of user needs. This is possible by means of the previously mentioned GUI, which gives different users, experts and the public the possibility to augment the knowledge base depending on their knowledge or specific application domain (see feature-matrix in Table 1). Additional proof-of-concept services to be developed will be defined by the users involved in the project, but also for generic use-cases of interest to the broader public (e.g. cloud free mosaic generation for specified time spans, semantic content-based image retrieval in user defined AOIs). These services will be evaluated based on their performance (quantitative and qualitative) together with the Austrian semantic data & information cube.

3. FIRST EXPERIENCES AND OUTLOOK

Even though the main use of EO data cubes up to now has been as a storage engine for accessing ARD, a data and information cube leverages each scene's semantic enrichment together with the reflectance values and additional data (e.g. DEM). Incorporating information creates an integrated data management system that provides methods for accessing, analysing and writing data. The data cube becomes a central part of an expert system as a fact base rather than mere storage. The fact base is constantly and incrementally augmented by the provision of new data and virtual or actual information layers that are derived using the user's expert knowledge.

When queries are made possible directly in a data cube there is a definite increase in input/output (IO) demand in comparison to when it is being used solely for data storage and access. To cope with this increased IO demand the project will evaluate the state-of-the-art data cube technologies such as rasdaman [13] and the ODC and how they handle different queries, whether prioritising spatial, temporal, both, neither or other aspects. One of these aspects is that semantic queries are not conceptually limited to pixel-based evaluation, but may require object-based or spatial analysis, either topological or non-topological. Evaluation will potentially be conducted in tandem with relational, object-based or graph databases, depending on the defined user requirements.

One example of a generic data and information data cube was already demonstrated by Augustin et al. [9] implementing a completely automated workflow for Sentinel-2 data acquisition, semantic enrichment and integration of scenes and generated information layers. The workflow utilises automated semantic enrichment generated using SIAM incorporating them into an implementation of the ODC, thus enabling semantic queries through time. They successfully demonstrated a semantic query of the prevalence of water-like observations through time, excluding observations categorised as being cloud-like.

The user's access point to the data cube in Sen2Cube.at is a graphical Web-based inference engine, which allows information extraction from the data and information cube independently of the client, either a Workstation, Laptop or mobile phone. The GUI intuitively connects the knowledge base, consisting of physical world models (ontologies of the world), and the fact base, i.e. the data cube made of data and information products, with the inference engine [14] (cf. Figure 2).

Once the semantic data & information cube is scaled up, it will result in the first SCBIR system in operational (working) mode for analysing EO data directly in big image databases - never before demonstrated on national scale. The innovative approach will prove the big data paradigm, "bringing users to the data and not data to the user", for EO data and associated information. This will open novel ways to exploit Sentinel-2 data and data-derived information, in particular for non-EO experts. The project outcome could be very relevant for many different user groups, who strive to make better use of Sentinel-2 data, and the generic, semantic EO data cube concept will enable a range of novel applications.

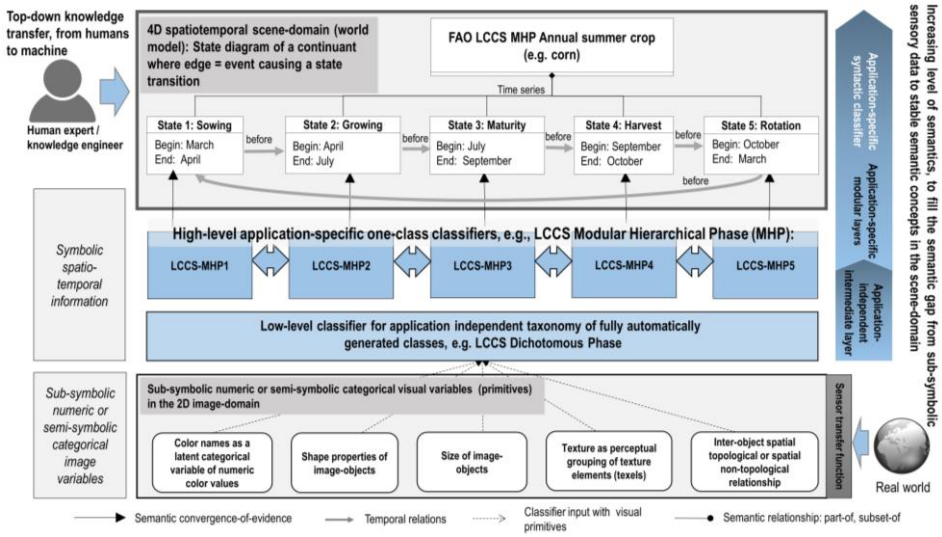


Figure 2: The graphical inference engine, implemented as a Web-based graphical user interface, is part of the expert system, which allows model-driven production of information on top of the data-driven automated extraction of information layers. The schema shown here illustrates the example of a temporal succession of different visual appearances of an annual summer crop, brought into the system by the user's domain knowledge and stored into the knowledge base. From [7].

4. ACKNOWLEDGEMENTS

The research has received funding from the Austrian Research Promotion Agency (FFG) under the Austrian Space Application Programme (ASAP) within the project Sen2Cube.at (project no.: 866016). We also thank the project partners and users *Spatial Services*, *Zentralanstalt für Meteorologie und Geodynamik* (ZAMG) and *Agrarmarkt Austria* (AMA) for their valuable contribution to the project design.

5. REFERENCES

[1] ESA, "Mission Status Report 139," https://sentinel.esa.int/documents/247904/3347201/Sentinel-2-Mission-Status-Report-139-25-August-07-september-2018.pdf, 2018.

[2] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, 2017.

[3] A. Lewis *et al.*, "The Australian Geoscience Data Cube - Foundations and lessons learned," *Remote Sens. Environ.*, vol. 202, pp. 276–292, 2017.

[4] P. Baumann *et al.*, "Big Data Analytics for Earth Sciences: the EarthServer approach," *Int. J. Digit. Earth*, vol. 9, no. 1, pp. 3–29, 2016.

[5] B. Killough, "CEOS Analysis Ready Data for Land (CARD4L) Description Document Preamble. online: http://ceos.org/document_management/Meetings/Plenary/30/Documents/5.5_CEOS-CARD4L-Description_v.22.docx," 2006.

[6] G. Giuliani *et al.*, "Live Monitoring of Earth Surface (LiMES): A framework for monitoring environmental changes from Earth Observations," *Remote Sens. Environ.*, vol. 202, pp. 222–233, 2017.

[7] D. Tiede, A. Baraldi, M. Sudmanns, M. Belgiu, and S. Lang, "Architecture and Prototypical Implementation of a Semantic Querying System for Big Earth Observation Image Bases," *Eur. J. Remote Sens.*, vol. 50, no. 1, pp. 452–463, 2017.

[8] B. Wilkinson, Mark D. ; Dumontier, Michel ; Aalbersberg, IJsbrand Jan ; Appleton, Gabrielle ; Axton, Myles ; Baak, Arie ; Blomberg, Niklas ; Boiten, Jan-Willem ; da Silva Santos, Luiz Bonino ; Bourne, Philip E. ; Bouwman, Jildau ; Brookes, Anthony J. ; Clark, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, pp. 1–9, 2016.

[9] H. Augustin, M. Sudmanns, D. Tiede, and A. Baraldi, "A Semantic Earth Observation Data Cube for Monitoring Environmental Changes during the Syrian Conflict," *GI_Forum*, vol. 1, no. 1, pp. 214–227, 2018.

[10] M. Sudmanns, D. Tiede, L. Wendt, and A. Baraldi, "Automatic Ex-post Flood Assessment Using Long Time Series of Optical Earth Observation Images," *GI-Forum J. Geogr. Inf. Sci.*, vol. 1, pp. 217–227, 2017.

[11] A. Baraldi, L. Durieux, D. Simonetti, G. Conchedda, F. Holecz, and P. Blonda, "Automatic Spectral-Rule-Based Preliminary Classification of Radiometrically Calibrated DMC / SPOT-1 / -2 Imagery — Part I : System Design and Implementation," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1299–1325, 2010.

[12] A. Baraldi, M. L. Humber, D. Tiede, and S. Lang, "GEO-CEOS stage 4 validation of the Satellite Image Automatic Mapper lightweight computer program for ESA Earth observation level 2 product generation - Part 1: Theory," *Cogent Geosci.*, vol. 4, no. 1, pp. 1–46, Apr. 2018.

[13] P. Baumann, A. Dehmel, P. Furtado, R. Ritsch, and N. Widmann, "The multidimensional database system RasDaMan," in *Acm Sigmod Record*, 1998, vol. 27, pp. 575–577.

[14] M. Sudmanns, D. Tiede, S. Lang, and A. Baraldi, "Semantic and syntactic interoperability in online processing of big Earth observation data," *Int. J. Digit. Earth*, vol. 11, no. 1, pp. 95–112, Jan. 2018.

FROM ANALYSIS-READY DATA TO ANALYSIS-READY SERVICES: CHALLENGES AND HELPERS FOR EO SERVICE PROVIDERS

Peter Baumann

Jacobs University, rasdaman GmbH

Abstract – Despite the current wave of providing data analysis-ready we claim that some essential properties for easy, non-EO-expert and non-programmer exploitation of EO data are not usually considered in service design. These properties relate to the quality of service (human or machine) users experience, and conversely to the burden that is imposed when accessing archives. We spot some critical features and propose solutions.

Index Terms— Analysis-ready data, datacubes, standards, coverages, Web Coverage Service (WCS), Web Coverage Processing Service (WCPS), OGC

1. INTRODUCTION

The term Analysis-Ready Data (ARD), originally coined by the USGS Landsat team in 2017 [14], has seen a rapid uptake in the Earth Observation (EO) community. Not surprisingly, we encounter a variety of different interpretations which, however, all agree that EO data need to be offered in a way better suitable for consumption in particular by non-programmers and non-EO experts.

CEOS recently started to propagate CEOS Analysis Ready Data for Land (CARD4L) as data processed to allow "immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets" [9]. Among some metadata requirements CARD-4L implies radiometric and geometric calibration plus solar and view angle correction and atmospheric correction (optical) and topography and incidence angle correction (radar).

Obviously, among the core features of such data is to offer EO data in a homogenized, aggregated manner which abstracts away from particular storage organizations and encodings which traditionally pose problems to users – sometimes described as going "from files to pixels" to indicate the different semantic level of EO offerings. Standards are helpful here if they establish an abstraction not based on files and scenes, but on higher-level objects, such as the OGC Web Coverage Service (WCS) suite [2][1][6][7].

As temporal analysis constitutes today's killer application in EO it is indispensable that analysis readiness does not only address horizontal spatial extent (as has been achieved with seamless maps) but also time axis. Ultimately, all spatio-temporal axes should be included thereby having elevation and bathymetry, too. In the end, spatio-temporal analysis readiness inevitably leads to the concept of multi-dimensional datacubes, first presented in [8]; see also [1].

However, while the advantages of such a data organization for access (i.e., simple download) of data are imminent. Even ftp download, however, constitutes a service API, albeit with rudimentary functionality – and this is what we observe many organizations still focusing on. However, users today want to get away from a service philosophy of "go take the data and do the analysis yourself" but rather expect server-side analysis capabilities. Obviously, the quality of service is of crucial relevance for user uptake. We claim, therefore, that in parallel to analysis-ready data we need to consider *analysis-ready services*. In this contribution we first inspect the state of the art, based on the ESA Sentinel archives. By doing so we spot several shortcomings which allow us to propose corresponding steps towards better EO service quality. To demonstrate feasibility of these ideas we present their realization in the European Datacube Engine, rasdaman.

The remainder of this paper is organized as follows. In Section 2 we exemplarily describe EO archive structures which complicate access. In Section 3, we introduce steps for improvement, and in Section 4 we describe a sample implementation of such steps. Section 5 concludes the paper.

2. EO SERVICES: STATUS AND IMPEDIMENTS

In contrast to many discussions about analysis-ready data we adopt a holistic approach and consider consequences of design decisions for the user experience. The central question guiding us is: How much knowledge and work is needed by the client in order to perform a particular task in some server? Knowledge includes aspects such programming skills required for performing a given task; work refers to the number of steps to be performed by the client, their complexity, and their resource needs.

We are of the opinion that such questions are applicable to both human users – typically accessing a service through some visual point-and-click client – and machine users where some algorithm – possibly deeply hidden in some service mash-up – when accessing a service. For the service as such this does not make any difference as it invariably "sees" the client through protocols and API invocations. Therefore, we prefer talking about clients than users in the sequel. Based on these considerations we establish some sample service request situations which will form our test cases subsequently.

doi:10.2760/848593

2.1. EO Archive Access Use Cases

The service features commonly discussed go substantially beyond downloading of objects or parts of it ("subsetting"), but include various aspects of server-side processing in the widest sense (not that already reformatting into another encoding involves CPU cycles). We find the following classification useful:

Data access: complete download of a particular object which has been identified through some search, link, or metadata reference. Implementations emphasizing simplicity of the server code often require that the object be returned in its exact original byte steam representation, such as the data format in which the object is stored in the server. A typical example is OGC Web Map Tiling Service (WMTS).

Data extraction: download of a part of an object identified. As this means drilling into the object anyway this use case is often combined with re-encoding into some client-selected data format. A typical example is OGC Web Coverage Service (WCS) Core.

Data filtering: prior to downloading find out whether some data object is fit for your purpose. This can require inspection of both data and metadata. OGC Web Coverage Processing (WCPS) can do this on sets of datacubes [5].

Data processing: apply some computational steps to an object in the server (following the Big Data paradigm of "ship code to server") and ship the resulting (new) object to the client. This can be a fixed, predefined process (such as through an OGC Web Processing Service (WPS) process) or an ad-hoc, flexible query (such as through an OGC WCPS processing request). For reasons of differentiation we assume processing of always one object, as the case of combination is addressed separately, coming next.

Data fusion: recombine a result object from two or more server-side objects. In the most general case these objects can reside on different servers, obviously subject to different, independent regimes of data presentation in terms of extent, resolution, Coordinate Reference System (CRS), etc.

Data maintenance: modify the offering of a remote service by either creating a new object, update all or part of an existing object, or delete an object from this server. Such updates must be possible concurrently to other client access and therefore need to adhere to the well-known ACID transaction properties.

2.2. ESA EO Archive Data Provisioning Case Study

For the Sentinel archives ESA suggests the SAFE format for uniform access to data offered. EO data are preprocessed into so-called granules which can be seen as tiles. From our perspective, some properties of SAFE are in particular practically relevant; we discuss these in turn.

A granule covers an area of 100 x 100 km. This leads to file sizes of typically 600 – 800 MB meaning that users must download files of these sizes for any processing. Moreover, it also means that any service working on such granules must load units of this size into main memory before processing of any request can start, be it a simple WCS GetCoverage or a complex WCPS analytics request. Detailed benchmarks have shown that an optimal tile size for general-purpose extraction and processing is in the area of 3 – 5 MB [10]. Hence, the units of storage access are by about two orders of magnitude too coarse for being efficient.

Further, a SAFE file is a zip archive containing the pixel payload plus a series of metadata. In terms of storage access this means that the zip file needs to be opened and the image file(s) need to be extracted. Depending on the implementation of the zip decoder this may mean significant extra processing which slows down server-side result generation.

Finally, image files are provided in the JPEG format following a lossless encoding regime. Using JPEG – despite its lossless storage – has several relevant consequences. As JPEG applies a transformation from time to spectral space, reconstructing a pixel from a JPEG stream requires (i) accessing several memory locations and (ii) significant CPU cycles. Altogether, albeit such data will already be in RAM this means extra overhead slowing down response generation in the server.

Yet another consequence of the wavelet-based storage of Sentinel products is the inability to optimize spatio-temporal subsetting, one of the most basic and widely used access operations at all. Some formats, like TIFF and NetCDF, support internal tiling which an intelligent server may exploit to load less than the whole file for subsetting requests. Obviously, considering the hundreds of Megabytes of file sizes, this can mean a significant difference in data loading. In contrast, with JPEG such a data load optimization is not possible as data are structured in a completely different manner on disk. In passing we note that all these computational steps may require intermediate representations in main memory or, even worse, on disk which additionally impacts request response time and server resource consumption.

All these considerations also hold for updates, in particular partial updates which are common when building and maintain a datacube.

3. RECOMMENDATIONS TOWARDS ANALYSIS-READY SERVICES

In this Section we set up a set of requirements aiming at making services more analysis-ready. Our guidance is simple: *how much effort – again, in terms of knowledge and resource requirements – does it take for a client to access and process a particular pixel set in the course of decision making?* Based on the observations we propose the following set of recommendations for EO service providers in order to achieve a high level of service quality. We differentiate between data and service modeling aspects, bearing in mind though that both are tightly intertwined.

Requirement 1: Provide data access in a granularity suitable for efficient storage access across all spatio-temporal dimensions, i.e.: x/y/z/t. This can be achieved by either re-

doi:10.2760/848593

tiling of data into a scheme that best fits client access patterns or at least utilizing some file format that supports internal tiling, such as NetCDF. Tile shape and size must be adjustable for the server architecture and workload – there is no "one size fits all" tiling for spatio-temporal data. Particular algorithms (like convolutions) and user scenarios (like disaster mitigation) will lead to different most suitable tilings. As normally more than just one application should be supported there will regularly be conflicting optimal tiling schemes, in which case a tradeoff will have to be found. For datacubes this applies to all x/y/z/t dimensions, hence traditional 2D GeoTIFF archives will show degraded performance.

Requirement 2: Minimize the number of CPU cycles and storage / memory access required for reconstructing a given pixel set in main memory. This rules out wavelet-based encoding options.

Requirement 3: Store data analysis-ready. Reconstruction of analysis-ready data on the fly is not only inefficient (if almost every query will require the same processing) but may introduce numerical inconsistencies. The authoritative values should be readily available in the database / archive. In terms of the usual processing levels, this excludes Level 1a and 1b; analysis-readiness in the sense of "we can logically aggregate into user-centric units such as datacubes without any loss of precision" starts with Level 1c (error corrected, radiometrically corrected, orthorectified).

Requirement 4: Ship code to data. Surprisingly, this well-known Big Data principle is not always implemented today. Low-level ftp, RESTful subsetting APIs, etc., do not allow server-side processing, but leave that to the client. However, also many python-based APIs, as well as WPS-based approaches, require application code to run on the client with just procedural calls to fixed server-side functionality. Instead, clients should be able to ship their processing requests for execution on the server, close to the data to avoid expensive data shipping round trips.

Requirement 5: High-level server-side filtering and processing language. While "ship code to data" is a must implementations vary widely in the API quality. Sometimes procedural source code, such as python, is shipped to the server for execution – obviously, a major security hole. Instead, a high-level, declarative language should be provided at the abstraction level of, say, SQL with its tremendous success. An equivalent is given by the WCPS Earth datacube analytics language [5] which is declarative, has a well-defined semantics, and is adopted OGC standard.

Requirement 6: Transparent federation. Data fusion often requires combination of objects sitting in different data centers. Ideally, the task of extraction, download, homogenization, and combination should not be with the client, but on the server. This requires intelligent ad-hoc orchestration of arbitrary servers, including optimization of data exchange and processing distribution. Obviously, federation has a potential for massively boosting ease-of-use and performance.

Requirement 7: Open standards. In the spirit of interoperability data and service APIs should adhere to well defined and curated standards. Looking at the rigor of maintenance required this calls for standards, e.g., by OGC, ISO, and OASIS Open; in contrast, e.g., the W3C Spatial Data on the Web group has disbanded after releasing its documents. For EO data, specifically, the OGC and ISO standards apply which have the additional advantage of being kept in lockstep synchronization (e.g., OGC CIS [2] is identical to ISO 19123-2). Notably, the WCS suite allows both ingest and retrieval based on the same conceptual model, OGC coverages [2].

4. IMPLEMENTATION FEASIBILITY

In this Section we demonstrate feasibility by inspecting a service tool which, among others, offers the features recommended for efficient, client-friendly access. This is the European Datacube Engine and OGC datacube reference implementation, rasdaman ("raster data manager"), which has been developed over two decades into a cross-domain datacube engine [12][3][10]. A general survey of datacube tools has been published by RDA [13].

The rasdaman engine resembles a complete software stack, implemented from scratch to support fastest management and retrieval on massive multi-dimensional arrays in a domain agnostic way. Its array query language, rasql, meantime is adopted as the ISO SQL Multi-Dimensional Arrays (MDA) standard [11]. For EO datacubes rasdaman supports the declarative spatio-temporal datacube analytics language standard, OGC WCPS [5] (Requirements 4 & 5).

The overall system architecture centers around the multi-parallel rasdaman worker processes which operate on arbitrarily tiled arrays (Requirements 1 & 2, see Figure 1) stored in a database or read from some legacy archive (hence avoiding copies). When ingesting data they can be stored in a number of formats, including the CPU's main memory array format (Requirement 2), through a WCS-T based ETL layer (Requirement 3) which homogenizes data and metadata, provides defaults, as well as the target tiling strategy [10]. Further tuning parameters include compression, indexing, cache sizing, etc. The resulting OGC compliant coverages represent analysis-ready space-time EO objects.

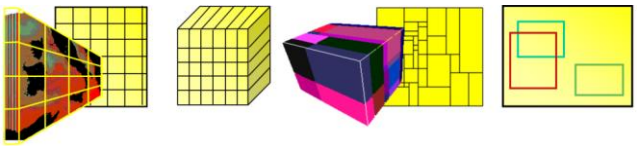


Figure 1 - Sample rasdaman datacube partitioning strategies (source: rasdaman).

In a rasdaman federation (Requirement 6), worker processes can fork subqueries to other cloud nodes or other data centers for load sharing and data transport minimization [4] (Figure 2). Figure 3 shows a visualization of actual federated query processing between the European Centre for

Medium-Range Weather Forecast (ECMWF) in the UK and National Computational Infrastructure (NCI) in Australia - both running rasdaman - for determining heavy rainfall risk areas from precipitation data at ECMWF and Landsat8 imagery at NCI [3]. The two query paths all lead to the same result for the user, thereby achieving location transparency.

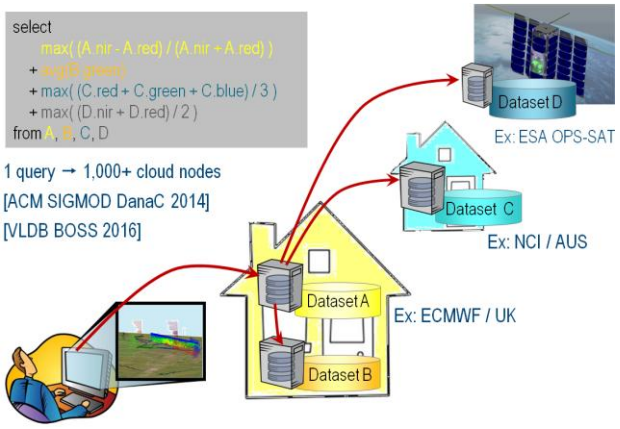


Figure 2 - rasdaman transparent distributed query processing (source: rasdaman).

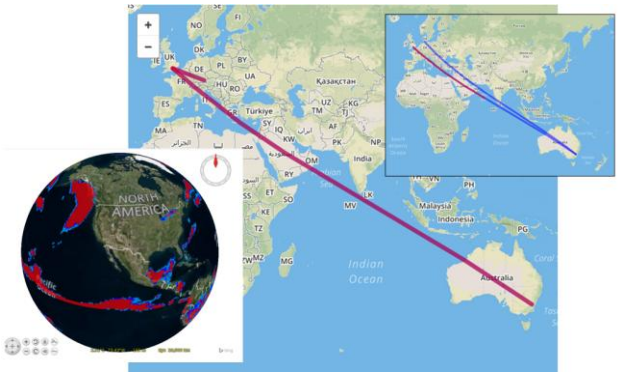


Figure 3 - Sample rasdaman intercontinental federation query [13].

Being official OGC WCS Reference Implementation, rasdaman at the same time, to the best of our knowledge, is the most comprehensive WCS suite implementation and the only tool supporting all WCS extensions (Requirement 7).

5. CONCLUSION

In this contribution we motivate a less data-centric and more service-centric view, acknowledging that both are just two sides of the same coin. In a nutshell, data are ready for analysis when common math can be applied on the data without tweaking it for sensor or archive characteristics. This is no rocket science as we have shown; rather, the resulting requirements are available in implementation, thus underlining technical feasibility.

We do not claim that our list of EO service requirements is final, rather it is likely that more service quality facets will come up in future. However, from our experience with multi-Petabyte datacube service federations we feel that the requirements listed are all essential. As such, it is the hope that this contribution stimulates further discussion, shedding more light on service aspects than has been done up to now.

6. ACKNOWLEDGEMENT

This work has been supported partially by EU H2020 Land-Support, EU H2020 EOSC-hub, German BMWi BigData-Cube, and German BMEL BigPicture.

7. REFERENCES

[1] P. Baumann, D. Misev, V. Merticariu, B. Pham Huu: "Datacubes: Towards Space/Time Analysis-Ready Data". In: J. Doellner, M. Jobst, P. Schmitz (eds.): Service Oriented Mapping - Changing Paradigm in Map Production and Geoinformation Management, Springer Lecture Notes in Geoinformation and Cartography, 2018.

[2] P. Baumann, E. Hirschorn, J. Maso: "Coverage Implementation Schema, version 1.1". OGC document 09-146r6, www.opengeospatial.org/standards/wcs.

[3] P. Baumann, A.P. Rossi, B. Bell, O. Clements, B. Evans, H. Hoenig, P. Hogan, G. Kakaletris, P. Koltsida, S. Mantovani, R. Marco Figuera, V. Merticariu, D. Misev, B. Pham Huu, S. Siemen, J. Wagemann: "Fostering Cross-Disciplinary Earth Science Through Datacube Analytics". In. P.P. Mathieu, C. Aubrecht (eds.): Earth Observation Open Science and Innovation - Changing the World One Pixel at a Time, International Space Science Institute (ISSI), 2017

[4] P. Baumann, V. Merticariu: "On the Efficient Evaluation of Array Joins". Proc. IEEE Big Data Workshop Big Data in the Geo Sciences, Santa Clara, US, October 29, 2015

[5] P. Baumann: "The OGC Web Coverage Processing Service (WCPS) Standard". Geoinformatica, 14(4)2010, pp 447-479.

[6] P. Baumann: "Web Coverage Service (WCS) Interface Standard – Core, version 2.0". OGC document 09-110r4, www.opengeospatial.org/standards/wcs.

[7] P. Baumann: "OGC Coverages Domain Working Group Public Wiki". http://myogc.org/coveragesDWG.

[8] P. Baumann: "Language Support for Raster Image Manipulation in Databases". Proc. Int. Workshop on Graphics Modeling, Visualization in Science & Technology, Darmstadt / Germany, April 13 - 14, 1992, Springer 1993, pp. 236 – 245.

[9] CEOS: CEOS Analysis Ready Data for Land (CARD4L) Description Document. http://ceos.org/ard

[10] P. Furtado, P. Baumann: "Storage of Multidimensional Arrays Based on Arbitrary Tiling". Proc. ICDE'99, March 23-26, 1999, Sydney, Australia.

[11] D. Misev, P. Baumann: "Enhancing Science Support in SQL". Proc. IEEE Big Data Workshop Data and Computational Science Technologies for Earth Science Research, Santa Clara, US, October 29, 2015

[12] hidden rasdaman team: "rasdaman: Datacubes on Steroids". Proc. ACM SIGSPATIAL, Seattle, USA, November 06, 2018

[13] RDA: "Array Database Assessment Working Group Report". https://www.rd-alliance.org/groups/array-database-working-group.html.

[14] USGS: "U.S. Landsat Analysis Ready Data (ARD)". https://landsat.usgs.gov/ard

DATACUBES AS A TOOL FOR ANALYSIS READY DATA INTER-COMPARISON

Simon Oliver[1], Lan-Wei Wang[1], Medhavy Thankappan[1], Tina Yang[1], Fuqin Li[1], Joshua Sixsmith[1]

[1]Geoscience Australia, Corner of Jerrabomberra Avenue and Hindmarsh Drive, Canberra, Australia - simon.oliver@ga.gov.au

ABSTRACT

There is an increasing push by data suppliers and users to access Analysis Ready Data (ARD) as a standard product package for moderate resolution Earth Observation (EO) data. The United States Geological Survey (USGS) is progressing work to deliver a global ARD product from Landsat observations, as part of their upgrade to Collection 2. The European Space Agency (ESA) plans to release a global surface reflectance product based on the Sentinel-2 mission data. Many other agencies have developed a significant capability for routine generation of similar products at continental and regional scale. This paper introduces the Open Data Cube (ODC) as a basis for exploring and inter-comparing long time-series records of co-registered surface reflectance measurements in a consistent approach. Geoscience Australia, through the Digital Earth Australia (DEA) program, has developed an inter-comparison tool to enable users to assess the fitness for purpose and effectiveness of global and regional scale correction methodologies employed to derive surface reflectance. The inter-comparison tool enables investigation into the consistency of ARD products through time and also the absolute accuracy of the products compared to field-based measurements.

Index Terms— Analysis Ready Data, surface reflectance, Digital Earth Australia, Open Data Cube, validation

1. INTRODUCTION

Since digital product delivery began, radiance at sensor, or Level 1 - [3], has been viewed as the standard moderate resolution EO product offered by satellite data providers. This baseline shifted significantly with the automation of precision ortho-correction of at sensor radiance, which led to spatially aligned observations becoming the norm. Increasingly, users have been calling for even further advances and delivery of higher order processing and standardisation of measurements such as that offered by surface reflectance correction (often categorised as a Level 2 product [3]). The radiometric alignment of observations provides for inter-comparison within a time-series of measurements from a given sensor. Once seen as the domain of value-adders in the EO marketplace, space and associated agencies managing the delivery of public good mission data have increasingly looked to produce surface reflectance as the new benchmark. The ability to directly apply such products to analysis without further processing led to a global movement and development of a set of guidelines by the Committee on Earth Observation Satellites (CEOS) around what is now commonly referred to as Analysis Ready Data (ARD) [5]. CEOS through the Land Surface Imaging Virtual Constellation team has led the work on defining specifications for the CEOS Analysis Ready Data for Land or CARD4L covering multiple product families. CARD4L are satellite data that have been processed to a minimum set of requirements and organised into a form that allows immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets. Through this effort also came recognition of the value of additional auxiliary products within the ARD product package that would further facilitate direct analysis of the data (per-pixel quality assessment, metadata, and sensor geometry etc.).

For a number of years, Landsat Level 2 data has been available on-demand via the United States Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center Science Processing Architecture (ESPA) On Demand Interface [7]. The United States (conterminous United States) Landsat Analysis Ready Data (ARD) are consistently processed to the highest scientific standards and level of processing required for direct use in monitoring and assessing landscape change. A fundamental goal for Landsat ARD is to significantly reduce the magnitude of data processing for application scientists, who currently have to download and prepare large amounts of Landsat scene-based data for time-series investigative analysis [10]. The USGS is currently working towards delivery of global Landsat Collection 2 ARD.

The Sentinel-2 Level-2A product from ESA provides Bottom Of Atmosphere (BOA) reflectance images derived from the associated Level-1C products. Each Level-2A product is composed of 100 km2 tiles in cartographic geometry (UTM/WGS84 projection). Level-2A products are systematically generated at the ground segment over Europe [9], with production to be extended to global coverage by the end of 2018.

Geoscience Australia, through the Digital Earth Australia and predecessor programs, has a long heritage in production

of entire Landsat and Sentinel-2 time-series of surface reflectance at continental scale. Through the techniques developed by Li et al. 2012 [2] and its implementation in the wagl pipeline code [6], GA utilises surface reflectance as the foundation product from which the value to government programs is realised. In comparison with USGS and ESA products, the GA Level 2 product includes, amongst other differences, Bi-directional Reflectance Distribution Function (BRDF) and Terrain Illumination correction. One perceived advantage of GA's approach over those provided by the space agencies, is in the consistency of algorithm and inputs to the correction - wagl applies a common correction methodology across all Landsat and Sentinel-2 sensors. It is theorised that this consistency leads to a reduction in barriers to product interoperability and interpretation of analyses, and facilitates improved data fusion across the range of sensors.

The Open Data Cube open source software suite, pioneered by Geoscience Australia and now driven by a growing international community, provides an integrated gridded data analysis environment for decades of analysis ready EO and related data from multiple satellite and other acquisition systems. The Data Cube is a system designed to: catalogue large amounts of EO data, provide a Python based API for high performance querying and data access, give scientists and other users easy ability to perform exploratory data analysis, allow scalable continental processing of the stored data, track the provenance of all the contained data to allow for quality control and updates [8].

The open data cube platform, and specifically the core component of the code base, has been used as the basis for development of a number of tools supporting applications and tools which include: a web map server, plugin for QGIS and the ODC Cubedash dashboard.

2. THE CHALLENGE

As space agencies evolve the benchmark for standard products, entities who have their own expertise and technologies in production of Level 2 data are seeking ways to comprehensively compare them against the standard Level 2 products. The fitness for purpose evaluation and inter-comparison of corrected data is being used to provide an evidence base for strategic decision making. As the standard Level 2 products [2] are developed using different algorithms (LEDAPS vs LaSRC [2] in the case of Landsat Level 2 / ARD), and with different ancillary and geometric sources (USGS Landsat vs ESA Sentinel-2), understanding where and how these differences might have a negative impact on derivative products is important.

The inter-comparison tool was initially developed as part of an experiment to examine the relative performance of USGS LEDAPS and LaSRC corrected Landsat Level 2 [7] compared to the GA Lambertian, NBAR (Normalised BRDF Adjusted Reflectance) and NBART (as per NBAR plus terrain illumination correction) products (Li et al. 2012) [2].

3. METHODS

3.1. Data preparation

Initially, 12 Landsat WRS2 (Landsat World Reference System 2) path rows were identified for the development of the tool to assess low/medium vegetation cover areas and high BRDF (forest) areas. Acquisitions in paths adjacent to the target path/row were also acquired and processed in order to allow examination of the impacts on the temporal consistency of the products. As DEA's production workflow was based on a different Level 1 specification (differences in projection, pixel resolution, scan gap infill method and pixel reference), Level 1 and 2 (Collection 1) data were sourced from USGS. Code was developed to automate the process of ordering and retrieval of the full time-series of data leveraging the ESPA Application Programming Interface (API). GA's wagl code generated an ARD package outputting both Lambertian (equivalent to the USGS Level 2), NBAR and NBART products. Datasets were prepared for datacube indexing through generation of a datacube compatible metadata files using preparation scripts available at https://github.com/GeoscienceAustralia/dea-notebooks/tree/tinaY/11_Inter_comparison [15].

The wagl code was changed to use the same solar irradiances as the USGS method and to retain a Lambertian surface reflectance in the output product. The additional processing steps included in the NBAR and NBART were considered to result in distinct differences which would prohibit a reasonable inter-comparison for key areas of product performance.

3.2. Data Cube construction

An ODC instance was deployed in a test environment and product descriptions for the USGS and GA surface reflectance products developed and added. The USGS and GA Level 2 products were then indexed to make them available to Data Cube processes [8].

3.3. Inter-comparison tool development

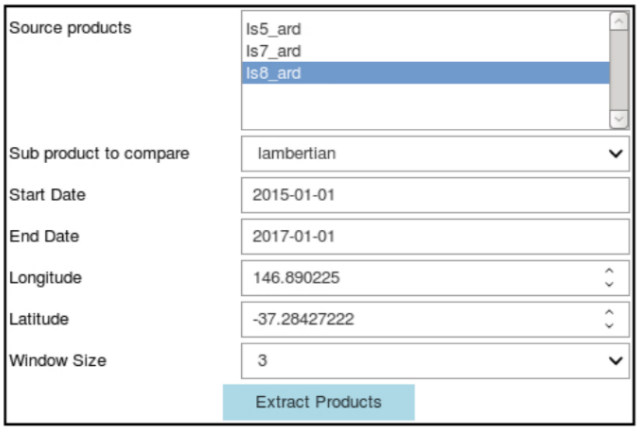


Fig 2. The inter-comparison tool running a single point location query based on a 3x3 pixel window size.

The inter-comparison tool was written in Python and heavily leveraged the Jupyter Notebook framework, ODC API and Plotly library [14]. Development focussed on enabling side by side comparison of two sets of measurements through time with a key feature being interaction with graph. The inter-comparison tool offers "area of interest" selection from a vector file or a point location (Figure 2) with user defined variable buffer (3x3, 5x5, 7x7 pixels); it can be run in batch processing mode.

The inter-comparison tool leverages integral components of ARD by including the ability to apply masks based on quality assurance layers, to the selected area of interest in order to eliminate cloud and other artefacts. The two basic parameters introduced into the visualisation of results include the mean and standard deviation of both sets of data.

In addition to the surface reflectance measurements, all relevant attributes associated with the area of interest can be extracted as CSV outputs for further analysis. These attributes include the minimum, maximum, and surface reflectance variance, valid pixel percentage after masking, aerosol, BRDF parameters (if available), ozone, water vapour, cloud cover percentage, solar azimuth/zenith angle, incident/exiting angle.

4. RESULTS

The inter-comparison tool has also been developed to allow plotting of in-situ spectra collected coincident with satellite overpass to provide a point of truth in ARD product inter-comparison.

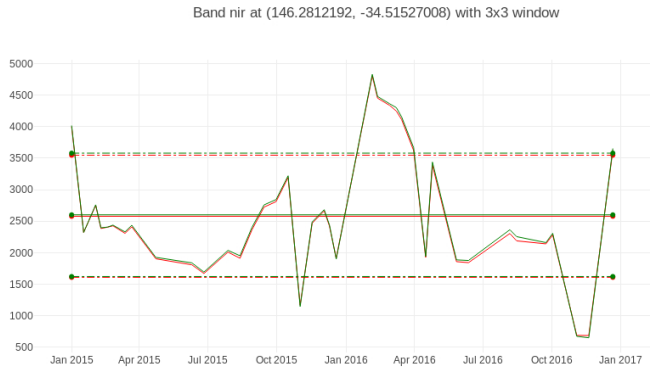


Fig 3. The inter-comparison tool demonstrating similarities in performance of USGS Level 2 (green) and GA Lambertian (red) Landsat Near Infrared over a flat agricultural site over a two year period. The complete straight and dashed horizontal line represents the mean and standard deviation of the time series.

The inter-comparison tool allows for investigation of phenomena in the landscape. Figure 3 demonstrates an equivalent performance of the USGS and GA Lambertian Near Infrared surface reflectance products. The GA NBART products exhibit less variation in reflectance compared to the USGS product in locations where terrain illumination is an influence as shown in Figure 4.

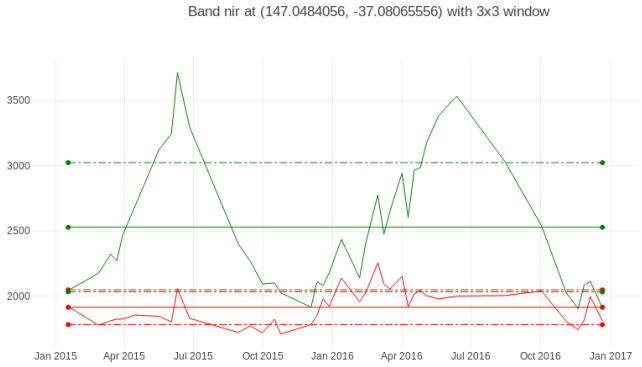


Fig 4. The inter-comparison tool demonstrating differences in performance of USGS Level 2 (green) and GA NBART (red) Landsat Near Infrared on a north east facing hill slope over a two year period. The complete straight and dashed horizontal line represents the mean and standard deviation of the time series.

5. DISCUSSION AND CONCLUSIONS

A key driver for the development of the inter-comparison tool was the need to establish an evidence-base to underpin the Geoscience Australia's approach to production of surface reflectance products. The tool has been effective in conveying the relative advantages of the different correction methods and products available. The inter-comparison work is just one component of a much broader body of work that is also examining how different surface reflectance

algorithms impact the quality of derivative biophysical parameters which will be elaborated in subsequent reports. The Data Cube inter-comparison tool makes effective use of existing open source software libraries and open data to provide crucial information to underpin a consistent approach to assess the relative performance of different surface reflectance algorithms derived from a common baseline (Landsat Collection 1 in this case).

The tool has been extremely effective in demonstrating how the algorithmic approach employed in creation of a given product can impact on the retrieval of accurate and consistent surface reflectance. It also provides a convenient means for users to quickly evaluate new sources of data to determine their fitness for purpose.

The inter-comparison tool represented in this paper is a component of a larger range of activities currently underway within the DEA program surrounding ARD inter-comparison and investigations into the fundamental requirements of interoperable data from different sources. As entities such as GA move towards adoption of Level 2 products from space agencies, a full comprehension of the impacts, both positive and negative, of adopting them in lieu of existing in-house capabilities. Ultimately, shortcomings of the standard products can be fed back to the provider in order to develop a solution which enables the broadest adoption of the data. GA plans to extend the analysis to Sentinel 2 inter-comparison as the global products become available.

More broadly, the accuracy of surface reflectance depends on the ancillary data, e.g., aerosol data, water vapour and surface BRDF parameters. Future studies will aim (i) to conduct sensitivity analysis for different parameters and to see how these parameters impact the overall results; (ii) to see what level of the processing is needed for consistency in the data and whether BRDF normalization is necessary for time series analysis.

The techniques employed here can be used to characterise other products such as cloud classification algorithms across both spatial and temporal dimensions. The development of the tool also lead to a proposal to refine the target data selection in order to better characterise and target the quantitative comparison.

6. ACKNOWLEDGEMENTS

The authors wish to acknowledge the contribution of Landsat data by USGS EROS.

7. REFERENCES

[1] A. Lewis, S. Oliver, L. Lymburner, B. Evans, L. Wyborn, N. Mueller, G. Raevksi, J.Hooke, R. Woodcock, J. Sixsmith, W. Wu, P. Tan, F. Li, B. Killough, S. Minchin, D. Roberts, D.Ayers, B. Bala, J. Dwyer, A. Dekker, T. Dhu, A. Hicks, A. Ip, M. Purss, C. Richards, S. Sagar, C.Trenham, P. Wang, L.-W. Wang, "The Australian geoscience data cube — foundations and lessons learned", *Remote Sens. Environ.*, 202, pp. 276-292, 2017. https://doi.org/10.1016/j.rse.2017.03.015

[2] F. Li, D.L.B. Jupp, M. Thankappan, L. Lymburner, N. Mueller, A. Lewis, A. Held, "A physics-based atmospheric and BRDF correction for Landsat data over mountainous terrain", *Remote Sens. Environ.*, 124 pp. 756-770, 2012. https://doi.org/10.1016/j.rse.2012.06.018

[3] "Data Processing Levels | Science Mission Directorate", *Science.nasa.gov*, 2018. [Online]. Available: https://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products. [Accessed: 11- Oct- 2018].

[4] "Digital Earth Australia - Geoscience Australia", *ga.gov.au*, 2018. [Online]. Available: http://www.ga.gov.au/about/projects/geographic/digital-earth-australia. [Accessed: 10- Oct- 2018].

[5] "CEOS Analysis Ready Data", *Ceos.org*, 2018. [Online]. Available: http://ceos.org/ard/. [Accessed: 14- Oct- 2018].

[6] "GeoscienceAustralia/wagl", *GitHub*, 2018. [Online]. Available: https://github.com/geoscienceaustralia/wagl. [Accessed: 14- Oct- 2018].

[7] "Landsat Surface Reflectance Level-2 Science Products | Landsat Missions", *Landsat.usgs.gov*, 2018. [Online]. Available: https://landsat.usgs.gov/landsat-surface-reflectance-data-products. [Accessed: 14- Oct- 2018].

[8] "OpenDataCube", *opendatacube*, 2018. [Online]. Available: https://www.opendatacube.org/. [Accessed: 10- Oct- 2018].

[9] "User Guides - Sentinel-2 MSI - Level-2A Product - Sentinel Online", *Earth.esa.int*, 2018. [Online]. Available: https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/product-types/level-2a. [Accessed: 14- Oct- 2018].

[10] "U.S. Landsat Analysis Ready Data (ARD) | Landsat Missions", *Landsat.usgs.gov*, 2018. [Online]. Available: https://landsat.usgs.gov/ard. [Accessed: 14- Oct- 2018].

[14] "plotly", *Plot.ly*, 2018. [Online]. Available: https://plot.ly/python/. [Accessed: 15- Oct- 2018].

[15] "GeoscienceAustralia/dea-notebooks", *GitHub*, 2018. [Online]. Available: https://github.com/GeoscienceAustralia/dea-notebooks/tree/tinaY/11_Inter_comparison. [Accessed: 15- Oct- 2018].

DEVELOPING IMAGE PROCESSING CHAINS FOR THE THEIA LAND DATA CENTRE TO PROVIDE NEAR REALTIME MULTI-SATELLITE IMAGE PRODUCTS

Peter Kettig, Joelle Donadieu, Thibault Ducret, Simon Baillarin, Olivier Hagolle

Centre national d'études spatiales (CNES), Toulouse, France

ABSTRACT

The French public land data center Theia aims at providing advanced remote sensing products to ease user access to earth-observation satellite data. Within this data center, the Muscate processing platform uses several processors to produce and distribute high-quality multi-satellite earth observation images in near real-time. Muscate automatically orchestrates the work of these processors.

Each of them is aiming to maximize the value extracted for scientific exploration combining Sentinel, Landsat, Spot and recently Venμs images in order to increase availability. This paper will present Muscate, its currently implemented processors and finally the upcoming new processors and improvements.

Index Terms— Theia, Muscate, earth observation, big data, image processing

1. INTRODUCTION

The Theia land data centre[1] was established in 2011 by eleven French public organisms in order to promote the use of satellite remote sensing data by the scientific and public policy actors [1]. These actors request Earth observation data to monitor land surfaces. In response, Theia delivers a range of products and services allowing its userbase to maximize the profit of space missions.

Muscate (MUlti-Satellite, multi-sensor (CApteur in French) and multi-TEmporal data centre) is the dedicated processing platform to process the multi-satellite earth observation imagery in near real-time. With images from Spot, Landsat, Sentinel-2 as well as Venμs, the data is directed towards both the scientific community as well as public actors.

Inside Muscate, mulitple processing chains exist to enhance the satellite images, including:

- MAJA, a cloud screening and atmospheric correction processor [2] for Sentinel, Landsat and Venμs data
- LIS, a processor to detect snow on Sentinel-2 and Landsat-8 images [3]
- Iota2, a continuously updated land-cover map [4]

[1] Website: https://theia.cnes.fr

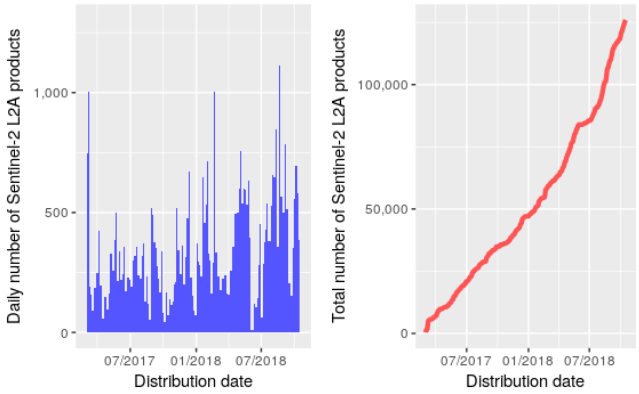


Fig. 1. Daily and total number of products since 01/03/2017 for the Sentinel-2 Level-2A collection

This paper will first show how Muscate is integrated into the high performance computing center at CNES in section 2. Afterwards, we present the currently existing algorithms (Section 3) as well as their main characteristics and finally their ongoing studies for improvements (Section 4).

2. THE ARCHITECTURE OF MUSCATE

Due to the high demand of satellite imagery, Muscate's goal is to offer near-realtime satellite images to the user.

The current continuous distribution is at around 500, spiking up to 1000 products a day for all collections combined. Figure 1 is showing the numbers for the Sentinel-2 Level-2A collection since its release to the public in March 2017.

The two main components of Muscate [5] are the acqusition-production module as well as the distribution module as depicted in Figure 2, with only the latter being interfaced with Theia. The acquisition and production is based on Phoebus, an orchestrator originally developed for the science ground segment of Gaia [6].

The satellite products arrive from their respective processing centers, such as Peps for Sentinel-2 and the VIP (Venus' processing centre) [7], mostly as Level-1C imagery: Orthorectified, top-of-atmosphere (ToA) images. The two main tasks of Phoebus are:

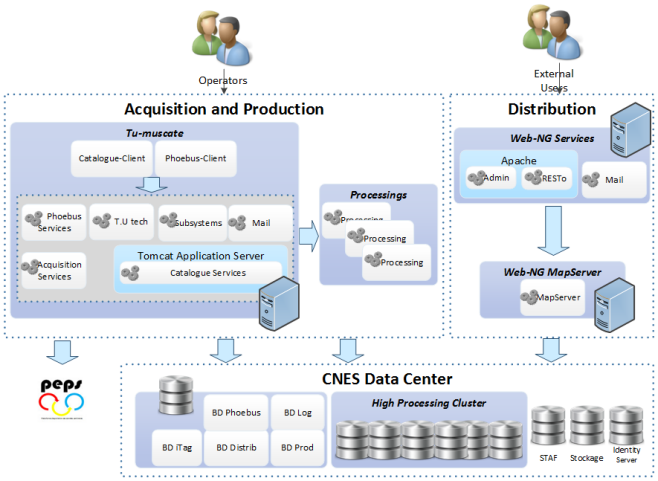


Fig. 2. Muscate architecture on the CNES HPC

- The creation of jobs, which indicate a single processing step and subsequently a workflow to be executed depending on the product

- The stable maintenance of the job-processing order, setting the priority for all tasks.

Further down the pipeline, the job (or workflow) is passed to the distributed resource manager (DRM) which interfaces directly to the high performance computing centre (HPC) located on-site at CNES in Toulouse. This creates an abstraction layer to interface more easily with the computing centre while keeping a high-effiency.

After a job is finished, the resulting product is automatically archived and sent to the distribution module which adds it to the given collection, allowing users to access it via the Theia MMI.

Using this architecture, the time from the acquisition of the product by the satellite until its availability within Theia is put down to a minimum while also keeping the needs for additional temporary storage low: More than 90% of the products acquired by Sentinel-2 between September and Mid-October 2018 were available on Theia's MMI within 3 days (cf. Figure 3).

3. PROCESSING CHAINS INSIDE MUSCATE

Once a product arrives in Muscate, it gets piped through multiple image processing chains, which will be further explained in the following.

All of them use the framework of the OrfeoToolbox (OTB) [8], which is an open-source image processing library for remote sensing applications developed by CNES.

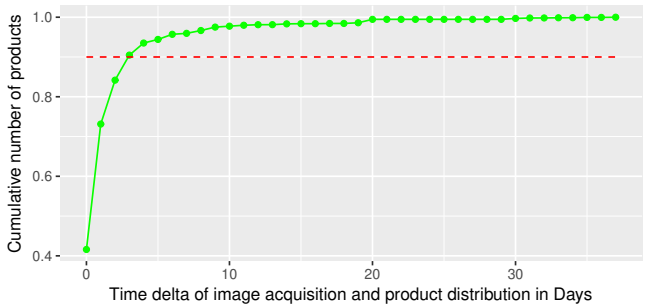


Fig. 3. Time lag between acquisition date and Theia production for Sentinel-2 products

3.1. MAJA (MACCS-ATCOR Joint Algorithm)

The basis for all image processing chains in Muscate is MAJA (MACCS ATCOR Joint Algorithm) [2]. The algorithm uses a multi-temporal approach to detect clouds, knowing that on two dates the probability of detection of clouds with the same shape is unlikely.

To detect clouds it uses a set of static and dynamic thresholds on the available spectral bands - mostly in the blue due to the higher contrast of clouds versus the rest. For some platforms, namely Venus and Sentinel-2 there are special cirrus spectral-bands available which are used to more reliably detect cirrus clouds [9].

In order to offer Theia's users a variety of data to choose from, three higher-level processing chains are currently implemented into Muscate which use the Level-2A outputs of MAJA. They will be presented in the following.

3.2. LIS (Let it snow)

This processing chain is used to detect snow throughout the year in central Europe and Canada's Quebec region, thus building up a comprehensive snow-cover history. The Level-2B outputs of the chain can be comprised of both Landsat-8 as well as Sentinel-2 images [3].

Its main difficulties are to distinguish the snow from surrounding clouds, which is why the algorithm incorporates a post-processing for the cloud masks provided in the Level-2A product in order to reduce false-positives.

3.3. WASP (Weighted Average Synthesis Processor)

This processing chain creates for every tile a monthly composite image [10], effectively removing clouds and directional effects from the Sentinel-2 and Venμs inputs.

Originally developed for ESA's Sen2Agri-Toolbox [11] it uses the provided cloud- and aerosol masks in oder to calculate a weighted average centered around a certain day in a month to recieve a Level-3A product with a fixed monthly-interval. Figure 4 shows a coud-free composite image of

WASP-products.

Fig. 4. Cloud free composite image over France on september 2018

The image collection is currently provided on Theia's website for metropolitan France, Belgium and Luxemburg as well as some selected regions around the world but will be extended to all available Sentinel-2 Level-2A tiles as well as to all the sites of Venμs. It forms a basis for studies on vegetation growth as well as the impacts of severe droughts such as this summer 2018 [12].

3.4. Iota2 (Land Cover Map)

Iota2 creates a land cover map for metropolitan France using a random-forest classifier [4]. The Level-3B map is created once per year at the full 10m resolution.

The processing chain was originally developed by the Cesbio institute in Toulouse for Landsat-8 but later adapted to Sentinel-2 providing continuous maps since 2015. Together with Landsat-8 this period is extended to 2014 [13].

3.5. WaterColor/OBS2CO

The newest addition to the existing processing chains will be the reliable detection of water-bodies and estimation of suspended particulate matter (SPM) concentration in inland water-bodies[14]. This is done by using a two-stage unsupervised classification to detect the water in a selected area, correcting sunglints, removing clouds and finally using the reflectance-values to estimate the SPM in the resulting polygon.

Eight zones around the globe were selected for a test-phase to monitor rivers and lakes. The processing chain will combine data from Sentinel-2, Landsat-5/7/8 and Venμs, resulting in the highest coverage possible.

Together with the Spot-World-Heritage (SWH) [15] images distributed by Theia, data from inland rivers of Africa up to the 1980s can be gathered.

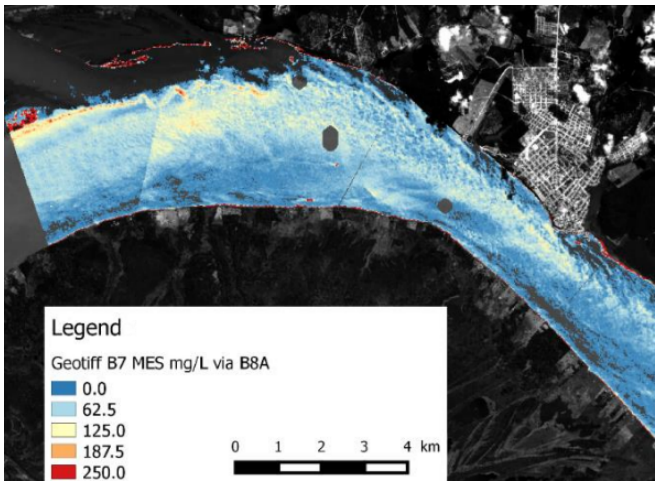


Fig. 5. Estimated concentration for the amazon river using the WaterColor/OBS2CO output for two distinct dates

Figure 5 shows the estimated concentration highlighted for a part of the amazon river in Brazil.

4. CONCLUSION AND FUTURE WORK

The current architecture of Muscate allows for an efficient processing of multi platform imagery. The platform will continue to grow in the future: There have been successful tests of running Muscate on a cloud-based insfrastructure such as AWS [5]. Also the number of products processed daily is steadily growing.

This offers the possibility to distribute Muscate to other scientific and commercial institions that wish to participate in the development of new earth observation services in order to attract more users.

At the same time the processing chains of Muscate add value to the products coming from multiple satellite platforms. In the future, the focus will shift towards improving the quality and accuracy of each of the underlying algorithms:

For MAJA, a study is ongoing about the usage of convolutional neural networks (CNN) for the detection of clouds using Sentinel-2 imagery - Consolidated numbers will be provided by early 2019. Using these improved cloud masks, all processing chains will benefit from the increased accuracy.

In the case of LIS, a Level-3B product is planned, which will perform the snow-detection on the basis of the Level-3A outputs of WASP, thus creating a monthly synthesis of snow cover for central Europe and eastern Canada.

For WASP, the synthesis on water-surfaces, especially in coastal areas shall be improved. This will lead to the possibility to run WASP more reliably on river deltas and around islands.

Finally, the pool of processing chains will keep on growing in the future as well, as shown in section 3.5.

REFERENCES

[1] Nicolas Baghdadi, Marc Leroy, Pierre Maurel, Selma Cherchali, Magali Stoll, Jean-François Faure, Jean-Christophe Desconnets, Olivier Hagolle, Jérôme Gasperi, and Philippe Pacholczyk, "The theia land data centre," in *Remote Sensing Data Infrastructures (RSDI) International Workshop. La grande motte, France*, 2015.

[2] Vincent Lonjou, Camille Desjardins, Olivier Hagolle, Beatrice Petrucci, Thierry Tremas, Michel Dejus, Aliaksei Makarau, and Stefan Auer, "Maccs-atcor joint algorithm (maja)," in *Remote Sensing of Clouds and the Atmosphere XXI*. International Society for Optics and Photonics, 2016, vol. 10001, p. 1000107.

[3] Manuel Grizonnet, S Gascoin, O Hagolle, C L'Helguen, and T Klempka, "Let it snow - operational snow cover product from sentinel-2 and landsat-8 data," in *ESA Living Planet Symposium. Prague, CZ 9-13 May 2016*, 2016.

[4] Jordi Inglada, Arthur Vincent, Marcela Arias, Benjamin Tardy, David Morin, and Isabel Rodes, "Operational high resolution land cover map production at the country scale using satellite image time series," *Remote Sensing*, vol. 9, no. 1, pp. 95, 2017.

[5] J Donadieu, S Baillarin, M Leroy, R Ngo, J Novasan, A Selle, and C L'Helguen, "Muscate - a versatile data and service infrastructure compatible with public cloud computing," in *Proceedings on the conference on Big data from Space 2017, Toulouse, France*, 2017, p. 279ppp.

[6] Pierre-Marie Brunet, "Big data challenges, an insight into the gaia hadoop solution," in *SpaceOps 2012*, p. 1275512. 2012.

[7] Vincent Garcia and Mireille M Paulin, "Peps: Plateforme dexploitation des produits sentinel," in *2018 SpaceOps Conference*, 2018, p. 2614.

[8] Jordi Inglada and Emmanuel Christophe, "The orfeo toolbox remote sensing image processing software," in *Geoscience and Remote Sensing Symposium, 2009 IEEE International, IGARSS 2009*. IEEE, 2009, vol. 4, pp. IV–733.

[9] Olivier Hagolle, Mireille Huc, D Villa Pascual, and Gérard Dedieu, "A multi-temporal method for cloud detection, applied to formosat-2, venμs, landsat and sentinel-2 images," *Remote Sensing of Environment*, vol. 114, no. 8, pp. 1747–1755, 2010.

[10] O Hagolle, M Kadiri, and D Morin, ," http://www.esa-sen2agri.org/wp-content/uploads/resources/technical-documents/Sen2Agri_DDF_v1.2_ATBDComposite.pdf, 2016, Accessed: 13/10/2018.

[11] Nataliia Kussul, Andrii Shelestov, and Andrii Kolotii, "Sen2-agri: Deployment of national sentinel-2 products distribution center in ukraine," 2016.

[12] Oliver Hagolle, Simon Gascoin, Michel Le Page, and Peter Kettig, "A seamless and cloudless sentinel-2 image of france in july 2018," http://www.cesbio.ups-tlse.fr/multitemp/?p=14192, Accessed: 14/10/2018.

[13] Oliver Hagolle and CESBIO, "Land cover map of france for 2014 from landsat8," http://www.cesbio.ups-tlse.fr/multitemp/?p=8009, Accessed: 29/09/2018.

[14] S Yepez, A Laraque, JM Martinez, J De Sa, JM Carrera, B Castellanos, Marjorie Gallay, and JL Lopez, "Retrieval of suspended sediment concentrations using landsat-8 oli satellite images in the orinoco river (venezuela)," *Comptes Rendus Geoscience*, vol. 350, no. 1-2, pp. 20–30, 2018.

[15] J Nosavan, A Moreau, and P Henry, "Spot world heritage: exploring the past," in *Sensors, Systems, and Next-Generation Satellites XXII*. International Society for Optics and Photonics, 2018, vol. 10785, p. 107850T.

doi:10.2760/848593

MACHINE LEARNING FOR CROP TYPE IDENTIFICATION USING COUNTRY-WIDE, CONSISTENT SENTINEL-1 TIME SERIES

Guido Lemoine, Wim Devos, Pavel Milenov, Raphaël d'Andrimont

European Commission, Joint Research Centre, Food Security Unit

ABSTRACT

Recent European Union (EU) legislation for the Common Agricultural Policy (CAP) introduces the concept of checks by monitoring [1], which requires the development of machine learning approaches for country-wide, full season use of the EU's Copernicus Sentinel time series at agricultural parcel level. The excellent consistency of dense Sentinel-1 time series over the EU make these a prime candidate to generate gap-free, consistent feature vectors for the full set of parcel declarations for use in machine learning. We demonstrate the use of tensorflow with a selection of $\sim$ 170,000 arable crop parcels selected from the full agricultural parcel data set over the Netherlands for the 2017 growing season. The prime focus is on identifying parcels for which our machine learning results suggest that their crop class label do not conform with that declared in the aid application. Repetitive tensorflow runs with the 7 major arable crop types result in separating the parcel set in 4.1% non-conform and the remainder (95.9%) conform. We discuss practical implementation details and impact in the checks by monitoring context.

Index Terms— Copernicus, Sentinel-1, SAR, agriculture, crops, CAP, machine learning, tensorflow

1. INTRODUCTION

The requirement to scale checks by monitoring to the full country (region) sample of the area-based aid applications with frequently available Sentinel data time series mandates new methods to handle the large data volumes in a continuous manner [2]. Machine learning provides a set of methods that are especially tailored to find common patterns in labeled deep data stacks and apply the "learned" patterns to new data series to predict its most probable label. The learning process is incremental, i.e. when new labeled series are added as training sets, the newly learned patterns can be (re-)applied to handle a wider range of conditions that are described by the labels. Thus, machine learning can be applied in checks by monitoring both in a reductive approach, i.e. to identify a small set of "outliers" for which the predicted label is different from the declared label, as well as to separate within a labeled category by some specific marker in the temporal signal. In this paper, we focus on the reductive approach.

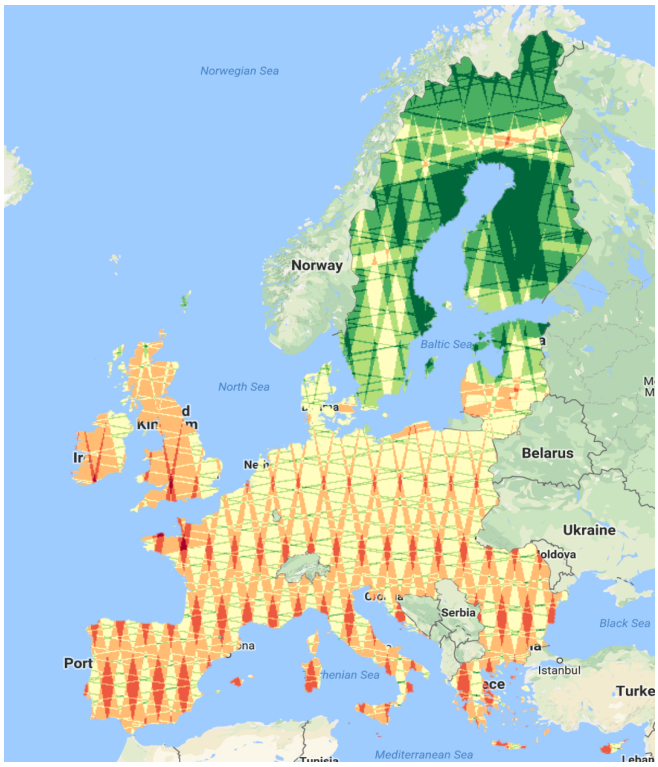


Fig. 1. Sentinel-1 revisit in the period 1-7 April 2018 (1 nominal revisit cycle of the combined Sentinel-1A and -1B). Number of revisits range from 1 (dark red) to 8 (dark green). Mid-latitude EU countries have 4 revisits (light yellow) for most of their territory.

A key difference between machine learning and classical Computer Aided (Photo-) Interpretation (CAPI) methods used in CAP control is that patterns are no longer "learned" by an experienced operator using a limited amount of prepared renderings of pre-selected image series, but by a machine that uses all available signal data over the period of interest. This eliminates a number of weaknesses that are inherent in the CAPI approach, such as the limitations in image availability, the incomplete use of radiometric information across sensor bands, variation due to non-harmonized renderings, operator

bias (and fatigue), etc. Machine learning can handle arbitrary amount of samples and depths of the time series data stack, which would be impractical to do in a CAPI approach.

A (current) drawback in machine learning is the need for consistently sampled, gap-free feature vectors that feed into the learning framework of the method (typically a neural network). Feature vectors are the records that are extracted for each declared agricultural parcel. The elements of the record are the individual signal values in the time series, usually in time order, and often reduced to a single value, usually the arithmetic mean, for all pixels that are included in the feature. "Consistently sampled" does not necessarily mean that a regular, equal interval sampling is required, but whatever sampling approach is chosen needs to be applied, consistently, for all features. In practice, a regular equal interval sampling is preferred. The requirement for "gap-free" (i.e. no missing data) feature vectors sets at national or regional level is a challenge. The orbiting Sentinels acquire imagery over fixed swaths which may cover different parts of the territory. Swaths of neighbouring orbits may overlap, but are acquired at different times, usually several days apart. Sentinel-1 can acquire data in both the descending orbit direction (local morning) as well as the ascending orbit direction (local evening). Sentinel-1 is insensitive to cloud cover and, since it is an active microwave sensor, acquires data independent from solar illumination (i.e. day and night). Acquisition over the EU land mass is at maximum capacity for both descending and ascending orbits (see Fig. 1). A major drawback of Sentinel-2 is cloud cover, which leads to significant and irregularly timed gaps in the consistent cover. Cloud masking combined with mosaicking and gap-filling methods (e.g. time series smoothing) is often insufficient to overcome these problems, esp. over extended cloudy periods. Thus, Sentinel-2 is not a prime source for machine learning approaches, but will play a role in the post-processing of machine learning results.

2. DATA AND METHODS

2.1. TIME SERIES EXTRACTION

For any location in the EU at least 2 images are acquired within the nominal 6 day revisit cycle. Due to orbit overlap, the revisit is actually more frequent than this, esp. towards higher latitudes, where revisit can be up to every day. To create a consistent, gap-free time series, it is sufficient to average the Sentinel-1 intensity values over a pre-defined period (e.g. a week) to extract a feature vector set for use in machine learning. This can be done for both the VV and VH polarization channels. The procedure to create the Sentinel-1 feature vector set currently relies on the use of Google Earth Engine (GEE, [3]), as it is the only "Big Data" repository that provides access to geocoded, calibrated S1 backscattering coefficients at 10 m pixel spacing, and for arbitrary selections. With the recent deployment of the Copernicus Data and Information Access Services (DIAS) instances, it can be expected that compatible European processing capacity will become available in the course of 2018.

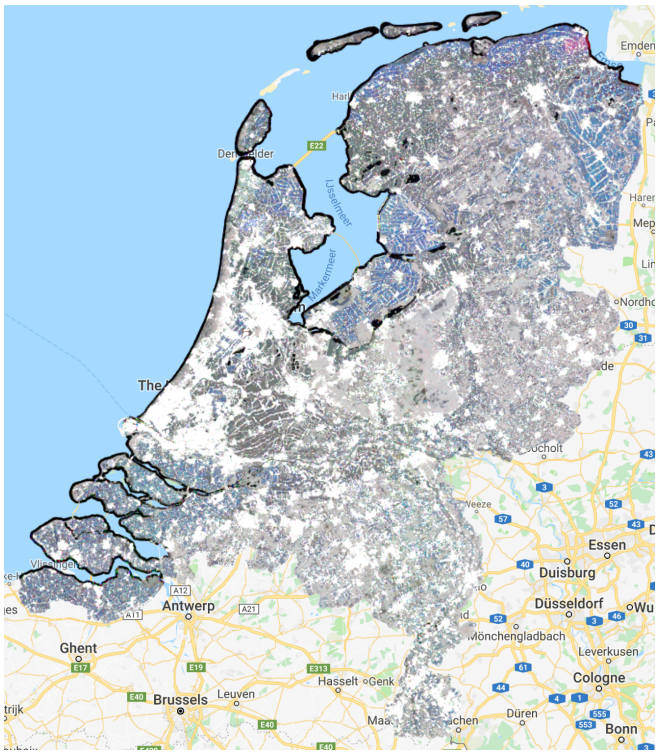
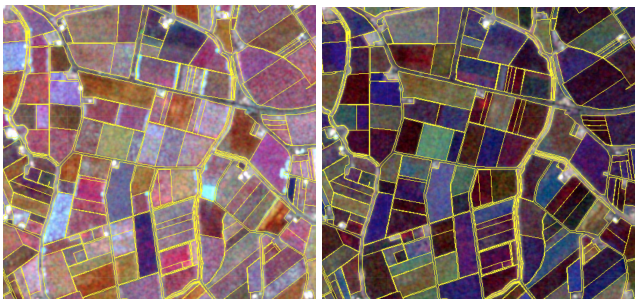


Fig. 2. Example weekly country-wide Sentinel-1 multi-temporal composite for the Netherlands, for the weeks starting on 6 May, 27 May and 17 June 2017 (VV polarization) in the Google Earth Engine JavaScript API. Contains modified Copernicus Sentinel data, 2017.

We illustrate results with a subset of the 2017 Netherlands open access parcel set (NL2017 pdok.nl). The parcel sets are imported as a table asset into GEE. Using standard functions in GEE, weekly averaged images are stacked for the period 1 April - 1 August, 2017 for both VV and VH polarizations (see Figure 2 and 3) after which mean values are extracted for each parcel. Optionally, parcels are buffered with an internal boundary of 10 m, to avoid including edge pixels. The complete set of parcel feature sets can then be exported to a CSV formatted table, retaining the original and calculated feature attributes for each parcel (e.g. including a unique ID and crop code, crop name, area, perimeter, etc.).

Based on the analysis of parcel statistics for the full set, those crop codes for which the summed area coverage is larger than 95% of the overall set are selected (see Table 1). These codes are then grouped, based on the crop category and crop name, into crop classes. Separation in crop classes is partially based on the expectation that these classes have distinct temporal signatures. For instance, silage maize and corn maize will be grouped in one class at this (reductive)



(a) VV polarization (b) VH polarization

Fig. 3. Full resolution extracts of Fig. 2 for an arable crop area west of Oud-Vossemeer (Zeeland, NL) overlaid with the NL2017 agricultural parcel vector. Contains modified Copernicus Sentinel data, 2017.

stage. This results in a set of 170454 parcels with 7 different crop classes.

2.2. MACHINE LEARNING USING TENSORFLOW

For machine learning, we have chosen tensorflow based on its growing reputation as a versatile open source toolkit for a wide range of machine learning problems. However, results reported in this document are likely to be reproducible in other (python based) open source machine learning libraries (theano, scikit-learn, etc.).

Tensorflow is installed by building from source, which optimises the use of specific hardware acceleration features of the platform. The tflearn module is required as ancillary library to run the deep neural network, which consist of 2 layers of 32 nodes and a softmax optimizer, for training and testing.

The parcel attributes that should not be included in the training and testing phase (e.g. area, perimeter, crop name) are removed from the feature set. The set has to be split into a training and testing samples. For large sets ($>$ 100,000 records), we choose a random selection of 20% of the overall set for training. This step is repeated 5 times to produce 5 distinct training sets with their complementary test sets.

3. RESULTS

Single tensorflow runs for the NL2017 record set require less than 5 minutes (100 epochs, 8 core Intel Xeon E3-1505M v6 @ 3.00 GHz, with 64 GB RAM and Quadro M2200 GPU). Training accuracy levels off beyond 80 epochs, and does not significantly increase with higher numbers of epochs. Tensorflow results are produced in the following format:

id,klass,prob0,prob1,...,prob6
000048dd00dda062f4a4,0, 98.39, 0.04, ..., 0.01
0000b3c0dcbe2d3eea3b,6, 0.03, 98.18, ..., 0.12

i.e. for each parcel, which has class label in column klass (class is a reserved word in python/pandas), 7 probabilities are estimated by the trained model, i.e. one for each crop class. The first entry has the highest value for prob0 (98.39%), i.e. predicted class (0) is matching the parcel label (0). The second entry is a clear mismatch (parcel label 6, but the predicted class with the highest probability (98.18%) is 1).

A single confusion matrix can now be created for each run, by accumulating the counts of each matching case (on the matrix diagonal elements) and each mismatch on the relevant off-diagonal element. Assignment is based simply on the maximum probability across the row for each parcel. An example confusion matrix, for a single run, is given in Table 1.

The counts in this confusion matrix are the number of parcels that are assigned to each matrix element. The overall accuracy (OA) for each of the 5 runs ranges between 95.4 and 96.5%. Each parcel is selected once to be part of the 20% training set, and classified 4 times for those cases when the parcel is in the complementary 80% testing set. For each parcel, the join of the individual runs can be generated, i.e. for each unique parcel ID the 4 predicted majority labels can be compared to the parcel label. The total number of parcels for which the majority of predicted labels is not the same as the parcel label is 6954 (out of 170454), i.e. 4.1%, which is more or less the same as $1-OA$ for each individual run. This shows that the method is very robust. For a subset of 3723 parcels all 4 predicted labels are the same and not equal to the declared label. These are prime candidates for further follow up.

Note that a number of parameter settings have been fixed in the reported tests as discussed above. Varying these parameters will have an effect on overall accuracy, for instance, stricter criteria for class probability will lower overall accuracy. Relevant code artifacts for the procedures outlined in this paper can be found in the Appendix of [4].

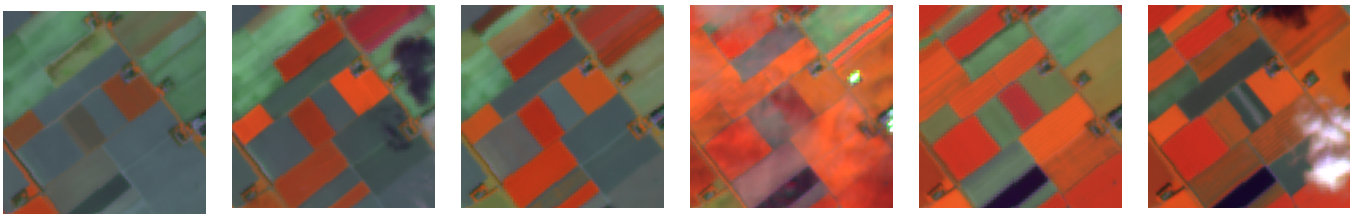
The tabular result can now be categorized to prioritize follow-up activities. From the confusion matrix it can be determined which cases of omission and comission are likely to have relevant impact on compliance to particular CAP support schemes (e.g. permanent grassland measures, [4]). Small and oddly shaped parcels may need to be excluded to reduce noise factors. Re-runs with fine-tuned parameter settings may help in eliminating or precising specific outlier categories. The combination of these analysis results help in defining follow-up inspection, such as the selection of Sentinel-2 imagery (see Fig. 4), generation of specific time series for analysis, and sorting cases that require extending the time series analysis or need to be followed up by a field visit.

4. CONCLUSIONS

Our analysis shows that it is fully feasible to process large amounts of parcel declaration data with deep Sentinel-1 image data stacks using a combination of standard GEE and

Table 1. Confusion matrix for a single tensorflow run for the 7 major arable crops in the NL2017 set (MAI=Maize; POT=Potato; WWH=winter wheat; SBT=Sugar beet; ONI=onions; SBA = spring barely; FLO=flowers). Overall accuracy is 96.1 %.

Crop	MAI	POT	WWH	SBT	ONI	SBA	FLO	sum	PA
MAI	**65260**	374	135	55	52	74	95	66045	98.8
POT	362	**26126**	41	77	25	12	75	26718	97.8
WWH	142	37	**15492**	7	25	125	12	15840	97.8
SBT	134	818	11	**12502**	38	3	67	13573	92.1
ONI	360	86	148	65	**4439**	136	67	5301	83.7
SBA	430	23	316	6	54	**3974**	21	4824	82.4
FLO	203	131	94	331	7	19	**2807**	3592	78.1
sum	66891	27595	16237	13043	4640	4343	3144	**135893**	
UA	97.6	94.7	95.4	95.9	95.7	91.5	89.3		



27 March 2017 | 6 May 2017 | 26 May 2017 | 22 July 2017 | 14 August 2017 | 23 September 2017

Fig. 4. Sentinel-2 time series of false colour chips generated in GEE and centered on a parcel which is labeled as winter wheat but for which tensorflow applied to weekly averaged Sentinel-1 feature vectors predicts onion as crop class. The Sentinel-2 sequence confirms the tensorflow prediction. Contains modified Copernicus Sentinel data, 2017.

tensorflow routines. We have already demonstrated similar results with full 2017 parcel data sets in Denmark and Flemish Belgium, achieving overall accuracies that are well above 90%, and for different crop class mixes. The tests in this study focus on the comparison of declared parcel labels with those predicted by a trained deep neural network. For other schemes, a stratified approach may be preferred over a full country, e.g. for agro-environmental areas that have more complicated cropping parameters. There are many permutations possible, though, for instance, approaches that may try to separate the more heterogeneous classes (e.g. grassland), compare distinct crop development by phenological progress and/or agronomically relevant factors (e.g. soil type), etc. The tools are rather generic and leave it up to the practitioner to device the test set-up and working hypothesis.

The overall accuracy of 96.1% produced for the NL2017 data set is excellent, as it exceeds the desired accuracy (95%) that was expressed by EU Member States to consider checks by monitoring as an efficient alternative to the current, sample based, on the spot controls.

Machine learning methods have considerable potential in other CAP control domains (e.g. Land Parcel Identification Systems (LPIS) quality control, physical block segmentation, etc.). A key advantage of using the consistent Sentinel-1 time series is that models trained with 2017 data would, in principle, be useful to predict class labels for 2018 data, even before the definitive declaration data would be available. Geographical and temporal transfer learning is one of our ongoing research topics.

REFERENCES

[1] European Union. “Commission implementing regulation (EU) 2018/746 of 18 May 2018 amending Implementing Regulation (EU) No 809/2014 as regards modification of single applications and payment claims and checks”. Off. J. Eur. Union, 2018, 61, L125-1-7.

[2] Devos, W.; Fasbender, D.; Lemoine, G.; Loudjani, P.; Milenov, P.; Wirnhardt, C. “Discussion Document on the Introduction of Monitoring to Substitute OTSC”, Technical Report, European Commission: Brussels, Belgium, 2017, ISBN 978-92-79-74279-8.

[3] Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. “Google Earth Engine: Planetary-scale geospatial analysis for everyone.” Remote Sens. Environ. 2017, 202, 1827.

[4] d’Andrimont, R.; Lemoine, G.; van der Velde, M., “Targeted Grassland Monitoring at Parcel Level Using Sentinels, Street-Level Images and Field Observations”, Remote Sensing, 10, 2018, 8, 1300, doi: 10.3990/rs10081300

FROM BIG COPERNICUS DATA TO BIG INFORMATION AND BIG KNOWLEDGE: THE COPERNICUS APP LAB PROJECT

Konstantina Bereta[1], *Hervé Caumont*[2], *Ulrike Daniels*[5], *Daems Dirk*[3], *Manolis Koubarakis*[1], *Despina-Athanasia Pantazi*[1], *George Stamoulis*[1], *Sam Ubels*[4], *Valentijn Venus*[4], *Firman Wahyudi*[4]

[1]National and Kapodistrian University of Athens, Greece; [2]Terradue Srl, Italy; [3]VITO, Belgium; [4]RAMANI B.V., The Netherlands; [5]AZO Anwendungszentrum GmbH, Germany;

ABSTRACT

We discuss the challenges of big Copernicus data and how our project Copernicus App Lab has dealt with them. Copernicus App Lab takes data from the land monitoring, global land and atmosphere services and makes it available on the Web and the Cloud using semantic technologies to aid its take up by mobile developers. We also discuss lessons learned for information retrieval, database and knowledge management research in the context of Copernicus.

Index Terms— big data, semantic technologies, linked geospatial data, Earth observation, satellite remote sensing

1. INTRODUCTION

Copernicus data is a paradigmatic case of big data which is acquired by the Sentinel satellites and contributing missions, together with in-situ data from sensors on the ground, at sea, or in the air. Copernicus is at the forefront of all big data challenges: *volume, velocity, variety, veracity,* and *value*. The H2020 project Copernicus App Lab (http://www.app-lab.eu/) targets the *volume* and *variety* challenges of Copernicus data, and it follows the path of previous research projects TELEIOS, LEO, and MELODIES, funded by FP7 ICT. Copernicus App Lab goes beyond these projects in the following important ways. First, it develops a software architecture that enables on demand access to big Copernicus data using the well-known OPeNDAP framework and the geospatial ontology-based data access system Ontop-spatial [2]. Now users and application developers do not need to download data or learn the details of sophisticated data formats for EO data. All they need to develop is an ontology describing the data they are interested in and R2RML mappings that capture the correspondence between the ontology and the data sources containing the data. Using traditional approaches, application developers would have to implement different clients/adapters in their applications corresponding to the different file formats their data is in, in order to process the data. Instead of implementing custom code, they can use the functionalities of the Ontop-spatial mapping language for all data sources regardless of their formats.

Secondly, it brings computing resources close to the data by making the Copernicus App Lab tools available as Docker images that are deployed in the Terradue cloud platform as cloud services. The platform allows application developers to access Copernicus data and carry out massively parallel processing without the need to download the data and carry out the processing locally. Thirdly, it enables search engines like Google to treat datasets produced by Copernicus as "entities" in their own right and store knowledge about them in their internal knowledge graph. In this way, search engines will be able to answer sophisticated users questions which is beyond the reach of modern search engines today. A more detailed description of the Copernicus App Lab project is given in [4].

This work has received funding from EU Horizon2020, Grant Agreement nr. 730124.

2. THE COPERNICUS APP LAB ARCHITECTURE

Figure 1 presents the conceptual architecture of the *Copernicus integrated ground segment* and the Copernicus App Lab software architecture.

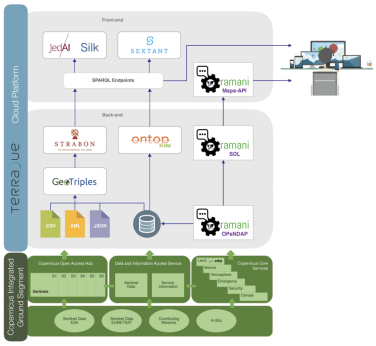


Fig. 1. The Copernicus integrated ground segment and the Copernicus App Lab software architecture

In the lower part of the figure, the Copernicus *data sources* are shown. These are Sentinel data from ESA, Sen-

tinel data from the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), satellite data from contributing missions and in-situ data. The next layer makes Copernicus data and information available to interested parties in three ways: via the Copernicus Open Access Hub, via the Copernicus Core Services and via the Data and Information Access Service (DIAS).

All the software components of the project run in the Terradue cloud platform (https://www.terradue.com/portal/). The platform allows cloud orchestration, storage virtualisation, and virtual machine provisioning, as well as application burst-loading and scaling on third-party cloud infrastructures. Within the Terradue cloud platform, the developer cloud sandbox service provides a platform-as-a-service (PaaS) environment to prepare data and processors. It has been designed with the goal to automate the deployment of the resulting EO applications to any cloud computing facility that can offer storage and computing resources (e.g., AWS).

In Copernicus App Lab, access to Copernicus data and information can be achieved in two ways: (i) by downloading the data via the Copernicus Open Access Hub or the Web sites of individual Copernicus services, and (ii) via the popular OPeNDAP framework (https://www.opendap.org/) for accessing scientific data. In the first case (workflow on the left part of the two top layers of Figure 1), the downloaded data should then transform into RDF using the tool GeoTriples [7] or scripts written especially for this task. GeoTriples enables the transformation of geospatial data stored in raw files (shapefiles, CSV, KML, XML, GML and GeoJSON) and spatially-enabled RDBMS (PostGIS and MonetDB) into RDF graphs using well-known geospatial vocabularies such as the Open Geospatial Consortium (OGC) standard GeoSPARQL [10]. The performance of GeoTriples has been studied experimentally [7] using large publicly available geospatial datasets. It has been shown that GeoTriples is very efficient especially when its mapping processor is implemented using Apache Hadoop.

After Copernicus data has been transformed into RDF, it can be stored in the spatiotemporal RDF store Strabon [6, 3]. Strabon can store and query linked geospatial data that changes over time. It has been shown to be the most efficient spatiotemporal RDF store available today using the benchmark Geographica in [5, 3]. Copernicus data stored in Strabon may also be interlinked with other relevant data. To do this in Copernicus App Lab, we use the interlinking tools JedAI and Silk. JedAI is a toolkit for entity resolution and its multi-core version has been shown to be scalable to large datasets [9]. Silk is a well-known framework for interlinking RDF datasets which we have extended to deal with geospatial and temporal relations [11].

The *novel way of accessing Copernicus data and information* in Copernicus App Lab is captured by the workflow on the right part of the two top layers of Figure 1, and it is based on the popular *OPeNDAP framework* for accessing scientific data. The *streaming data library (SDL)* implemented by RAMANI communicates with the OPeNDAP server and receives Copernicus services data as *streams*. In this way, SDL enables on-the-fly computation of spatial and temporal aggregations (e.g., a longterm moving average that is often of interest to EO applications). The SDL is accessible through a list of APIs that are enhanced with an API ontology, which directly links to a function ontology that describes the offered functionality and analytics. This ontology describes calls and responses of the API and assists users in determining valid functions over different data types. The API responses are provided as JSON-LD with direct references to the semantics of the returned variables, allowing easier interpretation. OPeNDAP and SDL are installed and configured by VITO on a virtual machine running on the VITO hosted PROBA-V mission exploitation platform (https://proba-v-mep.esa.int), which has direct access to the data archives of the Copernicus global land service. The installation of OPeNDAP was done using Docker and access to the Copernicus global land and PROBA-V datasets via OPeNDAP is realised by mounting the necessary disks on the virtual machine.

One of the main contributions of Copernicus App Lab is the extension of the ontology-based data access system Ontop-spatial [1] with OPeNDAP support. Ontop-spatial is a system that connects to existing geospatial databases and creates virtual semantic graphs on top of them using ontologies and mappings, without downloading files and transforming them into RDF. Mappings encode how we map relational data to RDF terms. As we describe in [2], the new version of Ontop-spatial is able to connect to non-relational external data sources (e.g., APIs like OPeNDAP) and enable users to pose GeoSPARQL queries on top of them without the need of importing the data in relational databases.

Finally, data can be visualized using the tools Sextant [8] or Maps-API (https://ramani.ujuizi.com/maps/index.html). Sextant is essentially a GIS for linked geospatial data. It enables users to build layered maps consisting of geospatial data made available in various formats (e.g., KML, GML etc.) and SPARQL or GeoSPARQL endpoints. The Maps-API is similar to Sextant in terms of visualization functionality, but it takes its data from SDL and it cannot deal with linked geospatial data sources accessed by SPARQL or GeoSPARQL.

All tools are open source and they are available on the following Web page: http://kr.di.uoa.gr/#systems

3. A COPERNICUS APP LAB CASE STUDY

A simple case study, which demonstrates the functionality of the Copernicus App Lab software, involves studying the "greenness" of Paris. This can be done by relating "greenness" features of Paris using geospatial data sources such as OpenStreetMap and relevant Copernicus datasets from the land monitoring service of Copernicus, which are the leaf-area index dataset (global), the CORINE land cover dataset

(pan-European) and the Urban Atlas dataset (local).

Leaf area index (LAI) is a dimensionless quantity that characterizes plant canopies and it is defined as the one-sided green leaf area per unit ground surface area in broadleaf canopies (https://en.wikipedia.org/wiki/Leaf_area_index). The *CORINE land cover dataset* covers 39 EU countries (https://land.copernicus.eu/pan-european/corine-land-cover). Land cover is characterized using a 3-level hierarchy of classes with 44 classes in total at the 3rd level. The *Urban Atlas dataset* (https://land.copernicus.eu/local/urban-atlas/view) provides land use and land cover data for European urban areas, and it covers 800 urban areas in 28 EU countries.

In addition to the above datasets, our case study utilizes data from OpenStreetMap and the global administrative divisions dataset GADM. OpenStreetMap is an open and free map of the whole world constructed by volunteers. GADM (https://gadm.org/) is an open and free dataset giving us the geometries of administrative divisions of various countries.

The first task of any case study using the Copernicus App Lab software is to develop INSPIRE-compliant ontologies for the selected Copernicus data. The *INSPIRE directive* (https://inspire.ec.europa.eu/) aims to create an interoperable spatial data infrastructure for the EU, to enable the sharing of spatial information among public sector organizations and better facilitate public access to spatial information across Europe.

Once all the ontologies are defined, we can easily translate them into RDF using a custom script. Then, they can be stored in Strabon and be queried jointly in interesting ways. For example, assuming appropriate PREFIX definitions, the following GeoSPARQL query asks for the LAI values of the area occupied by the Bois de Boulogne park in Paris.

```
SELECT DISTINCT ?geoA ?geoB ?lai WHERE {
  ?areaA osm:poiType osm:park.
  ?areaA geo:hasGeometry ?geomA . ?geomA geo:asWKT ?geoA .
  ?areaA osm:hasName  "Bois de Boulogne"^^xsd:string .
  ?areaB lai:lai ?lai .
  ?areaB geo:hasGeometry ?geomB . ?geomB geo:asWKT ?geoB .
  FILTER(geof:sfIntersects(?geoA, ?geoB))}
```

Similarly, in Figure 2, we have used Sextant to build a temporal map that shows the "greenness" of Paris, using the datasets LAI, GADM, CORINE land cover, Urban Atlas and OpenStreetMap. We show how the LAI values (small circles) change over time in each administrative area of Paris (administrative areas are delineated by magenta lines) and correlate these readings with the land cover of each area (taken from the CORINE land cover dataset or Urban Atlas).

All RDF datasets and ontologies that have been discussed above are freely available at: http://kr.di.uoa.gr/#datasets.

The "greenness of Paris" case study can also be developed using the workflow on the right in the Copernicus App Lab software architecture of Figure 1. In this case, the datasets can be queried using Ontop-spatial and visualized in Sextant without transforming any datasets into RDF. In this case, the developer has to write R2RML mappings expressing the correspondence between a data source and classes/properties in

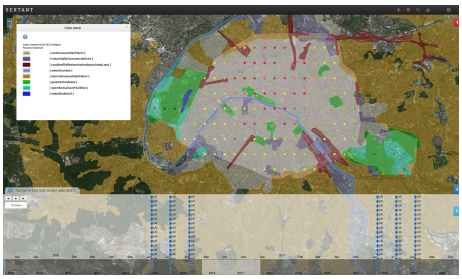


Fig. 2. The "greenness" of Paris

the corresponding ontology. An example of such a mapping is provided below (in the native mapping language of Ontop-spatial which is less verbose than R2RML).

```
mappingId opendap_mapping
target lai:{id} rdf:type lai:Observation .
       lai:{id} lai:lai {LAI}^^xsd:float;
                time:hasTime {ts}^^xsd:dateTime .
       lai:{id} geo:hasGeometry _:g .
       _:g geo:asWKT {loc}^^geo:wktLiteral .
source SELECT id, LAI, ts, loc FROM (ordered opendap
       url:https://analytics.ramani.ujuizi.com/
       thredds/dodsC/Copernicus-Land-timeseries-global
          -LAI%29/readdods/LAI/)  WHERE LAI > 0
```

In this mapping, the `source` is the LAI dataset, provided through the RAMANI OPeNDAP server of the Copernicus App Lab software stack. The dataset contains observations that are LAI values, the time and location for each observation. Operator `Opendap` retrieves this data and populates a virtual SQL table with schema `(id,LAI,ts,loc)`. Because of the fact that the `Opendap` operator is implemented as an SQL user-defined operator, it can be embedded into any SQL query. In the above mapping, we also refine the data we want to be translated into virtual RDF terms by adding a filter to the query to eliminate negative or zero LAI values. The `target` part of the mapping encodes how the relational data is mapped into RDF terms.

4. LESSONS LEARNED AND FUTURE CHALLENGES

The use of OPeNDAP offers better data access capabilities specifically for application developers that are not experts in EO, and thus it a clear benefit. OPeNDAP and SDL provide streaming data to the user and have some significant advantages over the OGC Web Coverage Service standard which is already offered by VITO. First of all, from a data provider perspective, OPeNDAP is easier to use, as it is able to deal with a wider variety of grid types. Furthermore, OPeNDAP can be easily extended with different conventions, allowing for easier integration of different datasets and without overhead like file conversion. Also, OPeNDAP enables the loose coupling of different Copernicus data sources into one data model, providing the user easy access through a single access

point that uses this data model. Finally, when using the Web Coverage Service, there is limited possibility to obtain client-specific parts of the datasets (one is limited to, for example, a bounding-box). In contrast, OPeNDAP allows for the caching of datasets by serialization based on internal array indices.

The most innovative aspect of using Ontop-spatial in Copernicus App Lab is its ability to give access to Copernicus data through the OPeNDAP framework. When data is stored in a database connected with Ontop-spatial, DBMS optimisations and database constraints are applied and query plans are optimized. This does not happen in the case where Ontop-spatial retrieves data on-the-fly from OPeNDAP, since data is preprocessed before it gets translated into virtual triples using Ontop-spatial. However, if we want to access Copernicus data that gets frequently updated, the virtual RDF graphs approach is useful as it avoids the repeated translation steps that have to be done by the data provider. For costly operations (e.g., spatial joins of complex geometries), it is better to materialize the data. To improve performance, we have implemented a caching mechanism so that queries that result in the same API calls for a time window w, whose length is a configurable parameter, can get cached data. We also extended our system with the ability to integrate other kinds of data e.g., HTML tables and social media data (e.g., twitter, foursquare). In our current work, we are developing further optimisation techniques to improve performance.

Participants of the ESA Space App Camp (www.app-camp.eu/) that was organised in September 2017 and 2018 had the opportunity to use the Copernicus App Lab technologies to implement demo applications. The objective was to make EO data, particularly from Copernicus, accessible to a wide range of businesses and citizens. The developers of the winning teams AiR and URBANSAT used Copernicus App Lab tools to access and integrate data from different sources.

It is important to point out that an approach very similar to our projects TELEIOS, LEO, Melodies and Copernicus App Lab is currently been taken by the CREODIAS platform, a cloud-based one-stop shop for all Copernicus satellite data and imagery, as well as the Copernicus services information (https://creodias.eu/). The CREODIAS approach is limited though since only *metadata* of Copernicus datasets are available as linked data and can be queried by relevant discovery tools. We, on the other hand, allow users to also make information and knowledge extracted from Copernicus data available as linked data. In this way it can be combined with other linked datasets (public or private) and enable the development of applications by mobile developers easily. In this way we contribute to the *value* dimension of big Copernicus data.

Google has recently activated the beta version of its dataset search (https://toolbox.google.com/datasetsearch), where the datasets that are indexed using *schema.org* (https://schema.org/), as proposed by Google, show up. We have followed these guidelines and annotated all the datasets used in the use case of Section 3, and made them available at the following link: http://kr.di.uoa.gr/#datasets. We have also recommended that the same practice is followed by the Copernicus services we have worked with (land monitoring, global land and atmosphere services). Our current work focuses on designing an extension to the community vocabulary *schema.org* appropriate for annotating EO data in general and Copernicus data in particular, by extending the class *Dataset* with subclasses and properties which cover the EO dataset metadata defined in relevant OGC standards.

5. SUMMARY

The Copernicus App Lab project targets the variety and volume challenges, and has developed a novel software stack that can be used to develop applications using Copernicus data even by developers that are not experts in EO. We presented a case study developed using the Copernicus App Lab software stack and discussed lessons learned and future plans.

REFERENCES

[1] K. Bereta and M. Koubarakis. Ontop of geospatial databases. In *ISWC*, 2016.

[2] K. Bereta and M. Koubarakis. Creating virtual semantic graphs ontop of big data from space. In *BiDS*, 2017.

[3] K. Bereta, P. Smeros, and M. Koubarakis. Representation and querying of valid time of triples in linked geospatial data. In *ESWC*, 2013.

[4] K. Bereta, H. Caumont, U. Daniels, D. Dirk, M. Koubarakis, D.-A. Pantazi, G. Stamoulis, S. Ubels, V. Venus, and F. Wahyudi. The copernicus app lab project: Easy access to copernicus data. In *EDBT*, 2019.

[5] G. Garbis, K. Kyzirakos, and M. Koubarakis. Geographica: A benchmark for geospatial RDF stores. In *ISWC*, 2013.

[6] K. Kyzirakos, M. Karpathiotakis, and M. Koubarakis. Strabon: A semantic geospatial DBMS. In *ISWC*, 2012.

[7] K. Kyzirakos, D. Savva, I. Vlachopoulos, A. Vasileiou, N. Karalis, M. Koubarakis, and S. Manegold. GeoTriples: Transforming Geospatial Data into RDF Graphs Using R2RML and RML Mappings. *Journal of Web Semantics*, 2018.

[8] C. Nikolaou, K. Dogani, K. Bereta, G. Garbis, M. Karpathiotakis, K. Kyzirakos, and M. Koubarakis. Sextant: Visualizing time-evolving linked geospatial data. *Journal of Web Semantics*, 35 (1), 2015.

[9] G. Papadakis, K. Bereta, T. Palpanas, and M. Koubarakis. Multi-core meta-blocking for big linked data. In *SEMANTICS*, 2017.

[10] M. Perry and J. Herring. GeoSPARQL - a geographic query language for RDF data. OGC Implementation Standard, 2012.

[11] P. Smeros and M. Koubarakis. Discovering spatial and temporal links among RDF data. In *LDOW*, 2016.

AUTOMATIC IMAGE DATA ANALYTICS FROM A GLOBAL SENTINEL-2 COMPOSITE FOR THE STUDY OF HUMAN SETTLEMENTS

Christina Corbane[1], Panagiotis Politis[2], Pieter Kempeneers[1], Martino Pesaresi[1], Dario Rodriguez[1], Vasileios Syrris[1], Pierre Soille[1]

[1]European Commission, Joint Research Centre

[2]Arhs Developments S.A., Luxembourg

ABSTRACT

The Copernicus Sentinel-2 mission offers new opportunities for mapping human settlements over large areas and for the update and improvement of the Global Human Settlement Layer. This paper presents the fully automated processing workflows tailored for large scale mapping of built-up areas from Sentinel-2 imagery. The first results provide insights into the capabilities gained either by analyzing separately optimally selected S2 tiles or by the processing a best-available-pixel composite over large areas.

Index Terms— Global Human Settlement Layer, built-up areas, Sentinel-2, pixel composite, JRC Big Data Platform

1. INTRODUCTION

The successful launch of the Copernicus Sentinel satellites marked a new era in the Big Data landscape and stirred the need for the development of operational image processing workflows that produce actionable, trusted, and robust information for different application areas based on free and open data. Mapping and monitoring of human settlements at a global scale is one particular application area that can greatly benefit from the Earth Observation data revolution brought by the Sentinels:

The two first missions, Sentinel-1 (S1) and Sentinel-2 (S2) operational since October 2014 and December 2015, respectively, provide free time series suitable for monitoring built-up areas changes at global scale. Sentinel-1 is designed as a constellation of two synthetic aperture radar (SAR) satellites, namely Sentinel-1A (launched in April 2014) and Sentinel-1B (launched in April 2016), offering a full systematic coverage of the land surface at a global level in the Interferometric Wide swath mode every six days. With such characteristics, Sentinel-1 gives the possibility to provide up-to-date global information on the status and evolution of human settlements and allows regular updates of built-up areas. Sentinel-2 satellites A and B, with the Multi Spectral Imager (MSI) instrument provide a 5-day revisit, 10 m pixels in visible bands: specifications which cover a number of human settlements mapping requirements. The complementarity of the two sensors can be used to compile a joint cloud-free global image database at a fine spatial resolution for mapping human settlements.

In 2016, the first map of human settlements (GHS_S1) to be fully derived from a global coverage of Sentinel-1 data was produced in the framework of the Global Human Settlement Layer (GHSL) project of the European Commission [1]. Two main components were key to the success of this Big Data challenge: 1) the advanced machine learning technology used for the automatic information extraction and which builds on the Symbolic Machine Learning (SML) classifier originally designed to deal with big data scenario and 2) the versatility of the Joint Research Centre Earth Observation Data and Processing Platform (JEODPP) [2] that allowed the selection, download, storage and mass processing of 6,721 S1 scenes used in the production of the GHS_S1 layer and the associated global mosaic [3] .

The latest developments presented in [1] in terms of the computationally efficient SML classifier combined with the growing capacity of the JEODPP to deploy consolidated information extraction workflows and the opportunity to leverage on the systematic coverage of Sentinel-2 are of great interest for the purpose of human settlements at a global scale [2]. The potential and added-value of Sentinel-2 data for improving high-resolution human settlement mapping was demonstrated in [4] in a pilot study covering selected areas in Italy. Scaling up the methods to cover large geographical areas involves new challenges related to: 1) the adaptation of the workflows to the characteristics of the large and heterogeneous coverage of Sentinel-2 imagery, 2) the need to optimize the selection the images to address the access, storage and computations requirements, 3) the automation of the information extraction methods while allowing flexibility in the choice of the area to be processed and efficiency in I/O.

The present work proposes two automated workflows tailored for large scale mapping of built-up areas from Sentinel-2 imagery. The workflows take into account the need to reduce the computations requirements while still enabling the coverage of large areas such as full countries, continents or even all landmass. The underlying idea is to provide solutions for automatic information extraction from Sentinel-2 data feeds that can work both in cloud environments or standard clusters.

2. OPTIMIZED GLOBAL COVERAGE OF SENTINEL-2 INPUT DATA

2.1. Optimized selection of S2 tiles

Since February 2018, S2 mission started fully exploiting the two satellite units of the constellation and delivering over 4 Terabytes of daily data on the Copernicus portals. For the purpose of mapping of human settlements at a global scale, it is needed to select from the millions of available S2 images, the best subset that covers the full landmass and minimizes the cloud coverage and the amount of data to be stored and processed.

The selection was performed at the 100 x 100 km tiles according to the Military Grid Reference System (MGRS) in which the S2 images are provided by the European Space Agency (ESA). The selection process itself was based on a floating forward search of all available quicklooks and cloud maks [5]. The quicklooks represent a spatial and spectral subset of the level 1C products. At each iteration the most significant quicklook image was included in order to obtain the minimum number of images required for a cloud free composite.
The maximum number of overlapping images was constrained to five. As a result, less than 5% of the available S2 images in 2017 were selected. On average, 3.16 overlapping images were needed for a cloud free global land cover composite (see FIGURE 1). Due to a lack of snow mask, an important number of selected images were covered with snow. Therefore, the selection process was repeated for latitudes above 45 degrees North, selecting only images acquired during summer season.

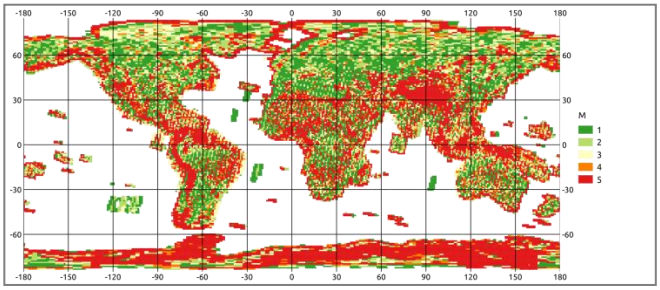


FIGURE 1. Number of overlapping image tiles (M) in the optimal subset obtained from the selection algorithm with max(M)=5.

2.2. Generation of a global cloud-free image composite

Based on the selection of quicklook images as discussed in section 2.1, the level 1C products at full spatial and spectral resolution were downloaded. A maximum composite was then calculated, based on the maximum normalized difference vegetation index (NDVI) for each pixel (see FIGURE 2).

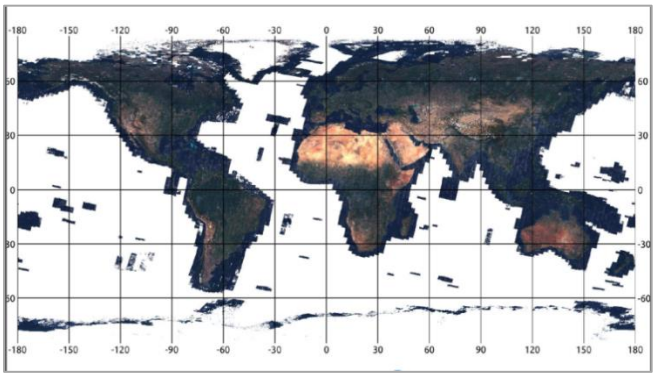


FIGURE 2. World composite based on selected Sentinel-2 quicklooks.

2.3. Atmospheric correction of input data

All 92,985 downloaded tiles were stored on the JEODPP and atmospherically corrected using the Sen2Cor L2A processor (version 2.5.5; European Space Agency. http://step.esa.int/main/third-partyplugins-2/sen2cor/ (accessed May 2018), which performs topographic correction and transforms top-of-atmosphere reflectance (TOA) to bottom-of-atmosphere reflectance (BOA). Scene classification and cloud masks are produced for each scene in the Sen2Cor process to allow for cloud and shadow masking prior to further analysis. On the basis of the scene classification, the percentage of cloud /shadow coverage over land was calculated.

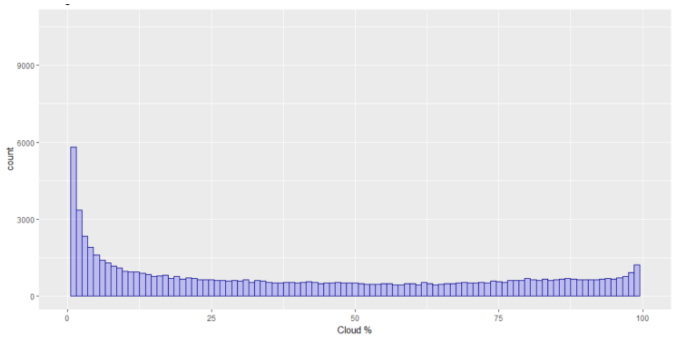


FIGURE 3. Distribution of the number of S2 tiles by percentage of cloud/shadow pixels over land derived from L2A scene classification.

Figure 3 shows the distribution of the number of tiles by percentage of cloud/shadow pixels over land. It shows that despite the quality check and the selection criteria, there are around 1,480 selected tiles with 100% cloud/shadow coverage over land. This is related either to the persistent cloud coverage (e.g. over mountainous areas, tropical zones) and to the non-consideration of the cloud shadows and the land surface in the selection schema that is based on image quicklooks and the poor quality of the cloud masks in vector format delivered with the imagery. A better cloud mask as well as the introduction of a new mask indicating cloud shadows could greatly improve the results.

3. PROCESSING FLOWS FOR BUILT-UP AREAS EXTRACTION FROM SENTINEL-2

The GHSL production workflow builds on the Symbolic Machine learning (SML) method that was designed for remote sensing big data analytics. The SML schema is based on two relatively independent steps:

(1) Reduce the data instances to a symbolic representation (unique discrete data-sequences);

(2) Evaluate the association between the unique data-sequences subdivided into two parts: *X* (input features) and *Y* (known class abstraction derived from a learning set).

In the application proposed here the data-abstraction association is evaluated by a confidence measure called ENDI (evidence-based normalized differential index) which is produced in the continuous [-1, 1] range. Details on the SML algorithm and its eligibility in the framework of big data analytics may be found in [6]. This classification technique has been successfully applied for the processing of Landsat data records of the past 40 years and for generating the first GHSL multi-temporal global product (GHS-Landsat) [1].

3.1. Tile-based processing workflow

A first proof-of-concept demonstrated the added-value of S2 data in improving global high-resolution human settlement mapping [4]. In the current study, the initial algorithm proposed and which builds on the SML classifier is extended to exploit the key features of S2 data: i) the availability of four 10 m spatial resolution bands (B2-Blue, B3- Green, B4- Red and B8- Near Infrared), ii) the availability of six bands at 20 m resolution especially in the Near Infrared and Shortwave Infrared (B5, B6, B7, B8a in Near Infrared and B11, B12 in Shortwave Infrared), iii) the output classification of Sen2cor that can be used for a stratified learning of built-up areas by landcover class.

The following features (*X*) derived from Sentinel-2 are used for the classification of the Sentinel-2 image with the SML approach: i) Spectral features: the three 10 m resolution and the seven 20 m bands , ii) Textural features: a textural feature derived from the brightness (corresponding to the maximum of the visible bands at 10 m) by applying the Pantex methodology [7]. The textural feature is used for refining the output confidence layer by eliminating overdetections, especially roads and open spaces identified as built-up. The learning set (*Y*) is based on the built-up as derived from the GHSL-Landsat. The rough classification output of the Sen2Cor is used during the associative analysis for stratifying the learning set of built-up derived from GHSL-Landsat. This allows tailoring the training set to the image under processing especially in the presence of clouds or cloud shadows and hence allows reducing commission and omission errors. The output confidence is further refined using a global annual composite of maximum Normalized Difference Vegetation Index (NDVI) derived from Bands B4 and B8 of all S2 images. This global layer was calculated in Google Earth Engine using TOA S2 images acquired in 2017. The diagram in Figure 4 presents a simplified version of the workflow for the classification of S2 image tiles. The processing chain is implemented using a massively parallel workflow that runs at tile level. The output confidences of overlapping and redundant are then tiled and mosaicked hence achieving reduced computation time, allowing easy replacement of image tiles in case of availability of better quality data and ensuring continuity across reference years in the case of updating the product specifications.

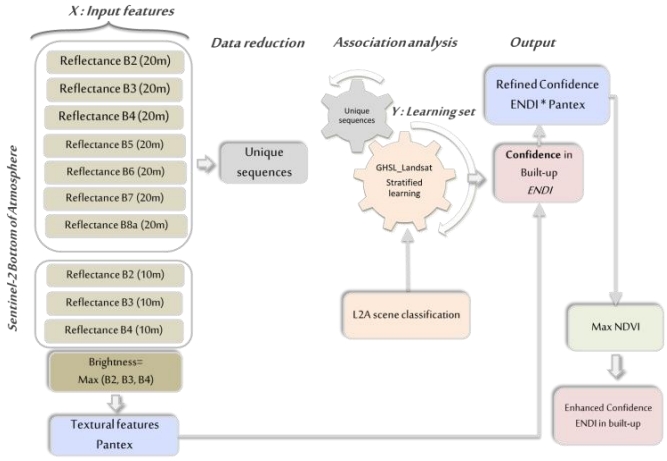


FIGURE 4. Tile-based fully automated workflow for built-up areas extraction from S2 surface reflectance data.

3.2. Composite-based processing workflow

An alternative workflow has been also developed for processing of pixel-based image composites as opposed to scene (tile)-based image processing. The workflow aims at exploiting the best-available-pixel composite by offering a novel opportunity to generate built-up information products that are spatially contiguous over large areas and in a manner that is dynamic, transparent, systematic, repeatable, and spatially exhaustive. The main differences with respect to the tile-based workflow are the following:

- The applicability to both TOA and BOA input data,
- The exclusion of the L2A scene classification from the learning schema because of its irrelevance in the case of the pixel-based composite,
- The use of the 10 meter bands as input, instead of all 10 and 20 m bands, as a compromise between memory efficiency and the need to cover large areas (i.e. equivalent to 5 x 5 S2 images tiles of 100 x 100 km each).

To avoid artifacts in the output confidence layer due to the arbitrary partitioning of the composite, the workflow is executed with a block processing approach including an

doi:10.2760/848593

overlap of 25% across neighbouring blocks. The outputs confidence layers from overlapping blocks are then merged using the average operator.

4. FIRST RESULTS

4.1. Visual assessment of the results

The workflows were tested in a pre-operational setting on large areas covering China, Italy, France and selected cities in Asia and Africa. The figures below show the output confidence layer of built-up areas obtained from the processing of 6,068 S2 tiles covering China (Figure 5). A close view over Beijing shows artifacts in the tile-based processing workflow due to the artificial partitioning of S2 images into footprints (Figure 4b). Despite the mosaicking of the results, these artifacts can still be observed especially in the case where two adjacent tiles are acquired in two different seasons with significant differences in the density of the vegetation coverage.

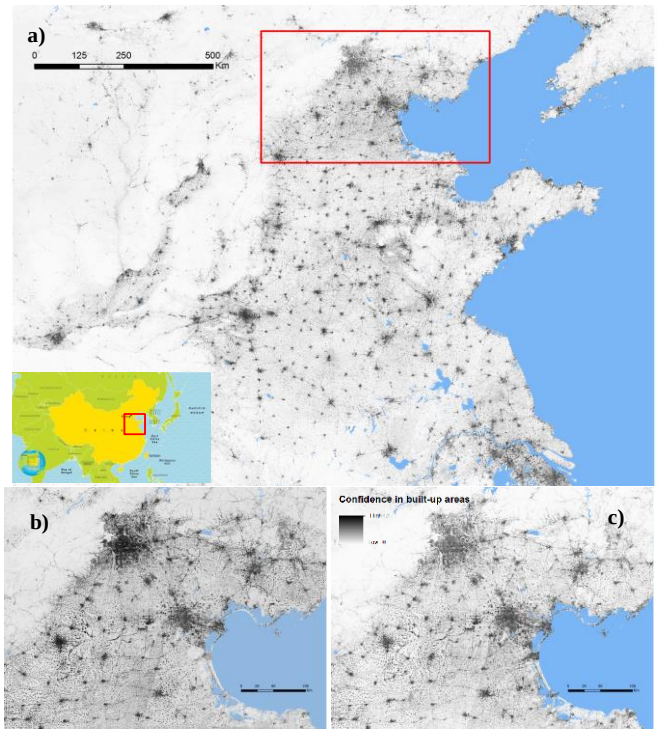


FIGURE 5. A) Results of built-up areas extraction in China shown in terms of confidence measure (ENDI rescaled in the range [0,1]). B) Close view of the output of the tile-based workflow over Beijing compared to C) the output from the composite based workflow.

4.2. Visual assessment of the results

The two workflows were compared in terms of performance. A large scale test for assessing the performance has been has been implemented for the same extent in China, covering a total area of 6,210,000 Km^2. The processing was accomplished using a conventional cluster, consisting of 16 processing nodes (E5-2650v2@2.60 GHz) with a total amount of 256 GB of RAM. The operating system is CentosOS 6.9. Memory usage constraints bounded the number of concurrent jobs to a total of 2 jobs for the composite workflow, resulting in a total processing time of ~12 hours. With 10 concurrent jobs, the tile-based workflow was completed in 15 hours. The results of the test are summarized in Table 1. They show the suitability of both workflows for the processing of large areas and give indications on the scalability potentials of the methods.

TABLE 1. Performance assessment of the two workflows

	Tile Based workflow	Composite based workflow
Input	1865 S2 tiles (100 x 100 km tiles)	276 blocks (150x150 km blocks)
Processing time	15 h	12 h
Number of concurrent jobs	10	2
RAM requirements per job	22 GB	120 GB

5. CONCLUSION AND OUTLOOK

In this paper two workflows for large scale automatic extraction of built-up areas from Sentinel-2 imagery were presented. Both methods build on the SML classifier, but are tailored to the processing of either single tiles or pixel composites of S2 tiles. The results for both methods are promising, suitable for information retrieval from big volumes of S2 data and offer new prospects for large scale mapping of built-up areas from S2 data. The combination of outputs from both methods is foreseen for the mapping of human settlements at the global level in view of updating the GHSL.

REFERENCES

[1] C. Corbane *et al.*, "Mass processing of Sentinel-1 and Landsat data for mapping human settlements at global level," in *Proc. of the BiDS'17*, 2017, pp. 52–55.

[2] P. Soille *et al.*, "A versatile data-intensive computing platform for information retrieval from big geospatial data," *Future Gener. Comput. Syst.*, vol. 81, pp. 30–40, 2018.

[3] V. Syrris, C. Corbane, and P. Soille, "A global mosaic from Copernicus Sentinel-1 data," in *Proc. of the BiDS'17*, 2017, pp. 268–271.

[4] M. Pesaresi, C. Corbane, A. Julea, A. J. Florczyk, and V. Syrris, "Assessment of the added-value of Sentinel-2 for detecting built-up areas in the frame of the Global Human Settlement Layer," *Remote Sens.*, 2015.

[5] P. Kempeneers and P. Soille, "Optimizing Sentinel-2 image selection in a Big Data context," *Big Earth Data*, vol. 1, no. 1–2, pp. 145–158, 2017.

[6] M. Pesaresi, V. Syrris, and A. Julea, "A New Method for Earth Observation Data Analytics Based on Symbolic Machine Learning," *Remote Sens.*, vol. 8, no. 5, p. 399, May 2016.

[7] M. Pesaresi, A. Gerhardinger, and F. Kayitakire, "A Robust Built-Up Area Presence Index by Anisotropic Rotation-Invariant Textural Measure," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 1, no. 3, pp. 180–192, Sep. 2008.

NEAR-REAL TIME DATA MANAGEMENT AND PROCESSING SYSTEM TO DEVELOP AND VALIDATE SPACE WEATHER SERVICES

A.F. Mulone[1], M. Casti[1], R. Susino[2], R. Messineo[1], E. Antonucci[2], G. Chiesura[1], D. Telloni[2], R. De March[1], E. Magli[3], A. Bemporad[2], G. Nicolini[2], S. Fineschi[2], F. Solitro[1], M. Martino[1]

[1]ALTEC S.p.A., Corso Marche 79, 10146 Torino, Italy

[2]INAF-OATo, Via Osservatorio 20, 10025 Pino Torinese, Torino, Italy

[3]Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

ABSTRACT

This paper describes the Heliospheric Data Center (HDC), a near-real time data management and processing system, designed and implemented to provide space weather services. The main system goal is to reduce the time between the space weather services definition and their activation in production environment. This goal is achieved providing several key elements to users: tens of integrated data sources from past and current missions, easy data integration through enriched standard data model, big data technologies used to overcome data management and processing challenges and integrated auxiliary tools and dataset for performing product validation. The first system version was developed within the framework of the Heliospheric Space Weather Center project, resulting from a joint effort between ALTEC S.p.A. and INAF-OATo, both located in Turin (Italy), for providing medium and short-term forecast of geo-effective space weather events, such as the coronal mass ejections (CMEs).

Index Terms— data store, data management, data processing, metadata, data model, neural networks, space weather, forecast

1. INTRODUCTION

Space weather refers to the environmental conditions due to the Sun activities that can influence the functioning and the reliability of space borne and ground-based systems and services or endanger property or human health. For this reason, it is important to be able to predict such phenomena, in order to limit possible damages. Space weather phenomena propagate from the Sun to the Earth, involving ambient plasma, the Earth magnetosphere, ionosphere and thermosphere. Therefore, signals related to them are observable in different datasets acquired by several space and ground-based instruments. Moreover, depending on the considered dataset, the evidence of the detected solar phenomena has different time-scales. As a result, the main features that a system design for space weather forecasting should have are the capability to manage several data types and reduced processing times.

Within this framework, we developed the HDC aiming at two main objectives: consolidate and evolve the heliospheric data center initially set up with the SOHO data coming from the ESA-approved SOLAR and develop a Heliospheric Space Weather Center for forecasting impacts of solar disturbances in the heliosphere and on the Earth's magnetosphere.

Management of different data products is one of the most important features of the HDC, which is hosted at ALTEC S.p.A., in Turin. Data products are different in format and in availability and they are stored in different repositories accessible using different protocols. These heterogeneous data shall be processed and accessible to scientists, so their metadata represent a key factor for the Center. In this respect, SPASE data model has been selected for a double reason: it allows to collect all useful information related to different data products (time series, images, etc.), and it permits to exchange products with other centers.

The management framework used in HDC, i.e., the ALTEC Space Data Processing (ASDP), is characterized by flexibility and extensibility. As a matter of fact, this framework is easily extensible in terms of: software that can be integrated, supported metadata and processing resources. It can be deployed on a distributed environment (cloud) too.

In ASDP for HDC, three different processing pipelines have been integrated and can run at the same time: remote sensing (medium-term forecast), in situ and neural networks (short-term forecast). In particular, each one of them is fed using datasets acquired by different on-going space missions. All generated products are stored and their metadata are described using a data model.

Medium-term forecast (by less than ~2 days) is based on remote sensing observations of the Sun and the heliosphere. Short-term forecast services (by less than ~2 hours) process in situ heliospheric observations at the Sun-Earth Lagrangian point L1. Short-term forecast is executed observing the magnetic helicity and also using neural networks.

doi:10.2760/848593

2. SYSTEM ARCHITECTURE

The architecture of the data management system (see Fig.1) is complex and consists of data stores to archive input data, a metadata data store configured in high availability and processing data stores to prepare data. The input archive is implemented through object storage technology. Moreover, due to its level of flexibility, we choose to base the metadata repository on Elasticsearch. Finally, the implementation of the processing data store is strictly dependent on the applications. A product manager component hands all different data stores and offers a transparent data access service.

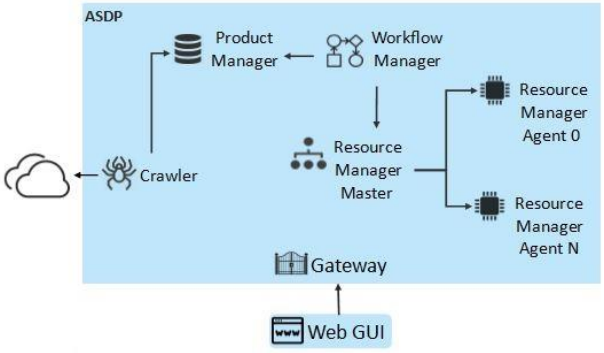


Fig. 1 System Architecture

ALTEC defined and developed a framework with the main aim to process a big amount of data allowing a seamless connection between the collected information and the analyses performed by end users. This is the ASDP environment.

2.1 ALTEC Space Data Processing (ASDP)

The ALTEC Space Data Processing (ASDP) is a distributed data processing framework designed for providing a flexible system capable to handle and process a large variety and amount of data.

ASDP is not a self-standing solution but it allows integrating both existing and new coded algorithms, enabling automatic processing of large datasets and complex pipelines. It enables to organize data in the most suitable domain data store in order to be ready for complex analysis. Innovative analytics algorithms can be easily activated in order to mine data and extract relevant information.

ASDP takes advantage of containers technologies. This simplifies its deployment in any distributed environments, allows runtime expansion of the system, and eases the integration tests. ASDP uses:

- AKKA framework for messaging and cluster managing [2];
- Docker as container technology [3];
- Elasticsearch as metadata repository [4];
- Apache Cassandra for database storage of the products [5];
- Apache Spark as computation framework, specifically used for advanced data analysis [6];
- Jupyter notebooks, written in Python language, as user interface to execute this analysis [7].

ASDP containers are deployed through docker-swarm technology that is a clustering and scheduling tool for Docker containers. With Swarm, IT administrators and developers can establish and manage a cluster of Docker nodes as a single virtual system.

3. NEAR REAL TIME DATA MANAGEMENT

3.1 Data Flow

Data flow starts with data crawling and ingestion. The crawler is the component that queries remote repositories for new data products and triggers internal ingestion and processing pipelines.

Remote repositories are heterogeneous and store different data formats. The crawler shall check whether data have already been ingested in ASDP; if they have not, the products will be downloaded, otherwise they will be skipped. The crawling can be enabled, disabled and scheduled for each type of data with different polling periods.

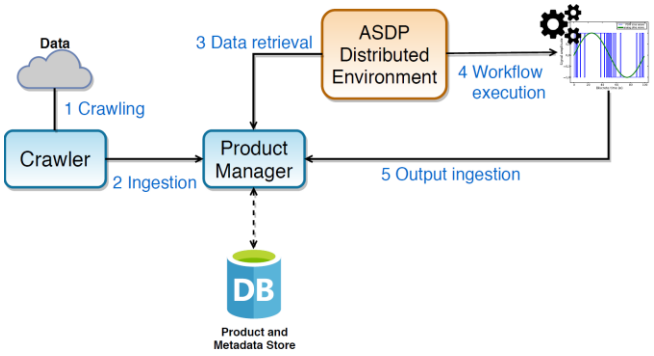


Fig. 2 Data Flow

Crawled products are remote sensing and in situ data. In particular, the in situ measurements relevant for the HDC are those acquired by the following space instruments: DSCOVR Faraday Cup (FC), DSCOVR Magnetometer (MAG), WIND/MFI, STEREO/Plastic, WIND/SWE, ACE/Mag and ACE/Swepam. The retrieval of these data is integrated in the Heliospheric Data Center, except for ACE/Mag and ACE/Swepam. On the other hand, remote sensing data ingested by the HDC consist of the images acquired by the solar coronagraph on-board of SOHO satellite, i.e., LASCO C2/C3.

All ingested products are retrieved by querying their related online archives, which are published in different server typologies like FTP, HTTP, etc., and are available with different frequencies according to the related instrument sampling time. For instance, remote sensing data acquired by LASCO C2/C3 are available with a frequency of about 12 minutes. On the other hand, in situ instruments are

characterized by a higher sampling frequency and new measurements are available each minute, as in the case of DSCOVR data.

Products are stored in an archiving file system shared by all the nodes of ASDP processing cluster, while metadata are stored in Elasticsearch where it is possible to enable and schedule snapshots of the repository.

3.2 Data integration and data model

Data integration is obtained through a flexible and scalable design and the implementation of a set of interoperable data stores. After their download, data products are ingested in ASDP and their metadata are extracted. Metadata is a key factor for organizing solar dataset. HDC uses the SPASE data model, which is the most widespread data model in virtual observatory.

SPASE is a set of terms and values along with the relationships between them that allows describing all the resources in a heliophysics data environment. SPASE aims at unifying and improving existing Space and Solar Physics data models. SPASE divides the heliophysics data environment into a limited set of resource types. A key resource type is Numerical Data.

Describing completely a Numerical Data resource requires other types of Resources, namely Observatory, Instrument, Person, and Repository, whose names are self-explanatory, and each one of them has its own set of attributes. Often, numerical data is presented in prepared images (gif or jpeg), and such presentations are referred to as Display Data resources. The other data related resource types are: Catalogue, listing events; Annotation, enabling experts to comment on data products; and Granule, describing individual files within another resource. Other types of resources include Document which can contain narratives or supporting information; Service that provide software to use data resources; Repository for storage locations; and Registry for metadata collections. Resource descriptions and the links therein are intended to make the Resource useful to scientific users.

In order to better describe the output generated by the pipelines, we integrated SPASE with ESPAS data model. As a matter of fact, ESPAS allows describing better the processing, while SPASE describes better the structure of the data.

ESPAS aims at building the e-Infrastructure necessary to support the access to observations, the modelling and prediction of the near-Earth space environment extending from the Earth's atmosphere up to the outer radiation belts. Data is described using a ‘scientific-friendly’ approach.

3.3 Data access and visualization

HDC has a dedicated interface for scientists. The homepage provides information related to the latest detected CME event and shows the latest data related to the Sun, the Heliosphere and the Earth magnetosphere.

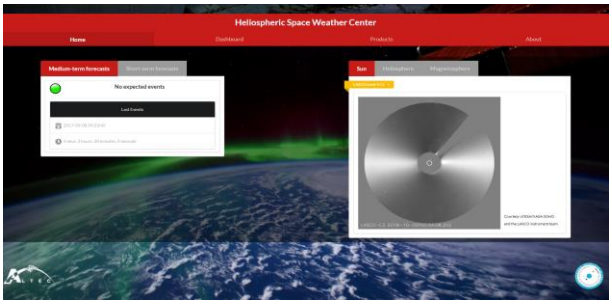


Fig. 3 HDC Home Page

For each event, it is possible to view the physical parameters computed by pipelines. Moreover, the interface provides the opportunity to download all data ingested in HDC and pipelines products, in order to execute off-line data analysis. Data can be accessed by product type, generation and sensing date (for ingested product).

The operation interface, instead, allows to check if the HDC pipelines are running correctly and interact with the system, if necessary. Moreover, it enables modifying the ingestion product pipelines altering the polling period for download or stopping the ingestion of a specific data product. The operator can modify metadata of each data products and can look for products using a flexible pseudo-SQL language.

4. DATA PROCESSING

Three different pipelines are integrated in ASDP: remote sensing, in situ and neural networks. Pipelines are developed using different technologies that have been integrated thanks to the flexibility of ASDP. Furthermore, data processed are different in format and multiplicity.

Remote sensing pipeline exploits the SolarSoft system, an IDL based system built from libraries related to several solar missions. The near real time ingestion of the latest available fits file starts the pipeline. The first step is the image calibration. The calibrated image is then analyzed and, if an event of interest is detected, an algorithm retrieves the CME physical parameters and computes its propagation time until it reaches L1 and the probability of impact on Earth.

The in situ pipeline starts with the ingestion of data in real time. For each run, the pipeline is fed with the latest 28 days data, which are processed in order to forecast solar geoeffective events.

The remote sensing and the in situ pipeline outcomes are then compared in order to improve the final event forecast.

The recurrent neural network pipeline uses TensorFlow library and python scripts [8]. This pipeline processes real time data; the recurrent network, trained using WIND data, stores a temporal model of solar wind employed to perform predictions.

4.1. Deep Learning Integration

The integration of neural network pipeline has involved python code, TensorFlow library, integration with

PostgreSQL database and the creation of a specific docker container [9]. The pipeline processes the in situ data acquired by the DSCOVR payloads, stored in database, in order to predict the DST geomagnetic index value that will be measured within few hours. HDC processes neural network TensorFlow trained models to tell if a weak, moderate or intense storm is about to happen in 2, 4 or 8 hours (all thresholds are customizable).

4.2. Scientific pipelines integration

The integration of the scientific pipelines within the HDC takes advantage of the ASDP flexibility. The pipelines are composed by different blocks that execute algorithms coded in different languages as python and IDL. Moreover, input data are read from different data stores: file system and database.

5. VALIDATION RESOURCES

Several tests have been performed in order to integrate algorithms and validate the scientific results. Tests have been executing exploits the features of ASDP framework that allow the creation of automatic test writing a common XML files.

Furthermore, the validation of the scientific results requests the development of suitable tools and selected different auxiliary data sets related to near real time measurements and historical data collected within CME catalogues available online.

5.1. Auxiliary data

HDC ingests data necessary for a backward verification of the pipelines outcome too. In particular, in order to verify whether the detected event reached the Earth's magnetosphere within the predicted time, the DST index is downloaded and compared with a threshold value. The availability frequency of this data is about one hour.

Furthermore, SOLAR – SoHo Long-term Archive, which contains SOHO data approved by ESA – is integrated and managed in HDC.

5.2. Validation tool

HDC flexibility allows the creation of tools to validate pipelines. These tools can exploit external framework and can be coded in different language (i.e. python). Moreover, tools can use different data products.

The remote sensing pipeline validation is based on the comparison between the obtained results and the CME catalogues available online. In this respect, two different catalogues have been considered: LASCO CME catalogue and CACTUS catalogue. The former is a manual catalogue and, for this reason, it is unfortunately out of date. On the other hand, CACTUS is an automated catalogue, which reports less information but it is available in near-real time.

In situ pipeline outcome is validated considering the DST geomagnetic index. New data is available every hour; therefore, after its ingestion, ASDP executes the validation pipeline.

6. FUTURE PROSPECTS

The capacity to manage different data products, different data stores, different frameworks and libraries allows improvements in terms of algorithms and services.

Additional data archives of both existing missions (e.g. SDO) and new missions that will be available in the next years (i.e. Metis, PROBA-3) will be managed at HDC. Radio-telescope data are under investigation before trying to integrate it in the Center.

New version of both remote sensing and in situ algorithms will be integrated as soon as possible. Moreover the crosscheck between these two pipelines shall be developed. The neural networks pipeline shall be improved to run in near-real time using DSCOVR data products. Furthermore, new neural networks pipelines could be developed and integrated. Now, this pipeline processes numerical data but it could run on image products.

From a technical point of view, the ASDP cluster could be deployed using Kubernetes instead of Docker Swarm which is currently used [10]. Another important feature to be addressed in the next future is the improvement of the backup techniques.

7. ACKNOWLEDGEMENTS

The authors would like to acknowledge Prof. Mauro Messerotti, Senior Advisor for Space Weather, INAF Science Directorate.

8. REFERENCES

[1] M. Casti, S. Fineschi, R. Messineo, E. Antonucci, A.F. Mulone, A. Bemporad, A. Fonti, R. Susino, F. Filippi, D. Telloni, F. Solitro, G. Nicolini et M. Martino, "Data Integration of Remote Sensing and In situ Data from Several Solar Space Missions for Space Weather Services", *Proceedings of the Conference on Big Data from Space (BiDS'17)*, Toulouse, France, 2017, pp. 359-362. doi: 10.2760/383579.
[2] Akka, https://github.com/akka/akka [October 22, 2018]
[3] Docker, https://www.docker.com/what-docker [October 22, 2018]
[4] Gormley, C. and Tong, Z., Elasticsearch: The Definitive Guide, O'Reilly Media, 2015.
[5] Apache Cassandra, http://cassandra.apache.org/ [October 22, 2018]
[6] Apache Spark, https://spark.apache.org/ [October 22, 2018]
[7] Project Jupyter, http://jupyter.org/ [October 22, 2018]
[8] TensorFlow, https://www.tensorflow.org/ [October 22, 2018]
[9] PostgreSQL, https://www.postgresql.org/ [October 22, 2018]
[10] Kubernetes, https://kubernetes.io/ [October 22, 2018]

MAPPING THE SURFACE DEFORMATION AT NATIONAL SCALE THROUGH THE AWS CLOUD IMPLEMENTATION OF THE S1 P-SBAS PROCESSING CHAIN

Zinno I. (1), Bonano M. (1,2), Casu F. (1,3), De Luca C. (1), Manunta M. (1), Manzo M. (1), Onorato G. (1), and Lanari R. (1)

IREA-CNR, Via Diocleziano 328, 80124 Napoli, Italy (1)
IMAA-CNR, C. da S. Loja 85050 Tito Scalo, Italy (2)
IREA-CNR, Via Bassini, 15 - 20133 Milano, Italy (3)

ABSTRACT

This work is aimed at showing that an effective integration of advanced remote sensing methods and new ICT technologies allows the full exploitation of large volumes of Earth Observation (EO) data, contributing to deeply investigate the Earth System processes and to address new challenges within the Big Data scenario.
In particular, we present an automatic pipeline implemented within the Amazon Web Services (AWS) Cloud Computing platform for the interferometric processing of large Sentinel-1 (S1) multi-temporal SAR datasets, aimed at analyzing Earth surface deformation phenomena at wide spatial scale. The developed processing chain is based on the advanced DInSAR approach referred to as Small BAseline Subset (SBAS) technique, which allows producing, with centimeter to millimeter accuracy, surface deformation time series and the corresponding mean velocity maps from a temporal sequence of SAR images. The implemented solution addresses the aspects relevant to i) S1 input data archiving; ii) interferometric processing of S1 data sequences, performed in parallel on the AWS computing nodes through both multi-node and multi-core programming techniques; iii) storage of the generated interferometric products. The experimental results are focused on a national scale DInSAR analysis performed over the whole Italian territory by processing 18 S1 slices acquired from descending orbits between March 2015 and April 2017, corresponding to 2612 S1 acquisitions.

Index Terms— DInSAR, P-SBAS, Sentinel-1, Deformation time series, Cloud Computing

1. INTRODUCTION

The Big Data paradigm is bringing revolutions in many scientific fields. A very relevant one is represented by Earth Observation (EO) where it is opening promising investigation opportunities and facing new challenges. Among several applications, the EO techniques have already shown to be very powerful for the detection and analysis of surface deformations due to their characteristics of large spatial coverage, high accuracy and cost effectiveness. The investigation of Earth surface deformation phenomena provides critical insights into several processes of great interest for science and society, especially from the perspective of further understanding the Earth System and the impact of human activities. In this scenario, Differential Synthetic Aperture Radar (SAR) Interferometry (DInSAR) is regarded as one of the key EO methods for its ability to investigate surface displacements affecting large areas of the Earth with centimeter- to millimeter-level accuracy [1].

Basically, the DInSAR technique allows generating spatially dense deformation maps by exploiting the phase difference (interferogram) between pairs of complex SAR images. Among several advanced DInSAR algorithms, a widely used approach is the Small BAseline Subset (SBAS) technique [2], which generates surface deformation time series and the corresponding mean deformation velocity maps by exploiting interferograms characterized by small temporal and spatial baselines between the acquisition orbits, in order to mitigate the decorrelation phenomena. The SBAS algorithm has already proven its effectiveness to investigate surface displacements with millimeters accuracy in different scenarios, such as volcanoes, tectonics, landslides, anthropogenic induced land motions, and it is capable to perform analyses at different spatial scales and with multi-sensor data [3]-[6].

Currently, the DInSAR scenario is characterized by a huge availability of SAR data acquired during the last 25 years, comprising the long-term C-band European Space Agency (ESA) archives (e.g., ERS-1, ERS-2, and ENVISAT), the RADARSAT-1 and RADARSAT-2 C-band data sequences, those provided by the L-band ALOS-1 and ALOS-2 sensors and by the X-band generation of SAR sensors, such as the COSMO-SkyMed (CSK) and TerraSAR-X constellations. Moreover, a massive and ever increasing data flow is nowadays supplied by the C-band Sentinel-1 (S1) constellation of the European Copernicus Programme that is composed of two twin SAR satellites, Sentinel-1A (S1-A) and Sentinel-1B (S1-B), which have been launched on April 2014 and April 2016, respectively. The main S1 acquisition mode on land, referred to as

Interferometric Wide Swath (IWS), implements the Terrain Observation by Progressive Scans (TOPS) technique, specifically designed for interferometric applications: indeed, the nominal footprint of the S1 TOPS mode extends for about 250 km, thus allowing the constellation to operate with a global coverage acquisition strategy. The S1 interferometric revisit time is either 6 days; moreover, thanks to both its intrinsically small spatial and temporal baselines, the S1 constellation is specifically oriented to DInSAR applications, thus it naturally fits the SBAS approach characteristics. Furthermore, the whole S1 archive is available with a free and open access policy, thus easing the data access and enlarging the scientific community interested in its exploitation, opening new research perspectives to understand Earth surface deformation dynamics at global scale. It is also evident that the S1 EO constellation, providing nowadays about 10 TB per day, is significantly contributing to move EO toward the Big Data "V" concept [7].

By considering the above described DInSAR scenario, it is clear that the development of effective solutions able to properly deal with the transfer, the storage, and, above all, the processing of such a huge SAR data flow is strongly needed. Within the framework of the advanced DInSAR processing, a parallel algorithmic solution for the SBAS approach, referred to as Parallel Small BAseline Subset (P-SBAS) [8], which implements a complete advanced DInSAR processing chain and is able to exploit distributed computing architectures, has been recently developed. P-SBAS permits to generate, in an automatic and unsupervised way, advanced DInSAR products by taking full benefit from parallel computing architectures, such as cluster, grid and cloud computing infrastructures [9][10].

We present in this work the implementation of an interferometric processing chain based on the P-SBAS approach dedicated to the processing of S1 data within the Amazon Web Services (AWS) environment. It supports both multi-node and multi-core scheduling policies and permits to generate surface deformation time series from large volumes of S1 data, thus allowing us to perform national-scale DInSAR analyses. It is worth noting that the proposed S1 data interferometric processing chain deals with all the aspects relevant to the i) S1 input data archiving, ii) their processing and iii) the storage of the computed interferometric products. In particular, we developed an automatic pipeline, which includes the download of the S1 input data from the AWS S3 archive towards the computing nodes, the launch and the completion of the P-SBAS DInSAR processing and, finally, the transfer of the generated results to the S3 long-term storage.

As experimental results we show in this paper the national scale DInSAR analysis performed over the Italian territory by processing 18 S1 slices (where a slice indicates an area on the ground of about 250x250 km^2) acquired from descending orbits during the March 2015 - April 2017 time span, corresponding on the whole to 2612 S1 IWS SLC images. Such an analysis was entirely carried out by exploiting AWS storage and computing resources.

2. THE S1 P-SBAS PIPELINE WITHIN THE AWS CLOUD ENVIRONMENT

The key issue related to the interferometric processing of S1 data is dealing with their huge volume, in terms of storage, data transfer and computational efficiency. First of all, a typical interferometric SAR dataset including hundreds of images can reach several hundreds of GB; moreover, throughout all the P-SBAS processing, a very large volume of intermediate and final products is generated (whose size is at least one order of magnitude larger than the input dataset). In order to cope with these issues, we selected, among the wide range of resources and services offered by the AWS environment, those better answering the requirements of the P-SBAS processing.

The first problem relevant to large interferometric data processing is their transfer towards the computing nodes that, in the case of S1 datasets made of hundreds of images, can take a significant time. The solution is to use computing resources in proximity to the data, i.e. connected to the data archive through dedicated access bandwidth. Therefore, we created a public S1 data archive on the Amazon Simple Storage Service (S3), the long-term storage service of Amazon, containing all the Sentinel-1 data acquired over Italy, both from ascending and descending orbits, which is updated with new acquisitions every time new data on the Copernicus Open Access Hub are available. Transferring data from the S3 storage to the instances of the Amazon Elastic Compute Cloud (EC2) is very fast thanks to the dedicated connection guaranteeing very high I/O performances.

To achieve good computational efficiency, among the available AWS EC2 instances, we selected the i3.16xlarge. Such a machine is, on the one hand, a storage-optimized instance and so it is very well suited to sustain the intensive Input/Output workload of the S1 P-SBAS processing. On the other hand, it is characterized by very good computational performances, indeed it allows splitting the parallel jobs of the processing among 64 vCPUs. Obviously, since the S1 P-SBAS processing encompasses several steps that are very different both from the algorithmic point of view (image registration, interferogram generation, interferogram filtering, phase unwrapping, time series generation) and for the type of input data (bursts, images, interferograms, stacks of interferometric products), the number of tasks that run in parallel on different CPUs is specifically designed for each step of P-SBAS taking into account the multi-threading implementation, the RAM occupation and the I/O workload.

In order to process all the 18 S1 slices acquired over Italy from descending orbits, we developed a completely automatic processing chain that starts querying the S1

archive on the S3 storage and downloading the data on the computing node and then launches the P-SBAS processing. Once it is completed, the final results are transferred and saved again on the S3 storage and the processing of a new slice is activated.

3. EXPERIMENTAL RESULTS

We present in this section the results of our national scale S1 P-SBAS DInSAR analysis carried out, through the AWS EC2 cloud platform, on the Italian territory. In particular, for such an analysis, we processed 18 S1 slices, covering an overall area of more than 300,000 km^2, as shown in Fig. 1. We exploited in total 2612 S1 IWS SLC images acquired from descending orbits within the time interval March 2015 – April 2017.

It is worth noting that the processing of a S1 slice lasts approximately 18 hours, thus suggesting that the overall analysis can be performed within one day by exploiting, in parallel, 18 i3.16xlarge AWS instances, with a cost of less than 1800 USD if on-demand instances are used.

Fig. 2 shows the overall mean deformation velocity map obtained by merging the 18 geocoded mean deformation velocity maps relevant to the considered S1 slices, computed with a spatial resolution of about 80x80 m^2. According to the color bar depicted in the Fig. 2, green color represents areas that are stable in terms of surface displacements, whereas the red and blue colors stand for negative and positive deformation velocity values, which correspond to an increase and decrease of the Line Of Sight (LOS) sensor-to-target distance, respectively. Moreover, in Fig. 2, we highlight four zones particularly relevant from the deformation viewpoint, which are delineated by the red rectangles (a), (b), (c) and (d) and are zoomed in the insets below. In particular, in Fig. 2a we report a sketch of the mean deformation velocity map relevant to central Italy. It is evident a deformation pattern characterized by a very large extent, which is associated to the seismic sequence that struck central Italy in 2016. Moreover, Fig. 2a highlights the presence of two lobes, characterized by both negative and positive LOS-projected displacement signals, respectively, which reveal a complex SW-NE oriented deformation pattern. Furthermore, we report the displacement time series relevant to two pixels (labeled as p1 and p2 in fig. 2a and marked by white stars), located in the maximum co-seismic deformation area. They clearly show the LOS-projected deformation signal associated to the occurred seismic sequence (see the red and blue vertical dashed lines that identify both the Amatrice and Visso/Norcia events). Fig 2b shows the mean deformation velocity map relevant to the Napoli Bay area. It is worth noting the significant deformation pattern corresponding to the area of the Campi Flegrei caldera, with the time series of a pixel located in the maximum deformation area clearly highlighting the uplift phenomena that have characterized this area during the 2015-2017 time period. Fig 2c represents the mean

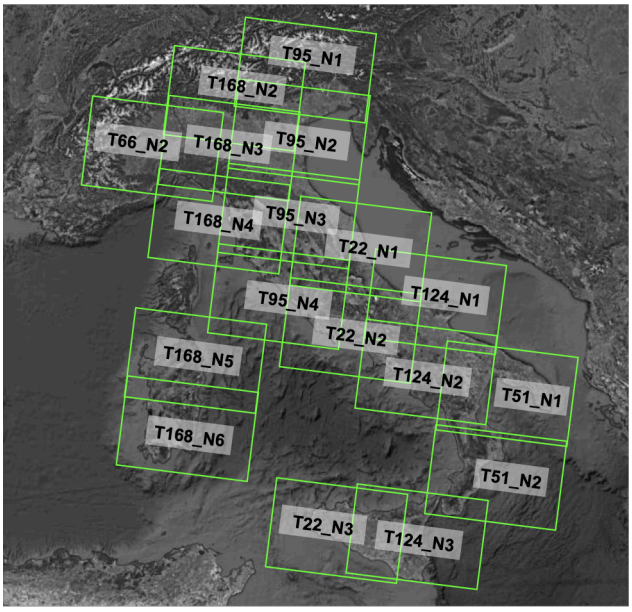


Fig. 1. Representation of the S1 slices acquired from descending orbits and processed through the S1 P-SBAS processing chain implemented within the AWS environment.

deformation velocity map associated to the extended slope movements affecting the little town of Plataci (southern Italy) and corresponding time series of displacement for two pixels located in the maximum deformation areas on the opposite mountainsides. Finally, in Fig. 2d the mean deformation velocity map of the area of Gioia Tauro (Calabria, Italy) is depicted, together with the displacement time series relevant to two pixels located on a highway and in the harbor area, both showing significant subsidence behavior

4. CONCLUSION

EO data archives are expanding at an unprecedented speed, both in size and variety, creating the opportunity to boost the study and the knowledge of the Earth System dynamics. In this paper we presented a Cloud Computing pipeline for carrying out national scale interferometric analyses from large multi-temporal SAR datasets acquired by the Sentinel-1 constellation. We dealt with the main relevant issues of Big Data processing, by including also the storage of both the input SAR images and the generated interferometric value added products.

To show the potentiality of the presented processing chain, we presented a national scale DInSAR analysis accomplished over the Italian territory by processing 2612 S1 IWS SLC data (the overall dataset size is almost 12 TB) acquired from descending orbits within the March 2015 - April 2017 time span. In particular, we produced the mean surface deformation velocity map of the whole Italian peninsula with a spatial resolution of about 80x80 m^2 and

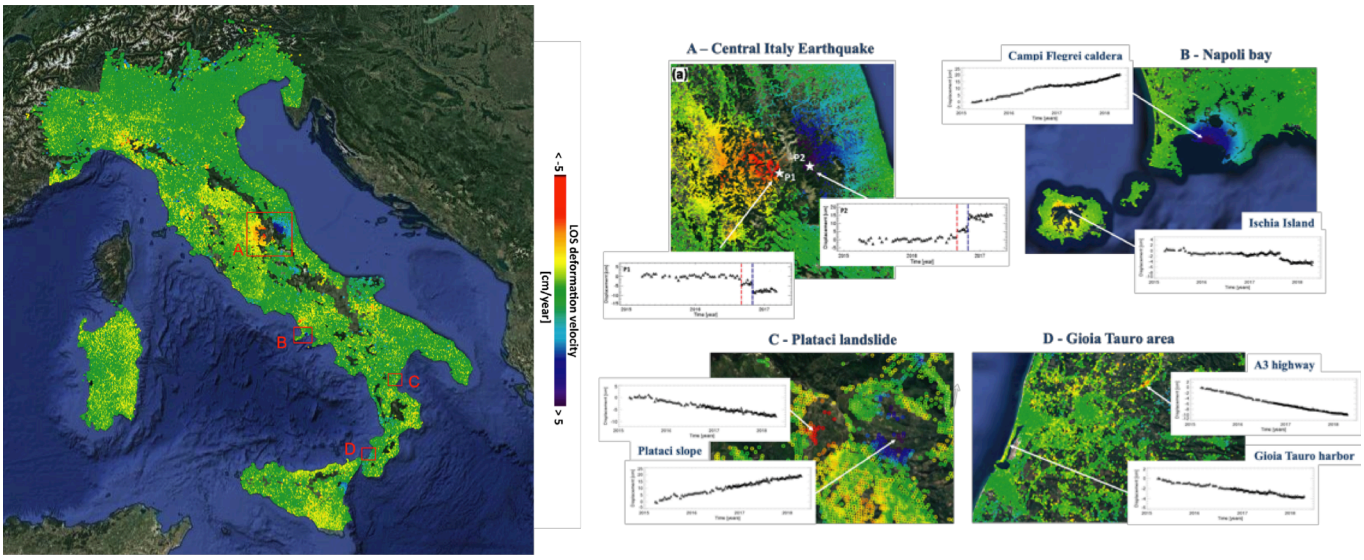


Fig 2. Overall mean deformation velocity map of the Italian territory generated through the S1 P-SBAS processing chain implemented within the AWS environment. The red rectangles represent four areas characterized by significant deformation phenomena.

the time series representing the evolution of the surface deformation within the considered time interval.

ACKNOWLEDGMENT

This work is supported by the Italian Civil Defence Protection Department, the EPOS-IP, the EOSC-hub, the ENVRI-FAIR and the OpenAIRE-Advanced projects of the European Union Horizon 2020 for research and innovation program (grant agreements: 676564, 777536, 824068, 777541, respectively), the I-AMICA (PONa3_00363) project, and the IREA-CNR/ Italian Ministry of Economic Development DGS-UNMIG agreement. Sentinel-1 SAR data are copyright of Copernicus (2016); the DEMs of the Italian territory are acquired through the SRTM archive.

5. REFERENCES

[1] A. K. Gabriel, R. M. Goldstein, and H. A. Zebker, “Mapping small elevation changes over large areas: Differential interferometry,” *J. Geophys. Res.*, 94, B7, 9183–9191, 1989.

[2] P. Berardino, G. Fornaro, R. Lanari, and E. Sansosti, “A new Algorithm for Surface Deformation Monitoring based on Small Baseline Differential SAR Interferograms”, *IEEE Trans. Geosci. Remote Sens.* 40, 11, 2375-2383, 2002.

[3] R. Lanari, O. Mora, M. Manunta, J. J. Mallorquì, P. Berardino, and E. Sansosti, “A small-baseline approach for investigating deformations on full-resolution differential SAR interferograms”, *IEEE Trans. Geosci. Remote Sens.,* 42, 7, 1377–1386, 2004.

[4] F. Casu, M. Manzo, and R. Lanari, “A quantitative assessment of the SBAS algorithm performance for surface deformation retrieval from DInSAR data”, *Remote Sens. Environ*., 102, 3/4, 195–210, 2006

[5] M. Bonano, M. Manunta, A. Pepe, L. Paglia, and R. Lanari, “From previous C-Band to New X-Band SAR systems: assessment of the DInSAR mapping improvement for deformation time-series retrieval in urban areas”, *IEEE Trans Geosci Remote Sens,* 51(4), 1973–1984, 2013

[6] M. Bonano, M. Manunta, M. Marsella, and R. Lanari, “Long-term ERS/ENVISAT deformation time-series generation at full spatial resolution via the extended SBAS technique”, *Int J Remote Sens,* 33, 4756–4783, 2012

[7] D. Laney, “3-D data management: Controlling data volume, velocity and variety”. Application Delivery Strategies by META Group Inc. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-DataVolumeVelocityandVariety.pdf, 2001

[8] F. Casu, S. Elefante, P. Imperatore, I. Zinno, M. Manunta, C. De Luca, and R. Lanari, “SBAS-DInSAR Parallel Processing for Deformation Time-Series Computation,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 7, no. 8, pp. 3285–3296, 2014

[9] I. Zinno, S. Elefante, L. Mossucca, C. De Luca, M. Manunta, O. Terzo, R. Lanari, and F. Casu, "A First Assessment of the P-SBAS DInSAR Algorithm Performances Within a Cloud Computing Environment," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*, 8,10, 4675-4686, 2015

[10] I. Zinno, F. Casu, C. D. Luca, S. Elefante, R. Lanari, and M. Manunta, "A Cloud Computing Solution for the Efficient Implementation of the P-SBAS DInSAR Approach," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens*, 10, 3, 802-817, 2017

FACING THE GEOSPATIAL INTELLIGENCE CHALLENGES IN THE BIG EO DATA SCENARIO

Sergio Albani, Paula Saameño, Michele Lazzarini, Anca Popescu, Adrian Luna

European Union Satellite Centre, Apdo de Correos 511, 28850 Torrejón de Ardoz, Spain

ABSTRACT

The global geopolitical situation is highly dynamic and security issues rise all around the world. In this respect, decision makers have to take suitable actions to respond in due time to challenging situations. The European Union Satellite Centre (SatCen) is an Agency of the Council of the European Union (EU) whose mission is to support the decision making and actions of the EU in the field of Common Foreign and Security Policy (CFSP) by providing products and services resulting from the exploitation of relevant space assets and collateral data. To support SatCen in accomplishing its mission, R&I activities are carried out to make the maximum benefit from Earth Observation (EO) by applying state-of-the-art solutions in the field of data management and exploitation as well as incorporating new data sources appearing in the market in the latest years to establish new services and products offering.

Index Terms— *Space and Security, Earth Observation, GEOINT, Big Data, ESA, GEO, RTDI*

1. INTRODUCTION

The entire world is facing challenges that are more diverse and less predictable than before. In particular, the domain of Security is presently fuzzing its boundaries, as areas like urbanization, social movements, political instability and climate change are challenging the current international state of play. International initiatives, such as the United Nations (UN) 2030 Sustainable Development Agenda[1], are working towards a more sustainable future, with clear targets defined including peace and wellbeing of the population. To address these targets, geospatial (and collateral) data constitute a large, reliable and sustainable resource for Security applications. However, as the EO data are constantly growing in terms of variety, volume, velocity, veracity and value, the key challenge in the Space and Security domain is to improve the capacity to access, process, analyse and visualize huge amounts of heterogeneous data to provide decision-makers with timely, clear and useful information.

The current geopolitical landscape and foreseen trends at international level for the coming period are challenging the way Geospatial Intelligence (GEOINT) is produced and delivered. Constantly, new channels are emerging with regard to data provisioning and how relevant information is being made available to each user, also in terms of how value is being added to the GEOINT products. Moreover the dynamics of the GEOINT landscape (including policies and regulations) demand preparedness to incorporate in the workflows new solutions and technologies. Traditionally, the production chain in GEOINT is following pipeline or waterfall approaches, where controlled processes are used to create well defined products, the core value unit being an intelligence report. With the current technological ecosystem, including the proliferation of new sensors and instrumentation, communication and social media, as well as the evolution in processing and storage solutions, the balance is shifting. Today, GEOINT is being rethought from managing serial-flow processes to properly managing the interactions between data, users and systems, and organizing ecosystem resources [1].

This paper discusses how SatCen is addressing the access and treatment of big geospatial data in an integrated framework that breaks away from the traditional waterfall approaches, considering two main factors: the increase of EO data and the way of managing them incorporating new technologies in an efficient manner.

1.1. A Big Ecosystem of Big EO Data

Big Data has become already a term used in almost every business sector, and most of the challenges that were being faced in the beginning of the hype have been, at least partially, resolved by the advances in storage solutions, cloud computing, communication systems, and the abundance of open Machine Learning and Data Science frameworks and tools. But while maturity has been reached to a certain extent at infrastructure and methodology level for conventional Big Data management, Big EO Data has still to reach its consolidation phase. This is not only due to the inherent complexity of the geospatial data itself, which requires specialized tools and understanding for proper information extraction and knowledge discovery, but also to the complexity of the sector with a visible increase in the number of new actors entering the market.

The free, full and open access to EO data provided by the Copernicus Programme[2] has been a game changer, creating the premises for new and innovative services in different domains, including Security (with a growing relevance of Sentinel-1 and 2 satellites for security applications [2], [3]).

[1] https://sustainabledevelopment.un.org/sdgs

[2] http://copernicus.eu/

doi:10.2760/848593

These Medium-High Resolution satellites are adding value to the current image interpretation practices of the Space and Security community, based mainly on Very High Resolution (VHR) data, as they can provide quick views of large areas, and support pro-active monitoring of Areas of Interest (AOIs) in view of current or future tasks.

In addition, in the coming years the access to – and availability of – satellite data is envisaged to follow a different approach from that of today as new data acquisition systems and new constellations are increasing the volume and the variety of EO data. In this context, the way to look at the classical 5 Vs of Big Data is advancing from data level to business level. Some operators already have in orbit or announced large constellation of EO satellites, which highlights the changing space market. But the dramatic expansion of available remote sensing data is not only due to the larger number of satellites in-orbit and expected satellite constellations [4]. Other acquisition systems such as Remote Piloted Aircraft Systems (RPAS) and High-Altitude Pseudo-Satellites (HAPS) are reaching enough maturity to be considered as alternative or complementary systems to run a specific analysis over an Area of Interest (AoI). Open data sources (e.g. OpenStreetmap) and collateral data such as social media data from Twitter or mobile data, in-situ and citizen science are complementing imaging systems and enable exploitation.

1.2. EO platforms for data management and processing

EO platforms have demonstrated their capability of enhancing the traditional GEOINT production based on serial-flow processes performed at users' workstations in terms of collaboration, efficiency and interoperability.

Traditional software architecture for EO Platforms and processes are following a layered, n-tier, monolithic approach. Monolithic applications are simpler and faster to develop, deploy and operate, but they also show important limitations as they tend to grow in size along years of development, accommodating changes within a single source code base, making them harder to maintain. They are also harder to scale, since the whole application needs to be modified, scaled and deployed at once. Performance can be improved by running multiple copies of the application behind a load balancer (x-axis scalability), but functional decomposition is usually hard. Also, monolithic applications are usually written in a single programming language and they are based on a single technology stack. Nowadays, new disruptive technologies emerge very quickly, which makes long-term commitment to a technology stack less agile.

The emerging approaches for EO Platforms are built upon innovative software architecture styles (e.g. microservices), that harness the power of new IT technologies and frameworks in response and anticipation of current and future needs coming from the rapid evolution of EO and cloud solutions [5]. Microservices, defined as "loosely coupled Service-Oriented Architecture with bounded contexts", show important advantages compared to the traditional monolithic style [6]. Loosely coupling forces services to be mutually independent, meaning that they can be updated, upscaled and deployed without requiring any modifications to other services in the architecture. This service independency opens the possibility to use the right technology for each specific service. These new architecture styles are also the base for cloud-aware systems which are able to elastically scale utilizing the underlying hardware resources in an efficient way. As EO processing tasks might be computationally demanding, but demand for processing is usually intermittent with peaks and valleys of demand, cloud-aware systems can benefit EO Platforms providing a more efficient usage of the infrastructure. To adopt these solutions, it is important to understand and carefully design strategies to tackle their operational and networking overhead due to the added complexity and their distributed nature.

2. STRATEGIES FOR SUSTAINABLE R&I IN THE SPACE AND SECURITY DOMAIN

Being SatCen an EU Agency with a recognized operational capability, R&I activities are strongly driven by operational requirements. The SatCen Research, Technology Development and Innovation (RTDI) Unit has the primary role to assess state-of-the-art technologies and deliver innovative geospatial management solutions that can improve the SatCen operational capabilities to offer EO products and services to Space and Security stakeholders.

To grow the SatCen innovation process, the RTDI Unit is building synergies mainly through cooperation with ESA (being responsible for the implementation of the cooperative activities defined in the ESA-SatCen Administrative Arrangement) and GEO (leading the Space and Security Community Activity[3] and contributing to the EuroGEOSS initiative). To achieve successful pre-operational results, the evaluation of suitable technologies and applications is also supported through the participation in H2020 projects. The accomplished projects BigDataEurope[4] and EVER-EST[5] constituted the pillars for the on-going projects NextGEOSS[6] and BETTER[7], which demonstrate solutions applied to different pilot cases in the Space and Security domain. This enables gathering and usage of expertise not available in-house, fosters cooperation, empowers the maximisation of the impact of the R&I activities within the Space and Security community and ensures the proper alignment with relevant European and global developments.

Building on the experience and expertise gained from these activities, the RTDI Unit is consolidating SatCen in-house innovation capabilities and implementing new

[3] GEO Space and Security Community Activity
[4] https://www.big-data-europe.eu/
[5] https://ever-est.eu/
[6] https://nextgeoss.eu/
[7] https://www.ec-better.eu/

operational solutions looking at the whole EO and collateral data lifecycle by focusing on four main work streams:

a) Use of new data acquisition systems such as HAPS, satellite constellations and RPAS (or even new sensors as thermal infrared or hyperspectral) that makes mandatory an adaptation of current SatCen operational flow to introduce these data sets;
b) Ingestion of alternative data sources (e.g. mobile networks and social media) to get indicators of activities happening in specific AoIs, linking them to other geospatial information;
c) Application of emerging technologies (e.g. Big Data, Cloud Computing, Interoperable Platforms, Artificial Intelligence and Machine Learning) to GEOINT activities;
d) Innovative EO solutions (e.g. SAR based Change Detection services) to turn data into relevant insights.

Figure 1 shows the RTDI paradigm to reach pre-operational phase of services and products, starting from stakeholders' requirements.

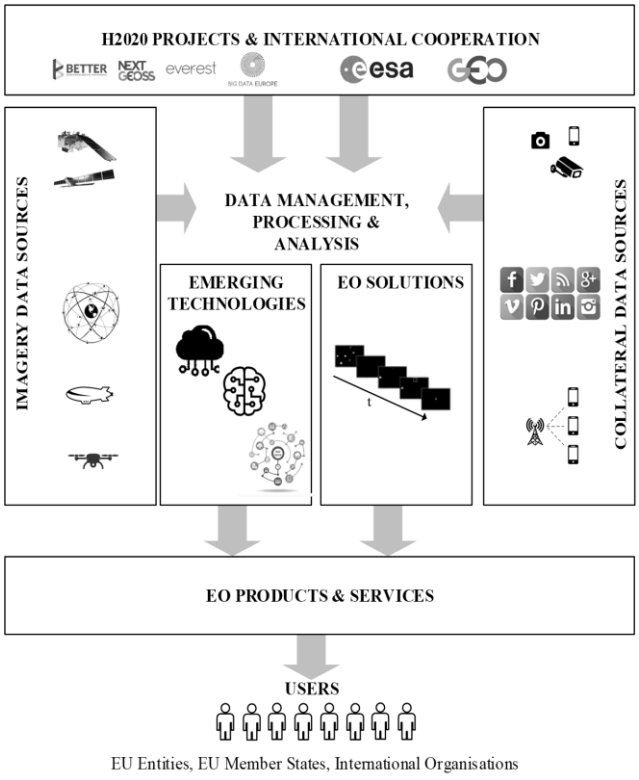


Figure 1. RTDI work streams

3. MOVING TOWARDS A PLATFORM APPROACH FOR GEOINT

To be able to efficiently validate new solutions and innovate service delivery in order to cope with the present-day world challenges, the RTDI Unit is moving from a process-driven to a platform-driven approach implementing an operational EO Platform to access and process relevant data for SatCen. This new Platform is based on the evolution of RTDI preliminary developments [7] and on the extension/customisation of existing commercial solutions[8], and it is envisaged as a main instrument for the conception, implementation and validation of new services and products to enlarge SatCen portfolio. Having this in mind, the definition of new EO applications and services is centered on the concept of enabling as many uses as possible of geospatial information.

From a functional perspective, the Platform aggregates vertical functionalities, dealing with the full data exploitation cycle, putting together data and processes from different sources, enabling the exploitation of the huge (and growing) amount of data generated. Contrary to the traditional analysis in which users download data to their workstations to analyse them, unnecessary downloads of raw data are avoided in favour of enriched and more digested data. The Platform is designed to abstract the underlying hardware infrastructure in order to make use of it elastically, upscaling or downscaling IT requirements as and when required to cope with the intermittent computational needs. Additionally, the proposed modular design enables the evolution of services as well as the integration of new services on the go. Services are implemented using heterogeneous technology stacks and utilizing the IT infrastructure best suited for them (e.g. SSD for I/O exhaustive tasks, GPUs for advanced calculations). Instead of relying in custom interfaces, the proposed architecture is built upon interoperability, making use of well-defined open standards for communication interfaces between services. Special attention is taken to the standards defined by the Open Geospatial Consortium (OGC)[9], since they are the most relevant for the domain (e.g. CSW and OpenSearch for Data Discovery; WCS, WMS, WFS for Data Access; and WPS for Data Processing).

The Platform will take advantage of new IT frameworks and technologies for this change of paradigm. Docker[10] is a good example of such technologies as it allows to run software components, securely isolated in containers, packaged with all its dependencies and libraries. Contrary to traditional Virtual Machines, containers can be faster and less resource heavy, presenting lower system overhead, so that taking up a container takes few seconds compared to minutes of a VM. Additionally, the Docker Engine provides an API for interacting with the Docker daemon with built-in functionalities to run containers and processes seamlessly. Therefore, containers are well-suited elements to encapsulate distributed services and processes that can be dynamically started, upscaled, or dismissed in seconds. Accompanying Docker, orchestration frameworks like Swarm and Kubernetes[11] are used to coordinate and run the multiple

[8] http://www.eodataservice.org/
[9] http://www.opengeospatial.org/
[10] https://www.docker.com/
[11] https://kubernetes.io/

components in a highly available and fault tolerant fashion. Finally, with the explosion of cloud computing, which offers a rapidly scalable infrastructure with seemingly infinite computational and storage resources over the Internet, minimizing as well operative efforts, the capacity of IT systems is rarely an issue. The approach followed in the Platform considers only vendor - and platform - agnostic solutions to ensure long-term sustainability as well as to support on-premises deployment in view of the risks derived from the management of classified information on the cloud. Cybersecurity and information assurance are also key issues when dealing with the distributed architecture of the Platform and the access to data. The Platform is designed to put in place authentication, authorization, access control and security mechanisms to ensure safe and secure access to data, processes and results generated, as well as convenient auditing capabilities by registering user actions.

4. INITIAL SERVICES

An initial set of services, built on the outcomes of a number of past and current RTDI initiatives, has been identified to be deployed as use cases to validate the Platform:

a) *Pre-processing of Sentinel-1 and Sentinel-2 data (PREP)* Chain of pre-processing modules to obtain an ortho-rectified image to be visualized or downloaded for further analysis on external tools;
b) *Amplitude Change Detection on Sentinel-1 (ACD)* Compare two Amplitude SAR images after pre-processing to produce Change Detection maps available in raster or vector format (after clustering);
c) *Multi-temporal Coherence on Sentinel-1 (MTC)* Generate a Coherence image from a couple of SLC images to be visualized with the Amplitude of the Master and Slave images in an RGB complex;
d) *Automatic Change Detection on Sentinel-2 (CDS2)* Compare two optical images after pre-processing to produce Change Detection maps available in raster or vector format (after clustering);
e) *Satellite Image Time Series (SITS)* Generate a multi-band raster product, where each band has the same physical meaning but different time indexes.

In the future, the adoption of Object Detection services based on Machine Learning algorithms, already explored within previous activities, will be considered, together with new innovative services.

5. CONCLUSIONS

To face the challenges of the Big EO Data Scenario, it is crucial to explore the best way to exploit EO and new geospatial data sources through the application of emerging technologies in order to create innovative EO solutions, harnessing at the same time the growing value and power of open innovation. Moreover, cooperation amongst key organisations in the Space and Security domain has to be fostered.

RTDI activities aim at ensuring that SatCen operational capabilities are maintained at the state-of-the-art. The Platform described in this paper, built on the experience gained through the participation in a number of R&I initiatives, will offer a solution to improve the SatCen capability to discover, access, process, share and interoperate huge amounts of heterogeneous geospatial data, fostering a more effective and efficient GEOINT production that will serve to address different challenges related to Security applications.

6. REFERENCES

[1] 2018 State and Future of GEOINT Report, *United States Geospatial Intelligence Foundation*, available online: https://usgif.org/education/StateofGEOINT

[2] R. Torres, P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Rommen, N. Floury, M. Brown, I. Navas Traver, P. Deghaye, B. Duesmann, B. Rosich, N. Miranda, C. Bruno, M. L'Abbate, R. Croci, A. Pietropaolo, M. Huchler, and F. Rostan, "GMES Sentinel 1 Mission", *Remote Sensing of Environment*, 120, pp. 9-24, 2012.

[3] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini, "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services", *Remote Sensing of Environment*, 120, pp. 25-36, 2012.

[4] Prospects for the Small Satellite Market, Euroconsult, 2018 Edition, *Euroconsult Executive Report*

[5] Big Data on Earth Observation Whitepaper, Big Data Value Association, TF7-SG5: Earth Observation and Geospatial http://www.bdva.eu/sites/default/files/TF7%20SG5%20Working%20Group%20-%20White%20Paper%20EO_final_Nov%202017.pdf, Nov 2017

[6] A. Wu, "Taking the Cloud-Native Approach with Microservices", 2017 https://cloud.google.com/files/Cloud-native-approach-with-microservices.pdf

[7] S. Albani, M. Lazzarini, P. Nunes, E. Angiuli "A platform for management and exploitation of big geospatial data in the space and security domain", *Proceedings of Big Data from Space 2017 Conference*, Toulouse, November 2017.

CLOUD BASED SPATIO-TEMPORAL ANALYSIS OF CHANGE IN SEQUENCES OF SENTINEL IMAGES

Allan A. Nielsen[a], Morton J. Canty[b], Henning Skriver[c] and Knut Conradsen[a]

[a]Technical University of Denmark, DTU Compute – Applied Mathematics and Computer Science
DK-2800 Kgs. Lyngby, Denmark
[b]Heinsberger Str. 18, D-52428 Jülich, Germany
[c]Technical University of Denmark, DTU Space – National Space Institute
DK-2800 Kgs. Lyngby, Denmark

1. INTRODUCTION

An important task in remote sensing Earth observation involves the detection of changes which may signal for example environmentally significant events. The Sentinel-1 synthetic aperture radar (SAR) and the Sentinel-2 as well as the Landsat optical/visible-infrared spaceborne platforms, with spatial resolutions of the order of 10-20-30 meters and revisit times of the order of days, provide an attractive source of data for change detection tasks. Specifically, the SAR imagery provide complete independence from solar illumination and cloud cover. A convenient source of such data is the Google Earth Engine which gives near real time data access and which has an application programming interface for the access and for processing the data. Here we make available open-source automatic change detection software and for optical data also automatic radiometric normalization software for both cloud and local processing.

The theory sections of this contribution are very similar (nearly identical) to sections in [1]. In this contribution, we exclude examples on radiometric normalization and include new developments in both stand-alone and cloud software implementation and we give new examples.

2. CHANGE DETECTION IN SAR DATA

In [2] a change detection procedure for multi-look polarimetric SAR data [3] is described involving a test statistic (and its factorization) for the equality of polarimetric covariance matrices following the complex Wishart distribution. The procedure is capable of determining, on a per-pixel basis, if and when a change at any prescribed significance level has occurred in a time series of SAR images. The procedure may also be applied to collections of pixels (segments, patches, fields). Single polarization (power data, dimensionality $p = 1$), dual polarization (for example vertically polarized transmission, vertical and horizontal reception, $p = 2$) and full or quad polarization (all four combinations of vertical and horizontal transmission/reception, $p = 3$) can be analyzed.

The term multi-look in SAR imagery refers to the number of independent pixels (termed the equivalent number of looks, ENL) of a surface area that have been averaged in order to reduce the effect of speckle, a noise-like consequence of the coherent nature of the signal transmitted from the sensor. The observed signals in the covariance representation, when multiplied by the equivalent number of looks, are complex Wishart distributed. This distribution is the multivariate complex analogue of the well-known chi squared distribution.

The complex Wishart distribution is completely determined by the parameters p (dimensionality), ENL, and $\mathbf{\Sigma}$ (the variance-covariance matrix). Given two observations of the same area at different times, one can set up a hypothesis test in order to decide whether or not a change has occurred between the two acquisitions. The null hypothesis, H_0, is that $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, i.e., the two observations were sampled from the same distribution and no change has occurred, and the alternative (change) hypothesis, H_1, is $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. Since the distributions are known, a likelihood ratio test can be formulated which allows one to decide to a desired degree of significance whether or not to reject the null hypothesis. Acceptance or rejection is based on the test's p-value, which in turn may be derived from the (approximately known) distribution of the test statistic when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ ("under H_0" in statistical parlance).

For analysis of the situation with data from two time points, $k = 2$ below, see [4, 5, 6, 7]. In [8] the authors describe bi-temporal region-based change detection for polarimetric SAR images by means of mixtures of Wishart distributions.

If we have data from more than two time points, $k > 2$, the procedure sketched can be generalized to test a hypothesis that all of the k pixels (or patches) are characterized by the same $\mathbf{\Sigma}$,

$$H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_k (= \mathbf{\Sigma})$$

against the alternative (H_1) that at least one of the $\mathbf{\Sigma}_i$, $i = 1, \ldots, k$, is different, i.e., that at least one change has taken place.

For the logarithm of the omnibus likelihood ratio test statistic Q for testing H_0 against H_1 we have, see [2]

$$\ln Q \quad = \quad n\{pk \ln k + \sum_{i=1}^{k} \ln |\boldsymbol{X}_i| - k \ln |\boldsymbol{X}|\}.$$

Here n is ENL, the $\boldsymbol{X}_i = n\hat{\mathbf{\Sigma}}_i$ (i.e., ENL times the observed covariance matrix) follow the complex Wishart distribution, $\boldsymbol{X}_i \sim W_C(p, n, \mathbf{\Sigma}_i)$, and $\boldsymbol{X} = \sum_{i=1}^{k} \boldsymbol{X}_i \sim W_C(p, nk, \mathbf{\Sigma})$. Also, if the hypothesis is true, $\hat{\mathbf{\Sigma}} = \boldsymbol{X}/(kn)$. $Q \in [0, 1]$ with $Q = 1$ for equality.

The probability of finding a smaller value of $-2 \ln Q$ is approximated by ($z = -2 \ln q$, where q is the actually observed value of Q)

$$P\{-2 \ln Q \leq z\} \quad \simeq \quad P\{\chi^2((k-1)f) \leq z\},$$

i.e., the probability of change at some time point. $f = 9$ for quad pol, $f = 4$ for dual pol, $f = 2$ for dual pol diagonal only.

Furthermore this test can be factored into a sequence of tests involving hypotheses of the form $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ against $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$,

$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ against $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 \neq \mathbf{\Sigma}_3$, and so forth. More specifically, to test whether the first $1 < j < k$ complex variance-covariance matrices $\mathbf{\Sigma}_i$ are equal, i.e., given that

$$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_{j-1}$$

then the likelihood ratio test statistic R_j for testing the hypothesis

$$H_{0,j} : \mathbf{\Sigma}_j = \mathbf{\Sigma}_1 \text{ against } H_{1,j} : \mathbf{\Sigma}_j \neq \mathbf{\Sigma}_1$$

is given by, see [2]

$$\begin{aligned} \ln R_j &= n\{p(j\ln j - (j-1)\ln(j-1)) \\ &\quad + (j-1)\ln|\sum_{i=1}^{j-1}\boldsymbol{X}_i| + \ln|\boldsymbol{X}_j| - j\ln|\sum_{i=1}^{j}\boldsymbol{X}_i|\}. \end{aligned}$$

Finally, the R_j constitute a factorization of Q such that $Q = \prod_{j=2}^{k} R_j$ or

$$\ln Q = \sum_{j=2}^{k} \ln R_j.$$

The probability of finding a smaller value of $-2\ln R_j$ is approximated by ($z_j = -2\ln r_j$, where r_j is the actually observed value of R_j)

$$P\{-2\ln R_j \leq z_j\} \simeq P\{\chi^2(f) \leq z_j\},$$

i.e., the probability of change at time point j with no previous change.

The tests are statistically independent under the null hypothesis. In the event of rejection of the null hypothesis at some point in the test sequence, the procedure is restarted from that point, so that multiple changes within the time series can be identified. For details including better approximations to the distributions of Q and R_j under the null hypotheses, see [2], visualization of change, and (some of) the software developed, see [9].

Since the omnibus method can detect not only if changes occur but also, within the temporal resolution of an image sequence, when they occur, long time series of frequent acquisitions over relevant sites are of special interest. One convenient source of such data is the Google Earth Engine[1] (GEE) [10] which ingests Sentinel-1 data (C-band, multi-looked VV/VH or HH/HV) as soon as they are made available by the European Space Agency (ESA) and provides an easy-to-use application programming interface (API) for accessing and processing the data.

3. CHANGE DETECTION AND RADIOMETRIC NORMALIZATION IN OPTICAL DATA

With respect to optical/visible-infrared imagery, a data-driven, statistical approach to change detection is provided by the iteratively reweighted multivariate alteration detection (IR-MAD) algorithm [11, 5]. This method applies iterated canonical correlation analysis (CCA) to geometrically co-registered multispectral images from two time points before calculating band-wise differences. The CCA orders the image bands according to similarity (measured by correlation), rather than spectral wavelength. The differences between corresponding pairs of canonical variates are termed the MAD variates. Specifically, a MAD variate Z is

$$Z = \boldsymbol{a}^T\boldsymbol{X} - \boldsymbol{b}^T\boldsymbol{Y}$$

where $\boldsymbol{X}$ represents the m-dimensional image at time point 1, $\boldsymbol{Y}$ represents the m-dimensional image at time point 2, and $\boldsymbol{a}$ and $\boldsymbol{b}$ are the eigenvectors from the CCA. Thus $\boldsymbol{a}^T\boldsymbol{X}$ is a canonical variate for time point 1 and $\boldsymbol{b}^T\boldsymbol{Y}$ is a canonical variate for time point 2. We have m uncorrelated canonical variates (CVs) with mean value zero and variance one from both time points, the correlation between corresponding pairs of CVs is ρ (termed the canonical correlation which is maximized in CCA), and we have m uncorrelated MAD variates with mean value zero and variance $2(1-\rho)$.

In each iteration the values of each image pixel j are weighted by one minus the current estimate of the change probability and the image statistics (mean and covariance matrices) are re-sampled. Since the MAD variates for the no-change observations are approximately Gaussian and uncorrelated, the sum of their squared values (after normalization to unit variance)

$$C^2 = \sum_{i=1}^{m} \frac{Z_i^2}{2(1-\rho_i)}$$

ideally follows a chi squared distribution with m degrees of freedom, $C^2 \sim \chi^2(m)$. The probability of finding a smaller value of C^2 is approximated by (c^2 is the actually observed value of C^2)

$$P\{C^2 \leq c^2\} \simeq P\{\chi^2(m) \leq c^2\}.$$

Small P-values favour rejection of the no-change hypothesis, so for each iteration, $1 - P\{\chi^2(m) \leq c^2\}$ is used to weight each pixel to gradually reduce the influence of the change observations on the MAD transformation. Iterations continue until the canonical correlations stop changing (or a maximum number of iterations is reached).

Furthermore, canonical correlation analysis is invariant to linear and affine transformations, a fact that can be used to perform automatic relative radiometric normalization of the two multispectral images [12, 1]. This is not pursued further here.

4. SOFTWARE

The authors have made available the necessary change detection software for interaction with the GEE on the open-source repository Github[2]. The client-side programs run in a local Docker container serving a simple Flask web application. Apart from the Docker engine[3] and a browser, no software installation is required whatsoever. After the user has been authenticated to the Earth Engine, he or she can carry out the following tasks: 1) run the IR-MAD algorithm on Sentinel-2 (or Landsat) bi-temporal imagery, 2) perform relative radiometric normalization in batch mode on an image sequence, 3) run the sequential omnibus algorithm on Sentinel-1 dual polarization image time series, 4) export imagery to his or her Earth Engine assets folder or to Google Drive for further processing or visualization.

JavaScript code[4] to run both the Wishart omnibus and the IR-MAD methods directly in the GEE code editor/playground is also available. The Wishart omnibus code also generates an MP4 movie showing where and when change occurred.

As a recent development, a Docker-based interface to the GEE for the Wishart omnibus algorithm is made available.[5] It talks to the GEE servers from a Jupyter notebook and is more flexible than the

[1] https://earthengine.google.com/ and https://developers.google.com/earth-engine/

[2] https://github.com/mortcanty/earthengine/

[3] https://docs.docker.com/

[4] http://fwenvi-idl.blogspot.de/

[5] http://fwenvi-idl.blogspot.com/2018/07/jupyter-notebook-interface-for.html

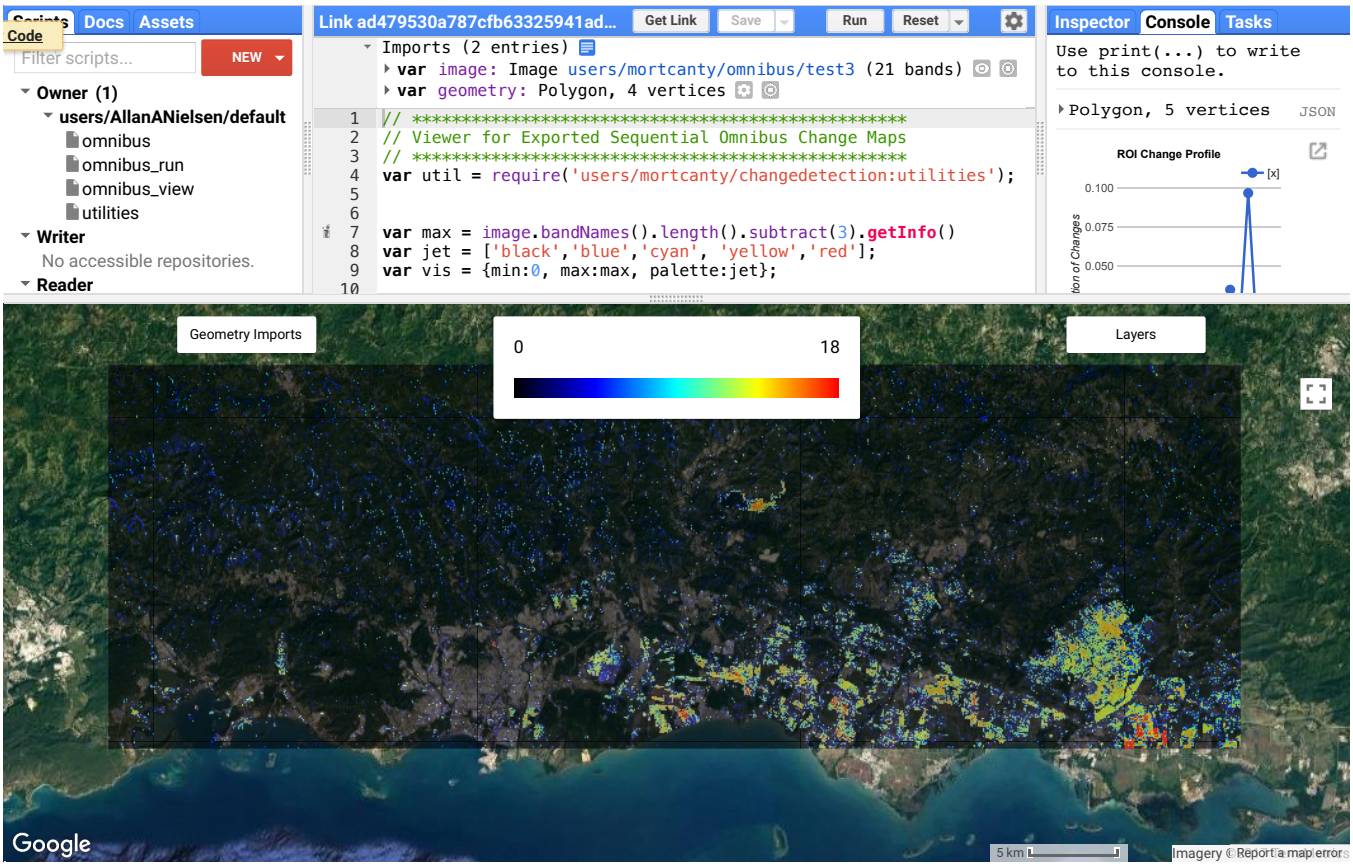


Fig. 1. Sequential omnibus change map for a region in southern Puerto Rico, showing the time of the most recent change (black none, blue early, red late). The time series consisted of 19 Sentinel-1 images from April to October 2017. Hurricane Maria struck on 20 September.

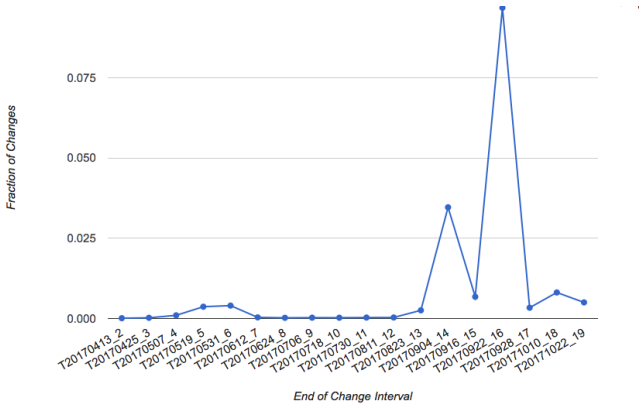


Fig. 2. Fraction of changed pixels in the south-eastern part of the change image shown in Figure 1. The peak occurs for the interval ending 22 September 2017, hurricane Maria struck on 20 September.

web interface, since the user is in a universal interactive Python programming environment.

Software is available also for local processing,[6] see [9]. Tutorials on how to install software and to do both the polarimetric SAR and the optical data processing locally on your own hardware are available on Github.[7,8] As another recent development, computer implementation work has been done within the Horizon 2020 project DataBio[9] DLV-732064 funded by the European Union (command-line and GUI executables[10] for Windows and Linux based on our Matlab code and on extended code from [13], a version for small images which fit into memory and a line-by-line version for big data exist), see proceedings from this meeting (first author Behnaz Pirzamanbein).

5. EXAMPLES

To illustrate, the Sentinel-1 multi-temporal VV/VH based change map in Figure 1 displays the color-coded time intervals in which the most recent changes in the 2017 hurricane Maria catastrophe in Puerto Rico occurred. Figure 2 shows the fraction of changed pixels which peaks in the interval ending on 22 September 2017. Maria made landfall in Puerto Rico on 20 September 2017. The change maps can be viewed interactively in the GEE Code Editor.[11]

Changes in one of several wildfires (the so-called Tubbs Fire[12] which took place on 9-30 October 2017 between Calistoga and Santa

[6] https://people.compute.dtu.dk/alan/software.html

[7] https://mortcanty.github.io/src/tutorialsar.html
[8] https://mortcanty.github.io/src/tutorial.html
[9] https://www.databio.eu/
[10] https://github.com/BehnazP/DataBio/
[11] https://code.earthengine.google.com/9374d69f4b0e3c11a7a14a9581f858d0
[12] https://en.wikipedia.org/wiki/Tubbs_Fire

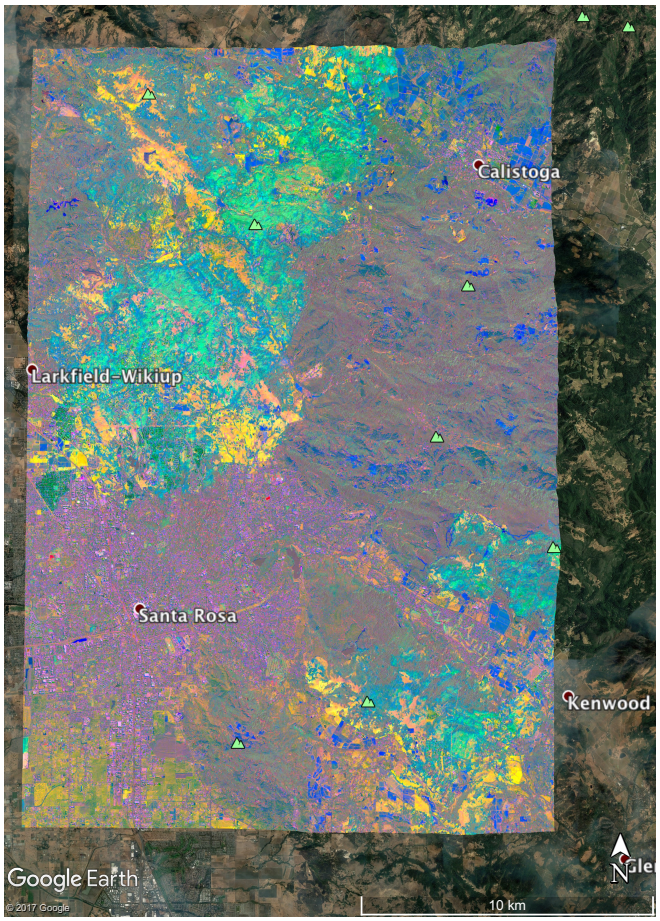


Fig. 3. The Tubbs Fire north of Santa Rosa, California, October 2017 (top-left; the bottom-right shows part of a larger fire around Kenwood). IR-MAD change variates associated with three greatest canonical correlations shown as RGB, burned areas in dark green (built-up areas), lighter green (mostly wooded) and bright yellow (mostly non-wooded), other non-fire related change mostly in blue (for example near Calistoga), and pale yellow (south of Santa Rosa). All variates are stretched over $\mp$ 16 no-change standard deviations.

Rosa and in which nearly 150 km^2 burned) in the northern California wine areas Napa Valley and Sonoma Valley are detected. The burned areas depicted in green (built-up and wooded areas) and bright yellow (non-wooded areas) in Figure 3 (where another fire down towards Kenwood is visible also) match well with published fire maps.[13,14] The Sentinel-2 images were acquired on 5 October and 1 November (bracketing the fire), only the four 10 m bands 2, 3, 4 and 8 were analyzed.

6. CONCLUSIONS

Examples based on both Sentinel-1 dual polarization synthetic aperture radar data and Sentinel-2 optical data show the usefulness of the generic, automatic change detection techniques sketched. Note, that for the optical change detection method, because of the orthogonality between the change variates, different types of change can be discriminated between.

The introduction of software for automated change analysis with polarimetric SAR as well as optical image data available to run either on your own hardware or to anyone authenticated to run on the Google Earth Engine is expected to be extremely useful to both researchers and practitioners. Generic, automatic techniques as these are expected to be useful in many other application areas also (other than natural disasters) where the study of spatio-temporal dynamics is important.

[13] http://abc7news.com/maps-a-look-at-each-north-bay-fire/2517694/
[14] http://fire.ca.gov/

7. REFERENCES

[1] M. J. Canty and A. A. Nielsen, "Spatio-temporal analysis of change with Sentinel imagery on the Google Earth Engine," in *ESA Conference on Big Data from Space (BiDS)*, pp. 126–129, Toulouse, France, 28-30 Nov 2017, https://doi.org/10.2760/383579.

[2] K. Conradsen, A. A. Nielsen, and H. Skriver, "Determining the points of change in time series of polarimetric SAR data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 5, pp. 3007–3024, 2016, https://doi.org/10.1109/TGRS.2015.2510160.

[3] J. J. van Zyl and F. T. Ulaby, "Scattering matrix representation for simple targets," in *Radar Polarimetry for Geoscience Applications*, F. T. Ulaby and C. Elachi, Eds. Artech, Norwood, MA, 1990.

[4] K. Conradsen, A. A. Nielsen, J. Schou, and H. Skriver, "A test statistic in the complex Wishart distribution and its application to change detection in polarimetric SAR data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 1, pp. 4–19, 2003, https://doi.org/10.1109/TGRS.2002.808066.

[5] M. J. Canty, *Image Analysis, Classification, and Change Detection in Remote Sensing, With Algorithms for ENVI/IDL and Python*, Taylor and Francis, Third revised edition, 2014.

[6] A. A. Nielsen, K. Conradsen, and H. Skriver, "Change detection in full and dual polarization, single- and multi-frequency SAR data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 8, pp. 4041–4048, 2015, https://doi.org/10.1109/JSTARS.2015.2416434.

[7] V. Akbari, S. N. Anfinsen, A. P. Doulgeris, T. Eltoft, G. Moser, and S. B. Serpico, "Polarimetric SAR change detection with the complex Hotelling-Lawley trace statistic," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 7, pp. 3953–3966, 2016, https://doi.org/10.1109/TGRS.2016.2532320.

[8] W. Yang, X. Yang, T. Yan, H. Song, and G.-S. Xia, "Region-Based Change Detection for Polarimetric SAR Images Using Wishart Mixture Models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 11, pp. 6746–6756, 2016, https://doi.org/10.1109/TGRS.2016.2590145.

[9] A. A. Nielsen, K. Conradsen, H. Skriver, and M. J. Canty, "Visualization of and software for omnibus test based change detected in a time series of polarimetric SAR data," *Canadian Journal of Remote Sensing*, vol. 43, no. 6, pp. 582–592, 2017, https://doi.org/10.1080/07038992.2017.1394182.

[10] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Tau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18–27, 2017, https://doi.org/10.1016/j.rse.2017.06.031.

[11] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 463–478, 2007, https://doi.org/10.1109/TIP.2006.888195.

[12] M. J. Canty and A. A. Nielsen, "Automatic radiometric normalization of multitemporal satellite imagery with the iteratively re-weighted MAD transformation," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 1025–1036, 2008, https://doi.org/10.1016/j.rse.2007.07.013.

[13] N. Falco, P. R. Marpu, and J. A. Benediktsson, "A toolbox for unsupervised change detection analysis," *International Journal of Remote Sensing*, vol. 37, no. 7, pp. 1505–1526, 2016, https://doi.org/10.1080/01431161.2016.1154226.

LOCAL AND AUTOMATED PROCESSING OF SENTINEL-2 TIME SERIES: ADRESSING THE BOTTLENECKS

Philipp Hochreuther, Nathalie Reimann, Matthias Braun

Institute of Geography, University of Erlangen-Nürnberg, Erlangen, Germany

ABSTRACT

Since the start of Sentinel-2 A/B in 2015/2017, high-resolution satellite images recorded at a five- to six day frequency enable the development of remote sensing time series over larger regions. The large amount of data coming with every single scene however is a big challenge for such tasks, as infrastructural-, hardware- and software bottlenecks hamper the efficient processing. We present a fully automated procedure based on a, local, non-cloud-based setup to handle several thousands of Sentinel-2 scenes based on open-source software and free ESA tools alone. The setup includes tools for fast data downloading, atmospheric correction and merging of adjacent granules. Download rates can be significantly boosted by making use of Google's cloud storage pool. Time consumption of (pre-)processing can be optimized by parallelizing using Ubuntu's built-in tools and R. The architecture is tested on the complete time series (2015 – 2018) of four granules over the 79N glacier, Greenland.

Index Terms— Sentinel-2, local platform, processing, optimization, time series

1. INTRODUCTION

Since the start of ESA's Sentinel-2 (S-2) A satellite in June 2015, high-resolution multi-spectral images have become freely available to everyone. S-2 B, started in March 2017, has shortened the revisiting interval to 5 days at the equator and less than two days in polar-near regions for each granule [1]. Each S-2 granule contains a wealth of metadata coded in xml format, plus JPEG 2000 images in all available resolutions (10, 20 and 60m). At this time, most of granules covering the planet are offered at the 1C processing level, meaning top of atmosphere (TOA) images in cartographic geometry. All images are 100x100km^2 ortho-images in UTM/WGS84 projection. Sentinel-2 data can be freely accessed via the Copernicus Open Access Hub, though there are various other sources where the data is mirrored; e.g. the French ground segment PEPS [2], or the USGS Earth Explorer. The Copernicus hub offers two ways of access: the OpenHub, which has a graphical user interface, and the API hub designed for automated downloads via scripts.

According to ESA, a typical S-2 scene is about 600 MB in size (https://sentinel.esa.int), though the data amount can vary dependent on the scene coverage. Though this data size is probably unproblematic for illustration purposes or local analyses, a current typical workstation is incapable of handling the raw data amount and processed products of larger areas or time series. Cloud platforms offer a solution for this dilemma through storing the imagery and substantial computational resources. Drawbacks of this are that, on the one hand, most cloud resources are not free, and on the other hand, processed data cannot be stored for longer terms, and be re-used outside of the cloud environment. Therefore, local resources may be a more attractive solution. We present an approach based on local hardware, with a stack of free software and tools aimed at efficient download and processing of S-2 data. The method is tested for a local setup consisting of four S-2 granules, but is potentially scalable, if sufficient computation and storage resources are available, for which we give recommendations based on empirical analysis of our work flow.

2. TEST DATA

We use a test set of four adjacent S-2 granules covering the tongue of the 79N glacier, northeast Greenland (Fig. 2). Prior to selecting suitable scenes, all available L1C scenes of the granules 26XMN, 26XMP, 26XNN and 26XNP were downloaded. Due to the availability of daylight, the annual period was restricted to March 15th to September 30th. This resulted in 5.817 scenes distributed over four years (Fig. 1), with a data size of 2.82 terabyte (TB), resulting in an average scene size of 508.64 megabyte (MB). As shown in Figure 1, 2015 contains the least scenes due to S-2 A started in June being the only satellite delivering data. For 2016, the number of scenes is still inferior to 2017/2018, as S-2 A is still the only satellite running.

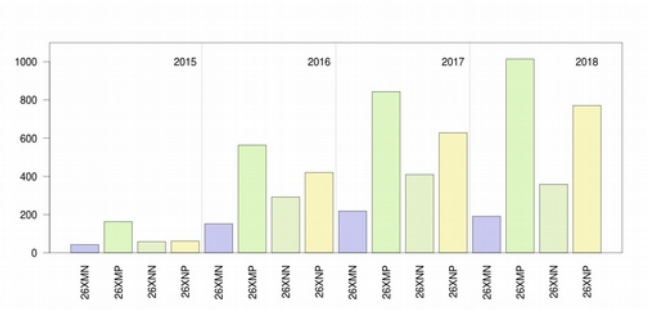


Figure 1: Granule count of the test dataset per year.

doi:10.2760/848593

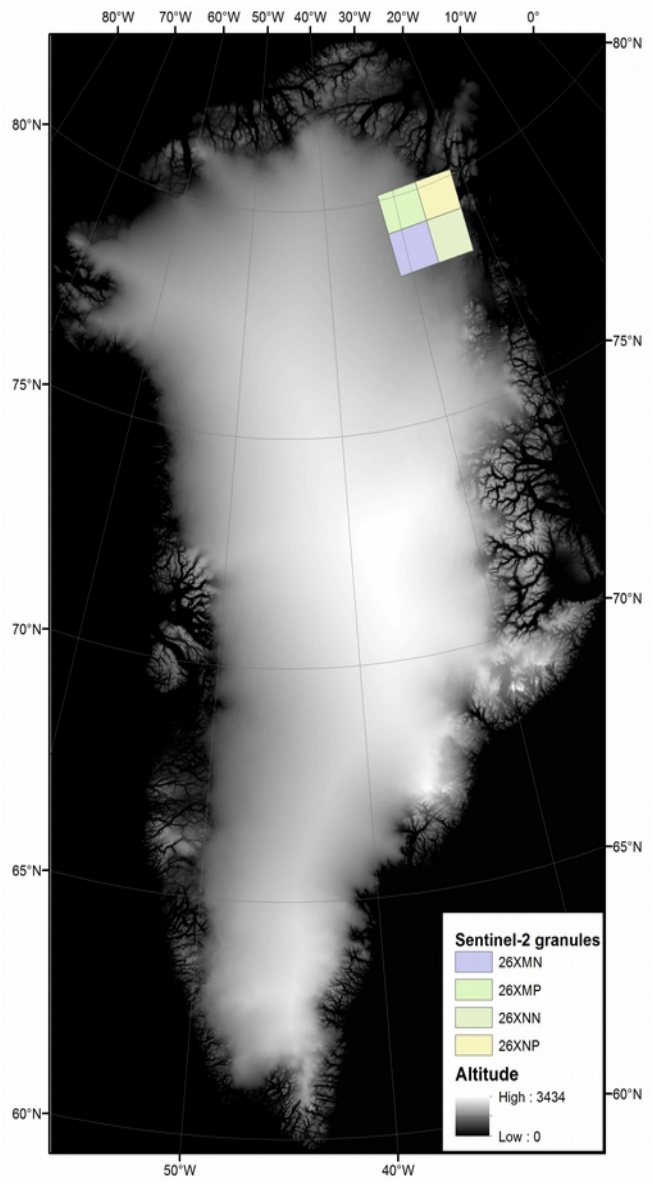


Figure 2: Location of the four Sentinel-2 granules covering the tongue of the 79N glacier, northeast Greenland. DEM: Greenland Ice Mapping Project (GIMP; [4]).

3. SYSTEM ARCHITECTURE

To ensure download and storage capabilities, we installed a processing- and a storage unit at the regional computing center Erlangen (RRZE). This ensures maximum download rates through direct connection to the core net of the German research network (DFN). The system runs on two processors with 12 cores/24 threads each and 128 GB RAM, with Ubuntu 16.04 server as operating system. Both servers are terminal servers without graphical user interface and can be accessed locally via SSH. The Samba protocol is employed to make the data available on the local workstation and to transfer scripts to the server. Additionally installed is the GDAL library [3], the standalone version of Sen2Cor [6] and the R environment [7]. The latter is used to script the whole process, making use of R's raster processing capabilities and, if needed, send commands to the bash. System workload is remotely monitored using the open source tool Munin.

4. PROCESSING CHAIN

As all processing is done locally, the first step requires downloading the Level 1C (L1C) data. Given the size of a single scene, this step is, with respect to the area of interest, the first possible bottleneck of the processing chain and is therefore described in detail in the respective chapter.
Within the following step, the tiles with a high percentage of no-data values are sorted out. Some download tools allow this exclusion by querying via the relative orbit (RO), which in return reduces the size of the data to download significantly. As our aim was to establish a fully automated procedure, the data is queried for the percentage of nodata values after downloading, requiring no a priori knowledge of the ROs. For some dates, more than one scene for one granule is downloaded; in these cases, the scene with the highest coverage is kept, the others are disregarded and deleted.

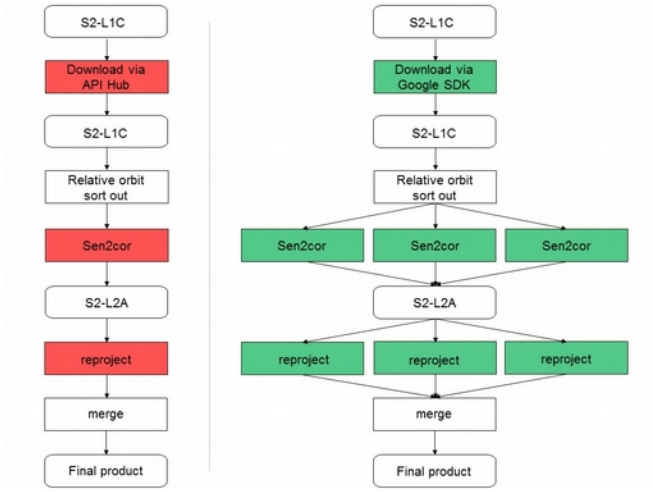


Figure 3: Sequential processing chain (left), with bottleneck processes color-coded in red, and partly parallelized processing chain (right).

Up to this point, the images kept still contain cloud cover. Potentially, over non-snow- or -ice-covered areas, the request for a cloud-free image can be met by a simple additional processing step querying the S-2 meta data for the percentage of cloud cover, or by counting cloudy pixels using the cloud mask delivered with the L1C product. For

snow-/ice-covered areas, the mask is sometimes not reliable, as the differentiation in the visible and thermal spectrum is difficult. We also tested the fmask algorithm [8] for L1C data, resulting in better, but still faulty cloud masks. Better results were achieved applying fmask over L2A data, delivering still far from perfect cloud masks, but already appropriate for decision making. For even better cloud masking and thus less error-prone results, we recommend looking e.g. into MAJA [5] or the Sentinel Hub cloud detector (https://medium.com/sentinel-hub/tools/home).

Next, the L1C data is transformed into Level 2A (L2A) data using the standalone version of Sen2cor, originally a processor within the ESA Sentinel-2 toolbox [5]. Sen2cor operates sequentially and includes atmospheric correction and scene classification. As optional parameters, we do not supply a Digital Elevation Model (DEM), as Sen2Cor only accepts the SRTM DEM (which is not available north of 60° latitude) or the commercial DTED-1 from PlanetDEM, which contradicts our goal to use only freely available tools and data. In addition, we do not restrain the processing to a certain resolution, thus L2A versions for all available resolutions are generated.

The S-2 data comes in UTM/WGS84 projection, in case of our test data UTM zone 26. This may hinder the direct composition (merging) of scenes, as this process typically requires the same projection, while two adjacent granules can spread over two UTM zones. Given the near-polar location of our test site, we reproject all tiles to the WGS84/NSIDC Sea Ice Polar Stereographic North projected coordinate system (EPSG 3413). This is achieved using the gdalwarp function from the GDAL library. To prevent inflation of the data size, a LZW compression is done afterwards using gdal-translate.

Before merging, a list containing all dates and corresponding L2A scenes is generated, which serves as input for gdalwarp. This list can be modified to contain only complete cases, i.e. all four scenes of the test data set.

5. BOTTLENECKS

5.1 Download rates

The Copernicus API hub can be addressed via various tools available. For this study, we use the script Sentinel-download by Olivier Hagolle (https://github.com/olivierhagolle/Sentinel-download). For this, a Copernicus account is needed, whose credentials are used by the tool to authenticate at the Copernicus server. Here, the data is downloaded per granule sequentially, though other methods exist, e.g. defining an area of interest using coordinates or geo-referenced vector data. With the above described configuration and data, we achieved download rates of 768.3 kB/s on average, resulting in a total download time for our test dataset of 1094.75 hours or 45.6 days. This can be reduced by parallelized downloading of all four granules in parallel. S-2 data is also mirrored by Google for the use in their Earth Engine, and can be downloaded without any login credentials using gsutil, a tool from Google Cloud SDK. The toolset can be installed free of charge by adding the package source to Ubuntu's repository list. Downloading the test data via Google cloud SDK resulted in a total download time of 56.65 hours or 2.36 days, thus an average download rate of 14.5 MB/s. Again, the process can be parallelized. This difference in speed arises from a) the authentication process and b) the http request sent for every single file in the .SAFE containers requested for download with the Copernicus server: The average download speed for the image files is with ~15 MB/s comparably high as with Google. Therefore, the 20-fold difference in download speed can be explained by the download tool, generating a list of files to download on the local client, and requesting authentication and download each file individually. With Google Cloud SDK, the file transfer is performed with a single copy command (gsutil -cp) and is thus significantly faster. The download command fetches all data up to date and thus is not running continuously; however, this can be implemented easily, as the algorithm checks for every S-2 scene if it is already present in the destination folder, and if so, skips the scene. A simple cron job starting the download script at a given interval will thus just download, and subsequently process, new data.

5.2 L2A processing

A single Sen2Cor run takes between 15 and 30 minutes, with 100% thread workload and up to 4 GB of main memory used. The given resources, especially memory usage, are used dependent on the processing step. An example is shown in Fig. 4: here, 24 threads are used for a parallelized run of Sen2Cor. During this run, the maximum memory usage of Sen2Cor adds up to 85 GB, equivalent to ~3.5 GB per single process. As this corresponds to the sequential run capacity usage, minimum RAM requirements of 4 GB per Sen2Cor thread for a parallelized run can be inferred. During the illustrated run, 1117 scenes worth of 424.1 GB were processed in 3 hours 40 minutes, resulting in a runtime of 5.1 minutes per scene.

5.3 Reprojection

Reprojecting all jpeg 2000 images within one S-2 scene sequentially using gdalwarp takes ~160 seconds. Thus, for 5000 thousand scenes, this process takes 9.25 days to complete if started sequentially. Parallelizing on 24 threads reduces this period to less than 24 hours.

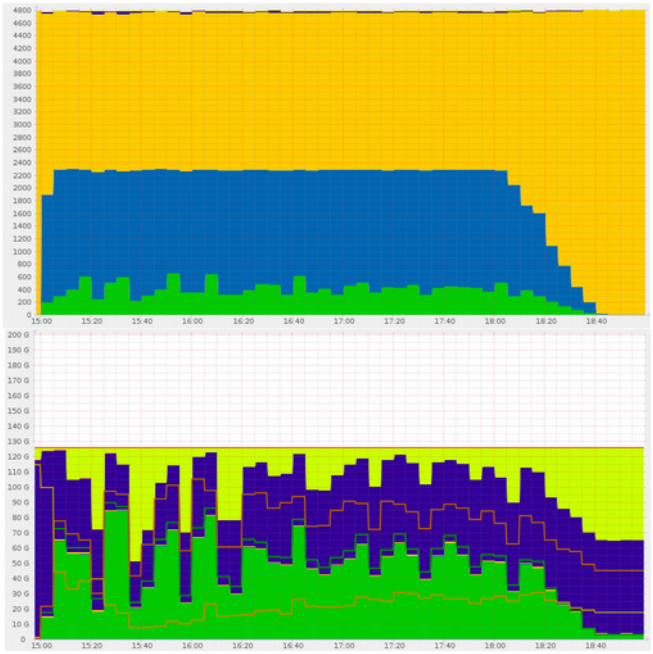


Fig. 4: System usage during a Sen2Cor run with 24 parallel threads. Upper graph: CPU usage in percent of all available threads (48 threads at 100%), consisting of system (green), user (blue) and idle (yellow). Lower graph: Memory usage (max. 128 GB), consisting of apps (Sen2Cor, green), cache (purple) and unused (yellow). Graphs and data produced with Munin.

6. HARDWARE RECOMMENDATIONS

Given the aforementioned system usage experiments, we recommend 4 GB of RAM per parallelized thread, e.g. for 24 parallel Sen2Cor runs 104 GB of RAM, plus a generously allocated SWAP partition. The size of the data partition should be calculated based on the number of granules required to cover the AOI and the length of the time series under consideration, times the average size of 500-600 MB per scene. During processing, the number of scenes will shrink, but L2A data have a around 33% higher data volume than L1C data. Thus, at least the same amount of space should additionally be calculated for storing all L2A scenes, if all temporal files are deleted during processing.

7. CONCLUSIONS

Given the availability of local processing and storage resources and a regionally limited area of interest, the setup presented is an interesting alternative to cloud processing of Sentinel-2 data. Bottlenecks in the processing chain are induced by the client-side data download tools and sequential processing. Those restrictions can be overcome by employing tools that offer server-side data querying, like Google cloud SDK, and by parallelizing processes that have to be executed hundreds of times similarly. Though the processing chain presented can still be optimized, e.g. through parallelizing the data download from Google cloud SDK, it already improves the time consumption by factors of up to 20 and thus makes processing of large amounts of Sentinel-2 data more feasible for non-cloud users.

7. REFERENCES

[1] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, P. Bargellini, Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services, Remote Sens. Environ. 120 (2012) 25–36.

[2] S. Duprat, D. El Maalem, M. Ferrer, V. Garcia, C. Louge, M. Paulin, E. Poupart, J. Gasperi, C. Taillan, PEPS - the French Copernicus collaborative ground segment, in: Proc. 2017 Conf. Big Data Space BIDS 2017 28th-30th Novemb. 2017, Toulouse (France), 2017: pp. 134–137.

[3] GDAL Development Team, GDAL - Geospatial Data Abstraction Library, Version 2.2.2, Open Source Geospatial Foundation, 2018.

[4] I.M. Howat, A. Negrete, B.E. Smith, The Greenland Ice Mapping Project (GIMP) land classification and surface elevation data sets, The Cryosphere. 8 (2014) 1509–1518.

[5] M. Main-Knorn, J. Louis, O. Hagolle, U. Müller-Wilms, K. Alonso, The Sen2Cor and MAJA cloud masks and classification products, in: 2nd Sentin.-2 Valid. Team Meet., Frascati, Italy, 2018.

[6] U. Müller-Wilm, Sen2Cor Software Release Note, (2018). URL: http://step.esa.int/thirdparties/sen2cor/2.5.5/docs/S2-PDGS-MPC-L2A-SRN-V2.5.5.pdf

[7] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008.

[8] Z. Zhu, S. Wang, C.E. Woodcock, Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images, Remote Sens. Environ. 159 (2015) 269–277.

WORLDWIDE MULTITEMPORAL CHANGE DETECTION USING SENTINEL-1 IMAGES

Elise Colin Koeniguer [(1)], *Jean-marie Nicolas*[(2)], *Fabrice Janez*[(1)]

(1) Onera, Chemin de la Hunière, 91123 PALAISEAU
(2) Telecom ParisTech

ABSTRACT

This paper discusses the visualization and detection of changes on Sentinel-1 images. The potentials of our change visualization algorithm REACTIV implemented on the Google Earth Engine platform are shown through many examples. A detection method is proposed and evaluated thanks to a Ground Truth performed in Palaiseau (France). The method demonstrates excellent performance compared to more traditional algorithms accumulating several bi-date tests.

Index Terms— multitemporal change detection, SAR, Google Earth Engine, visualization

1. INTRODUCTION

Since the launch of Copernicus data in open source, the platform services based on Earth observation are multiplying. Among the opportunities offered, the exploitation of time series enables monitoring of the entire globe, whether for environmental, civil, industrial, defense or surveillance needs. While optical images are in widespread use, SAR images are less often exploited because of the inherent speckle noise and the difficulty of interpretation. However, for the temporal monitoring, SAR images provide clear benefits: availability whatever the weather conditions, and temporal stability due to the use of an active sensor.

For these reasons, this article considers Sentinel-1 images to visualize and detect changes or activities. In section 1, we present the visualization method REACTIV (Rapid and EAsy Change detection in radar TIme-series by Variation coefficient). Several examples of applications are shown in Section 2. Then, we propose a detection scheme in section 3. Because the temporal behaviors describing events are varied, we will also propose to go towards the classification of different generic changes. The validation of this algorithm is ensured on a local study site for which a precise ground truth has been established. Finally, we will consider in section 4 the detection of a new event occurring in the last available acquired image, before concluding. To demonstrate our approaches, the Google Earth Engine (GEE) platform was chosen for its ease of use and its capability to test at a global scale [2].

This study is part of the Research Project MEDUSA, founded by Onera (w3.onera.fr/medusa)

2. VISUALIZATION AT A GLANCE

2.1. General Principle

It is now easy to obtain a hundred Sentinel-1 images over time for any location on the globe. Though we can visualize them as a video movie, this analysis is long and tedious. For this reason, we have developed a visualization method called REACTIV which gives a colorful visualization at a glance of all areas that have undergone changes [1]. In this representation, a bright color indicates a change, and the assigned color corresponds to a particular date.

In practice, some areas belong to different orbits. The variations of incidence, of the order of a dozen degrees, can generate variations of radiometry or projections. To avoid confusion with any change, we restrict to images from a single orbit.

2.2. Theoretical key points

REACTIV visualization goes through the use of the *Hue Saturation Value* (HSV) color space. The value is encoded by the maximum intensity signal. Saturation is encoded by the temporal variation coefficient, the key parameter. Finally, the hue is given by the date for which signal intensity is maximum.

The behavior is the following one. A change in images will induce a high saturation value and so a bright color in the visualization product. On the contrary, a lack of change will result in a totally reduced color and therefore in just a gray value.

The properties of the speckle coefficient of variation [4] largely explain the success of this visualization. These properties have been justified and described in detail in [3]. They were used in particular to automatically set the algorithm.

We synthesize them now. Among the behaviors of all areas that can be described as stable or unchanged, there are either decorrelated areas of speckle that follow a Nakagami law, or areas with a deterministic component, which follow a Rice law.

In the first case (**Nakagami Distribution**), for a **stable decorrelated speckle area**, the average amplitude value does not undergo a variation over time. The temporal amplitude distribution is the same as the spatial distribution; it is parameterized by a scale parameter μ, which depends on the

doi:10.2760/848593

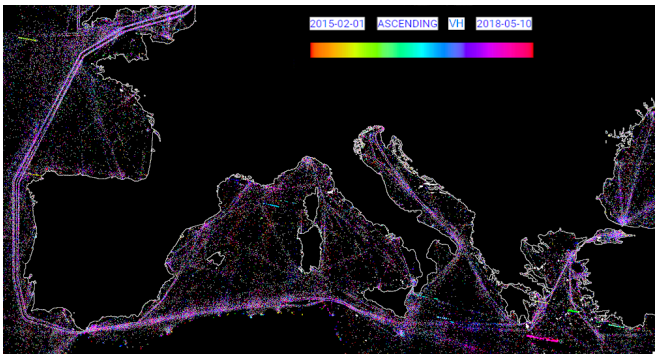


Fig. 1. Maritime roads are clearly visible on this REACTIV visualization

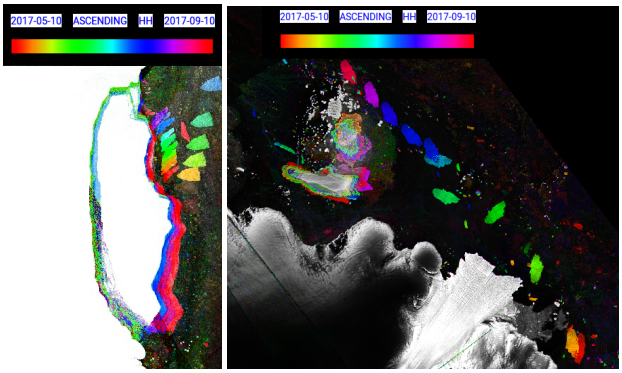


Fig. 2. REACTIV confirms calving from the Larsen C ice shelf and reveals displacements of smaller icebergs

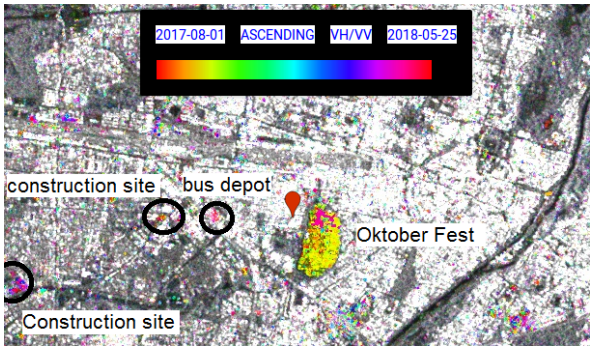


Fig. 3. REACTIV highlights Oktoberfest, held annually in Munich, Bavaria, Germany

average backscattering of the zone, and a shape parameter L, which corresponds to the Equivalent Number of Looks. These speckle realizations are decorrelated over time. In practice, all areas of forest or bare soil encountered most often satisfy this assumption, even if they are in interferometric conditions, because any small centimetric displacement cause decorrelation.

The second case (**Rice distribution**) corresponds either to a **Permanent Scatterer**, for which a very strong backscattering return dominates the signal in the resolution cell or to natural areas that can be considered as **stable correlated speckle area**, with exceptional immobility, for example, the Atacama desert in Chile, a non-sandy desert.

For stable decorrelated speckle areas, the theoretical average coefficient of variation can be expressed only in terms of L, and its variance in terms of L and N, the number of images in the stack. It is proportional to $1/N$: the larger the image number, the more precisely the coefficient of variation is estimated.

For stable correlated speckle areas, the theoretical average coefficient of variation is lower than in the previous case. It depends only on μ_c/μ, the ratio between the deterministic component amplitude μ_c, and the speckle shape parameter μ.

For both these **reference stable cases**, the inclusion of a deterministic target immediately introduces an increase of the theoretical value of the coefficient of variation. If the rupture occurs in a stable decorrelated speckle area, the gap depends only on μ_r/μ, the ratio between the amplitude μ_r of a point-event, and the speckle shape parameter μ.

Although literal expressions cannot be derived in all change cases, simulations of several scenarios have been undertaken: the mixture of two speckles, inclusion of a persistent event, etc. All these scenarios result in a coefficient of variation higher than that of an uncorrelated speckle law and thus allow to consider it for successful detection.

2.3. Applications

The implementation of REACTIV on Google Earth Engine platform has made it possible to highlight a large number of opportunities offered by the proposed visualization. For example in Fig. 1, it allows the immediate visualization of the boats. Furthermore, these results are obtained very quickly because of a purely temporal processing and not a spatial one. On GEE, it is near real time. On a large scale, we can also see the waiting areas as well as the maritime roads.

Fig. 2 shows the fracture in July 2017 of an iceberg of the Larsen glacier in Antarctica and its progression. The acceleration and rotation of the glacier appear from summer 2018.

Another application is in urban areas. In general, visualization can give an idea of the frequency of changes. In the example of Fig. 3, some changes in Munich are particularly visible. In particular, near the Congress Center indicated by the red mark, there is Theresenwiese, an open space of 420,000 square meters in the Munich borough of Ludwigsvorstadt-Isarvorstadt. It serves as the official ground of the Munich Oktoberfest. In our REACTIV product, the event is clearly visible with the yellow color associated to October month.

However, any visualization tool has its limits, and one may wish to have an automatic detection tool.

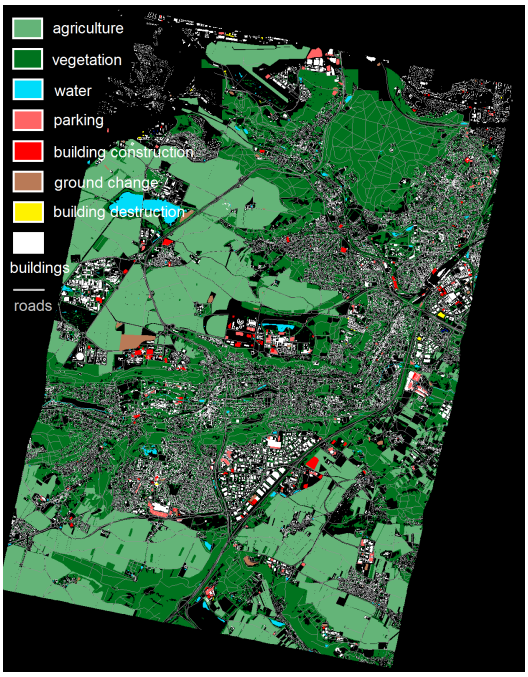


Fig. 4. Ground truth established on a test site around Saclay, France

3. CHANGE DETECTION IN TIME-SERIES

We now investigate a change detection algorithm based on the simple threshold of the temporal coefficient of variation. Qualitative analysis showed us that some changes are better seen in VH polarization, other ones in VV polarization. For this reason, we use the maximum coefficient of variation between VH and VV coefficient of variation as a decision criterion. An averaging of the criterion map is done prior the threshold, using a NL-mean filtering method. This is the only inclusion of the spatial dimension.

Quantitative performances were evaluated on the Saclay region (near Paris, France) that includes high-density construction areas. The considered area is about 15 km x 12 km. A precise Ground Truth Database has been established, first by making the difference between two vector databases (a BDTOPO base of the IGN, an OSM database) at dates before and after the observed time range, between June 2015 and June 2017. Then, all the changes found were manually validated or rebutted using optical archival images. The resulted ground truth in the Fig. 4 shows all the changes in red color, as well as parking areas, agricultural plots, water surfaces.

The ROC detection curve thus obtained on this test site is presented in Fig. 5. Each point of the ROC curve is obtained by fixing a threshold on the criterion. For a given threshold, False Alarm is given by the percentage of pixels detected that are not lying in the change class of the ground truth. The Probability of Detection is the percentage of change class objects that are detected; an object is considered as detected

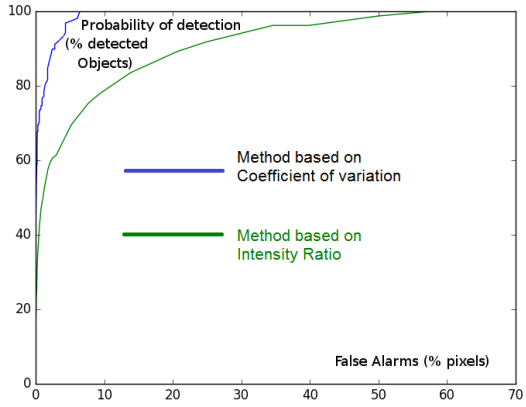


Fig. 5. Receiver Operating Characteristic curve for comparing ability of our method compared to a conventional approach to detect construction sites.

as soon as a minimum of 5% pixels of the entire object has been detected. Our criterion is compared to a more classical method where the intensity ratios are computed for each pair constituted by the first and the current image and then averaged together. The criterion map is also filtered by NL mean. This comparison shows that REACTIV method has a huge gain in terms of detection. Moreover, the method seems to be robust to a variety of different changes: some of them illustrated in Fig. 6 are different from VV and VH, and difficult to distinguish from natural changes over agricultural crops.

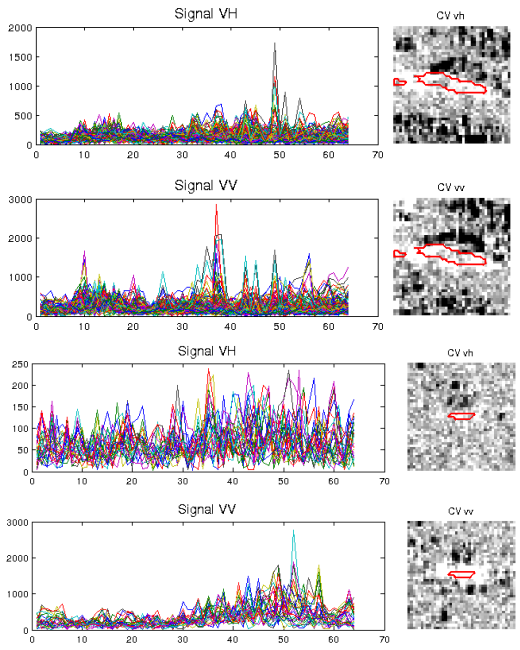


Fig. 6. Different types of time-profiles over changes

This detection method has been implemented under Google Earth Engine. One example result is illustrated in

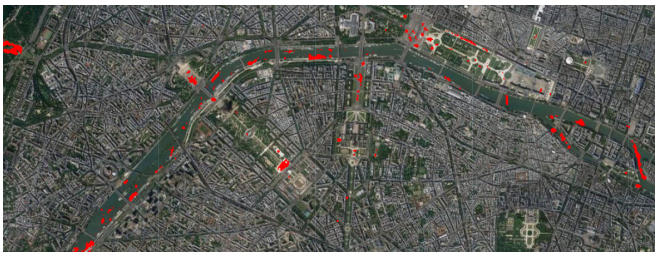


Fig. 7. Change detection in Paris over an optical map

the figure 7 for Paris in 2018. We have made the layer of pixels detected overlay on the optical image layer in red. Boat activity on the Seine can be seen, as well as the Invalides and Champ de Mars area, regularly subject to temporary installations. In practice, most detection failures are due to two main reasons:

- The high sensitivity of the response of an agricultural area. Depending on weather conditions, it is likely that the electromagnetic mechanisms and therefore the backscatter levels vary greatly from one date to another.

- So-called "point" events, which appear only on one date, and which do not appear in our Ground Truth. In practice, we have been able to verify that they actually correspond to zones of only a few pixels, for which we do not know exactly the origin of the rupture in the temporal profile.

4. DETECTION OF A NEW EVENT

For these latter behaviors, we propose to calculate the ratio of the coefficients of variation calculated with and without the maximal amplitude:

$$R_{isolated} = \frac{CV(A(k)_{k\in\{1...N\}\setminus\{k_{Amax}\}})}{CV(A(k)_{k\in 1...N})}$$

A threshold on this criterion detects in a very robust manner a certain number of temporal profiles containing a rupture on a single date. Results are most often different for VV or VH polarization, and spatially isolated. As an example, Fig. 8 shows an event detected in the area of Palaiseau (France), as well as the profile of amplitude associated, which reveals a sharp increase of the signal, without any associated explanation.

Detecting an event arriving on the last image of a stack, is a special case of isolated event detection. It can be performed by a threshold on:

$$R_{alert} = \frac{CV(A(k)_{k\in\{1...N\}})}{CV(A(k)_{k\in 1...N-1})}$$

This parameter has been implemented on the GEE platform. The set of plots detected and whose profiles have been manually checked, correspond to an event of this type, except for certain points that can be considered as artifacts: they are

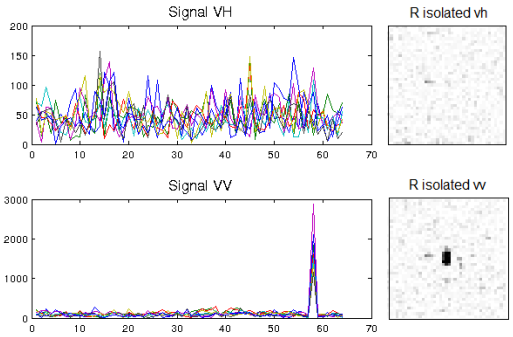


Fig. 8. A point-event detected in South of Paris.

points for which the amplitude profile is saturated. This is one of the drawbacks of the platform GEE that discards extreme values in images. In the future, using the data without changing the dynamics should avoid these artifacts.

5. CONCLUSION

In this paper, change visualization and detection methods have been proposed, using Sentinel-1 time series. The detected activities are varied: building construction/destruction, festive events, agriculture, vehicles, movements of icebergs, etc. The methods are based on the temporal coefficient of variation. Statistics properties of this criterion have made it a key parameter for deploying robust and extremely fast change detection strategies. The proposed methods have been demonstrated globally through the use of the Google Earth Engine platform. In perspective, we have proposed a method to detect a new break in any newly acquired image.

REFERENCES

[1] E. Colin-Koeniguer, A. Boulch, P. Trouve-Peloux, and F. Janez. Colored visualization of multitemporal sar data for change detection: issues and methods. In *EUSAR 2018; 12th European Conference on Synthetic Aperture Radar*, pages 1–4. VDE, 2018.

[2] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017.

[3] E Koeniguer, JM Nicolas, B Pinel-Puyssegur, JM Lagrange, and F Janez. Visualisation des changements sur séries temporelles radar: méthode REACTIV évaluée à léchelle mondiale sous Google Earth Engine.

[4] JM Nicolas. Application de la transformée de Mellin: étude des lois statistiques de l'imagerie cohérente. *Rapport de recherche, 2006D010*, 2006.

MICROCARB CNES MICROSATELLITE MISSION TO CHARACTERIZE CO_2 SURFACE FLUXES: SIZING OF THE MISSION CENTRE

C. L'HELGUEN[(1)], E. JULIEN[(1)], D. JOUGLET[(1)], P. LAFRIQUE[(1)], C. REVEL[(1)], S. CASTRO[(2)], F. HARMAND[(1)], E. JAUMOUILLE[(1)], B. VIDAL[(1)], C. PITTET[(1)], F. BUISSON[(1)], D. PRADINES[(1)]

[(1)]CNES, [(2)]THALES Services

ABSTRACT

The MicroCarb mission is currently one of the major projects in development phase led by the French Space Agency (CNES) in partnership with the United Kingdom Space Agency (UKSA). It aims at remotely measuring carbon dioxide atmospheric volume mixing ratios in order to characterize CO_2 surface fluxes. The launch is scheduled in 2021.

This article focuses on the MicroCarb mission centre. Definition of the data processing chains shows indeed that a large IT infrastructure is required to meet products availability need for the science community and to deal with the large volume of data planned to be produced during the lifetime mission. These MicroCarb major IT needs are explained and described with a first estimation of its sizing.

Index Terms— MicroCarb, CO_2, CNES

1. INTRODUCTION

Following the 21st Conference Of the Parties (COP 21), the MicroCarb mission was officially announced with the objective to remotely measure CO_2 column integrated volume mixing ratios (XCO2) in the atmosphere. The project is currently in development phase and, as the mission centre is being defined, the IT needs emerge to be high in terms of required processing cores and volume of data.

In particular, the spectral inversion of atmospheric radiometric measurements is one of the most demanding steps, as well as cloud detection.

This article first sums up the major facts about MicroCarb mission and the science objectives, before describing the processing chains integrated in the PayLoad Ground Segment (PLGS) and their impact on MicroCarb IT sizing.

2. MICROCARB MISSION

2.1. GOSat, OCO-2 and now MicroCarb

The MicroCarb mission is designed to globally monitor and characterize CO_2 surface fluxes, that is, the exchanges between sources and sinks. As CO_2 is the most important greenhouse gas produced by human activity, a better assessment of carbon fluxes is crucial for understanding the causes and consequences of climate change.

In 2009, JAXA was the first Space Agency to launch a satellite dedicated to greenhouse-gas-monitoring, called Greenhouse Gases Observing Satellite (GOSat). Then NASA launched the OCO-2 (Orbiting Carbon Observatory-2) satellite in 2014 to monitor CO_2. The Chinese missions TanSat ACGS, Feng Yun-3D GAS and, Gaofen-5 GMI also measure CO_2, even if data are not yet available. In 2021, MicroCarb will be the first European satellite for monitoring CO_2 with an expected mission duration of 5 years, with reprocessing of the data during two additional years.

2.2. Measurement precision

MicroCarb dispersive spectrometer instrument will deliver spectra which after processing by the PLGS will produce global measurements of the atmospheric concentration of CO_2 with an extremely high precision (of the order of 1 ppm, which is 0.25 %) and with a pixel size of 4.5 km x 9 km.

This performance is crucial for the success of the mission and the quality of MicroCarb data. It is indeed necessary to be able to estimate concentration gradients, which amounts to a few ppm, and to generate correct maps of CO_2 fluxes.

Spatial coverage and the repeat cycle of measurements are also important, which is why space-based observations are so valuable compared to a ground network that is difficult to deploy worldwide.

2.3. MicroCarb contributors

The MicroCarb mission combines scientific and technical teams who are bringing their respective skills to the project and working together to guarantee its success:

- CNES
- French scientific laboratories
- National and European industry
- UKSA, UK manufacturers and science laboratories
- EUMETSAT (European Organisation for the Exploitation of Meteorological Satellites).

The MicroCarb project is led by CNES in close partnership with the laboratories of IPSL, the joint research unit between CNRS (the French national scientific research centre), CEA (the French atomic energy and alternative energies commission), the LSCE (Climate and Environment Sciences Laboratory) and the LMD (dynamic meteorology laboratory).

The project is funded by the French Government (PIA).

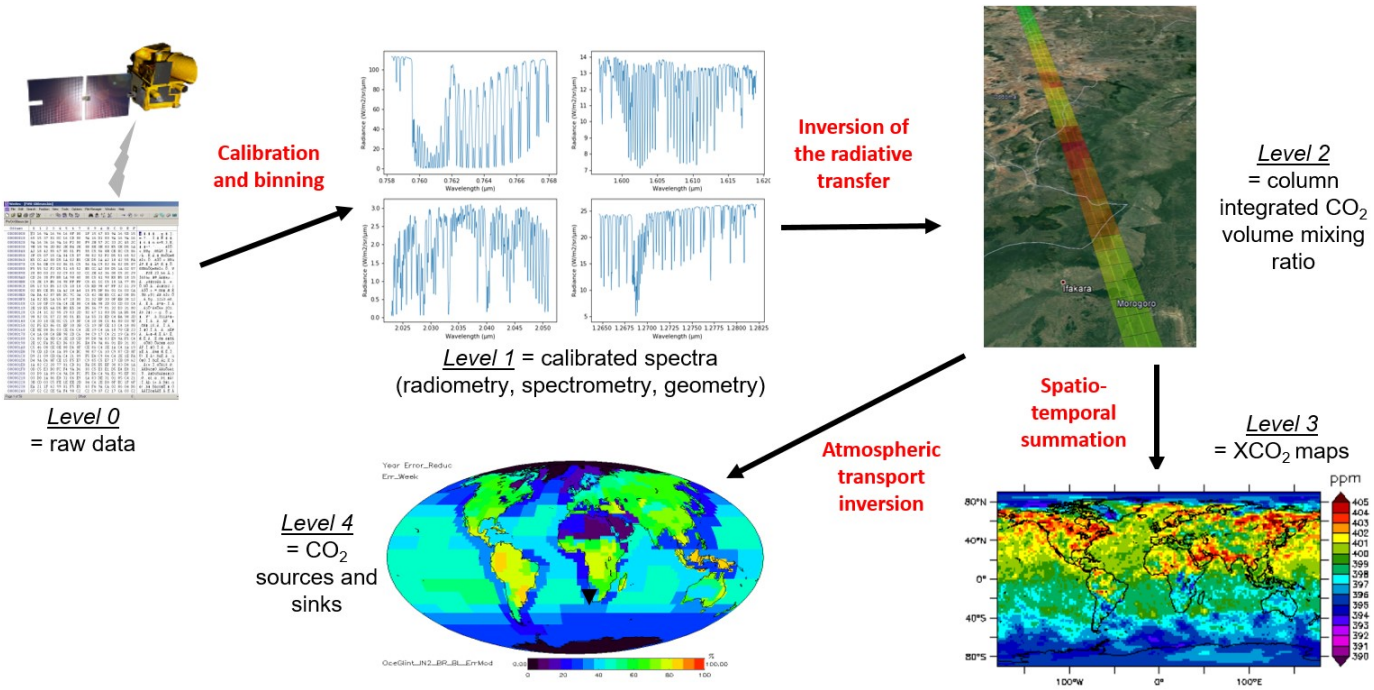


FIGURE 1 - MICROCARB DATA PROCESSING FLOW

3. SCIENCE OBJECTIVES

3.1. MicroCarb levels of product

CO_2 surface fluxes cannot be directly remotely measured (cf. Figure 1); MicroCarb will acquire atmospheric spectra in some wavelengths specific to CO_2 and O_2 (Level 1 data) which should then be inverted using a radiative transfer model (thanks to 4ARTIC software which integrates 4A/OP) to get column integrated CO_2 volume mixing ratio (Level 2 data). A spatio-temporal summation of these ratio generates XCO2 maps (Level 3 data). The surface fluxes can also be calculated from the atmospheric concentrations and the use of an atmospheric transport model (LMDZ). These products (Level 4 data) are global fluxes taking natural and anthropogenic fluxes into account.

3.2. Instrument

The instrument [2] on board MicroCarb is a compact infrared passive spectrometer operating in four spectral bands using a unique echelle grating (dispersive element) to achieve spectral dispersion and a unique NGP detector acquiring the four bands. CNES has tasked Airbus Defence & Space with developing and qualifying the instrument.
The instrument measures atmospheric spectra for the following species:

- Oxygen (O_2 at 0.76 and 1.27 µm) to retrieve the surface pressure and then normalize the computed CO_2 column concentration to dry air
- Carbon dioxide (CO_2 at 1.6 µm and 2.0 µm).

The instrument will be flown on a microsatellite built around CNES's Myriade spacecraft bus.

3.3. A precursor mission

In addition, MicroCarb aims to be a precursor of a future operational system able to accurately monitor global fossil emissions. Understanding the carbon cycle is important since it can help us to anticipate its evolution according to possible climate change scenarios.

4. PAYLOAD GROUND SYSTEM

4.1. General description

CNES is responsible for specifying and developing the mission ground segment, called PLGS (PayLoad Ground System). Its main activities are the control of the satellite and instrument, as well as the science data processing. This mission centre processes MicroCarb acquisitions from raw data to Level 3 products (monthly average maps of CO_2). The generation of Level 4 products (CO_2 surface fluxes) is computed outside of the PLGS by French scientific laboratories thanks to the French Atmosphere and Service Data Pole (AERIS).

4.2. Number of cores in routine phase

One of the main challenges to define MicroCarb PLGS is to be able to distribute Level 2 products (CO_2 concentration) to the scientific community in less than 7 days and hopefully in

doi:10.2760/848593

less than 48 hours after their acquisition in routine mode. This requirement implies to be able to generate all the Level 2 products from one day of MicroCarb acquisition in less than 24 hours.
Daily computation time for the production of Level 2 products from L0 products is estimated to be nominally around 33000 hours with one core. This estimation takes into account mandatory requirements relating to performances and optimisation.
In order to face possible delays or unavailability, the sizing of the IT architecture for processing chains from L0 to L3 is expected to be possibly double regarding to the needs. This means that about 3000 cores should be allocated to MicroCarb operational processing chains in routine phase.
These 3000 cores should be permanently allocated to MicroCarb mission from the Commissioning phase to the end of the mission (cf. Figure 2).

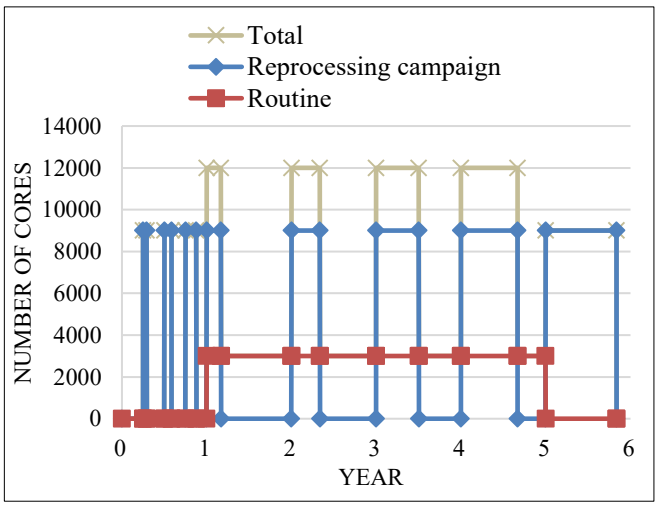


FIGURE 2 - REQUIRED NUMBER OF CORES DURING MICROCARB MISSION

4.3. Number of cores during reprocessing campaigns

In the meanwhile, parameters and algorithms will become more mature as the mission is in operation. This will require regular reprocessing campaigns in order to improve the quality of distributed MicroCarb products.
During the Commissioning phase, which is expected to last about one year after the launch, 3 reprocessing campaigns are planned to consolidate parameters and algorithms, as well as to validate the MicroCarb products before their distribution to the science community.
After the Commissioning phase, one reprocessing campaign per year is forecast with the following technical constraint; PLGS should be able to process 1 year of MicroCarb data in less than 2 months. As a consequence, the IT infrastructure necessary for reprocessing campaigns needs about 9000 cores during several months by the end of the mission (10 months for reprocessing all the data acquired during 5 years of mission; cf. Figure 2).

4.4. Large volume of data

In parallel, the volume of data generated and distributed by MicroCarb PLGS processing chains is also major. Almost 8 TBytes of data should be daily generated by the MicroCarb processing chains in routine phase.
In addition, data are not expected to be deleted after a reprocessing campaign; all the different versions of the same product due to change in parameters or software during the mission, could be available.
As a consequence, almost 4 PBytes of data should be produced by the end of MicroCarb mission (cf. Figure 3).

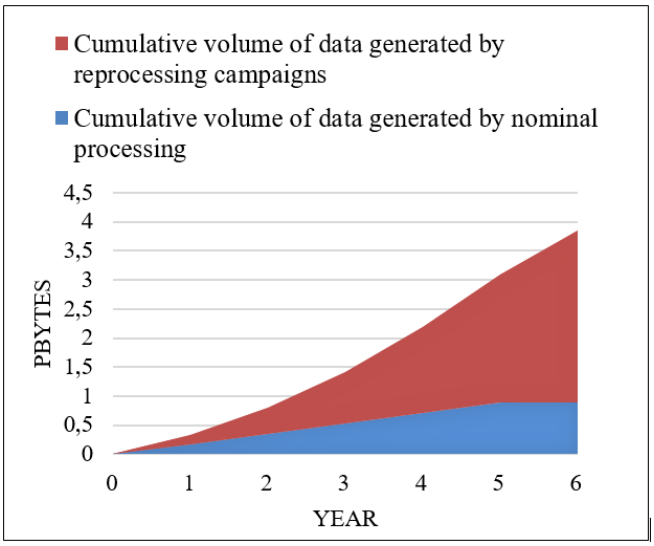


FIGURE 3 - GENERATED VOLUME OF DATA DURING MICROCARB MISSION

4.5. MicroCarb processing chains operation

MicroCarb mission centre should be able to process daily a large number of jobs (possibly 150000 jobs per day to process L0 data to L2 products) with high IT needs. In order to prepare future missions, EUMETSAT offers to operate L1, L2 and L3 processing chains on its premises by using its Processing Framework (PF) developed for Jason-CS mission. CNES could rely on EUMETSAT operational experiment over various missions for nominal processing as well as for reprocessing campaigns.

5. MAJOR ALGORITHMS

5.1. General overview

The current estimation of computation time for MicroCarb Level 1, Level 2 and Level 3 processing chains shows a predominance of 4ARTIC (4A Radiative Transfer Inversion Code). This software program enables to invert radiance spectrum to estimate the observed geophysical state. It relies on 4A/OP (Operation release for 4A, Automatized Atmospheric Absorption Atlas) [1][3] which enables to

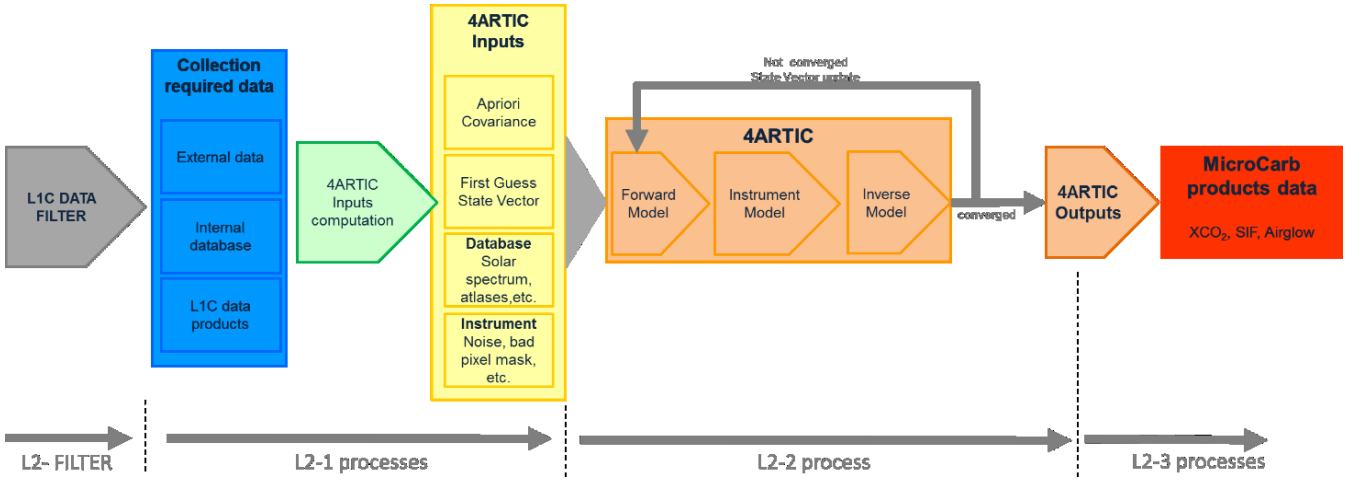


FIGURE 4 - MICROCARB LEVEL 2 PROCESSING CHAIN

compute radiative transfer. 4ARTIC is a CNES software program while 4A/OP is a CNES-LMD-NOVELTIS co-property. Diffusion in 4A/OP is computed using LIDORT [4].
About 85% of the whole processing time from Level 0 to Level 2 generation is dedicated to this software program, which is run several times in the processing chain to detect cloud and to compute CO_2 concentration. Thus, several studies are currently in progress to optimize 4ARTIC and reduce IT needs, without degrading output quality.

5.2. Cloud detection (Level 1)

Cloud detection is a major step to get products of good quality. As clouds prevent from correctly measuring and computing CO_2 concentration, cloudy acquisitions should be filtered in order not to degrade the quality of the MicroCarb distributed products.
At least 75% of the MicroCarb acquisition may be contaminated by clouds. A standard algorithm, based on radiometric analysis and comparison with reference data, should filter more than 80% of these cloudy acquisition.
In order to detect the remaining cloud contaminated acquisitions, a comparison between the computed and the expected clear sky surface pressure (from meteorological and altimetry data) will be performed. This algorithm implies to inverse part of MicroCarb O_2 spectrum with 4ARTIC. It needs to be improved and confronted to real MicroCarb data to ensure that cloudy acquisitions are correctly filtered with a low amount of bad cloud detections.

5.3. Computation of CO_2 concentration (Level 2)

Computation of CO_2 concentration in Level 2 processing chain consists of creating 4ARTIC working context before launching this program (cf. Figure 4).

As 4ARTIC has never been integrated in an operational mission centre of a previous CNES mission, several improvements still need to be implemented before MicroCarb launch. In particular, computing time might decrease, software quality needs to be improved and it should make a fair use of the CNES HPC (High Performance Computing) centre (i.e. Input/Output and memory management).

6. CONCLUSION AND PERSPECTIVES

Although MicroCarb is a probationary CNES project to measure atmospheric CO_2 concentration, its mission centre has high IT needs similar to a large-scale project.
A major challenge should be faced to specify, develop and integrate the processing chains respecting the schedule and the budget, while ensuring that high quality products are available for the science community.

7. REFERENCES

[1] L. Chaumat, C. Standfuss, B. Tournier, E. Bernard, R. Armante and N.A. Scott, “4A/OP Reference Documentation”, NOV-3049-NT-1178-v4.3, NOVELTIS, LMD/CNRS, CNES, 2012, 315 pp.

[2] V. Pascal, C. Buil, J. Loesel, L. Tauziede, D. Jouglet and F. Buisson, “An improved microcarb dispersive instrumental concept for the measurement of greenhouse gases concentration in the atmosphere”, Proc. SPIE 10563, International Conference on Space Optics — ICSO 2014.

[3] N.A. Scott and A. Chedin, “A fast line-by-line method for atmospheric absorption computations: The Automatized Atmospheric Absorption Atlas”, J. Appl. Meteor., 20, 1981, 802-812.

[4] R. Spurr, “User’s Guide to LIDORT Version 3.6”, RT Solutions, 2012.

PHENOLOGY AT CONTINENTAL SCALE: ONE SIZE DOES NOT FIT ALL

R. Goncalves [*], *V. Bakayov* [%], *R. Zurita-Milla* [$], *E. Izquierdo-Verdiguier* [#]

[*]Netherlands eScience Center; [%]University of Amsterdam, Amsterdam, the Netherlands
[$]Faculty ITC - University of Twente, Enschede, the Netherlands
[#]IVFL, University of Natural Resources and Life Sciences(BOKU), Vienna, Austria

ABSTRACT

Earth observation has a new boost with high resolution monitoring programmes and missions (e.g. Copernicus and the up-coming Landsat 9) that offer new opportunities for land cover, vegetation monitoring and phenology studies. At the same time, this data deluge brings new computational challenges that limit Scientists' search space. This challenge increases when working at Continental scales. Often the approach is to use a single product, such as a Vegetation Index (VI), and one set of pre-defined parameters to derive a new product for the entire search space. This *one size fits all* approach deviates the researcher from the truth, and thus barely exploit the real potential value of these high resolution data sets. In the context of Phenology analysis, here we quantify the variations on the *start of the season* (SOS) seen from space when using different VIs and SOS extraction methods (i.e. fitting functions and parameters). With a simple segmentation of the geographical space by Ecological regions, we show that land characteristics influence the choice of the VI and SOS extraction method. This pilot study is our seed to design a cloud-based platform that can combine different sensor data sets, algorithms and segmentation techniques to obtain a more accurate view of phenological metrics such as SOS at continental scales.

Index Terms— Start of season (SOS), TimeSat, high spatial resolution, cloud/distributed computing, Spark

1. INTRODUCTION

Phenology is the study of periodic plant and animal life cycle events and how these are influenced by seasonal and inter-annual variations in weather, climate and environmental conditions. Because of this, the timing of life cycle events vary from year to year and from place to place [1]. There are several sources of phenological data, such as ground observations, pheno-cameras and satellite sensors. The later provide remote sensing (RS) images that can be used to derive vegetation indices (VIs), which in turn can be used to characterize land surface phenology from continental to global scale. Such VIs, typically normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI), are used to extract various vegetation metrics. For instance the *start of the season*(SOS).

Motivation. Currently, there is no universally accepted method to extract phenological metrics from RS images. Applying different methods to the same RS data might result in a difference of the timing of phenological metrics of up to 60 days [2]. Moreover, multiple VIs can be used to study phenology, producing different results [3]. Also, the spatial resolution of each time series varies from sensor to sensor, making the integration of images a challenging task. With varying remote sensor and various phenology extraction methods, it is difficult to establish a uniform approach to study land surface phenology at large scale.

Satellite-derived phenology is also influenced by the temporal, spatial, and spectral resolutions of the RS images. To minimize the within-pixel variability, one can use high spatial resolution images. Moreover, in order not to penalize on the temporal accuracy, one might want to work with frequent samples; from 15-days composites to daily images. Therefore, to have best accuracies when producing phenology products, high spatio-temporal resolutions are required, resulting in a Big Data challenge.

Objectives. The main objective of this work is to study the validity and coherence of NDVI and EVI based Start-of-Season (SOS) phenology metrics at continental scale. To do such analysis, at large scale and high-resolution, we decided to use Apache Spark. This distributed computing framework was not only chosen because it allows us to handle the increased problem size, but also because it can easily scale when the problem size increases or decreases.

Using a Spark-based platform we study the validity and coherence of NDVI and EVI vegetation indices and compare them with different phenology extraction methods (fitting functions and parameters). First we study the impact of using one or another VI and fitting function on the estimated day of SOS. We compare the SOS products by calculating the average and standard deviation SOS values for all years. Then, we compare the difference between the SOS experiments for different Ecological Regions [4] showing that the results are not uniform.

In the following sections we show the impact of using NDVI and EVI, and different fitting functions to extract SOS met-

doi:10.2760/848593

rics for different Ecological Regions (Section 2). Then we introduce the platform architecture and its scaling capabilities (Section 3). Finally, we summarize our findings and present follow up activities (Section 4).

2. PHENOLOGY STUDIES

Here we study the validity and coherence of various SOS metrics derived using NDVI and EVI and different fitting functions. we compare these metrics by calculating the average and standard deviation SOS values for all years. Then we compare the difference in SOS values per ecological region.

2.1. Vegetation Indices

We used a dataset provided by Copernicus Global Land Service that spanned 19 years ($1999-2017$) by combining SPOT-VEGETATION ($1998-2014$) and PROBA-V ($2014-$present) satellite data. This dataset was used to calculate NDVI and EVI. Using the same input data minimizes any source of additional noise. The product has a spatial resolution of 1km, and is available as $10-$day composites.

The NDVI uses the near-infrared and red channels of the sensor [5]. The EVI also uses the blue channel and requires setting a set of coefficients $C1 = 6$, $C2 = 7.5$, $L = 1$, and $G = 2.5$ [6]. Both indices were calculated in a distributed mode using the computational platform presented in Section 3.

2.2. SOS with TimeSat

To compare the SOS derived from the time series of NDVI and EVI, we used 3 fitting functions provided by TimeSat for seasonality extraction: Asymmetric Gaussian (AG), Savitzky-Golay (SG) and Double Logistic (DL). The additional parameters provided by TimeSat are fixed for all experiments. TimeSat is executed in parallel over set of machines using the computational platform presented in Section 3.

Spikes in the time series, mostly caused by clouds, were removed if they were larger than 2 standard deviations from the running median (Timesat standard approach). The seasonal parameter was set to 1 indicating there is at most 1 vegetation season per year. Since noise reduction can bias the SOS estimates, the fit is adapted to the upper envelope of the data [1]. By setting the number of envelope iterations to 2, the fitting function is adjusted during the second iteration with weights derived from the first iteration. The amplitude cutoff was set to 0 to process all the available data.

For the SG fitting we wanted to remove sharp differences but without loosing the ability to capture sudden changes since in semi-arid areas the vegetation almost instantaneously responds to precipitation [7]. Therefore, the width of the SG moving windows was set to 4. For the valid data range we set the parameter to $[0, 1]$, the same set for the VIs described Section 2.1.

[1] http://web.nateko.lu.se/timesat/docs/TIMESAT33_SoftwareManual.pdf

Table 1: SOS min, max and mean value, and the stand deviation min and max

Experiment	min	max	mean	min SD	max SD
NDVI-AG	-10	204	113.8	4.92	137.80
NDVI-SG	-1.6	201.77	107.9	6.01	146.77
NDVI-DL	-8.92	205.64	113.5	5.04	137.88
EVI-AG	13	192	107.2	5.07	97.91
EVI-SG	8	186	100.6	4.98	106.81
EVI-DL	15	191	107.3	4.95	95.93

2.3. Compare SOS products and functions

Here we compare the SOS for the different VIs and fitting functions. Table 1 shows the minimum and maximum SOS values for each of the experiments. To remove outliers, the highest and lowest 2% of values were discarded. The average value per pixel was calculated for the whole range of the time-series from 1998 to 2017. Furthermore, to assess the seasonal spatial change across all the years the standard deviation (SD) was calculated for the same period.

Results show that NDVI has a higher overall spread than EVI. This results in lower minimum values, which in some cases can go negative indicating SOS in the previous year. The spread between the min. and max. observations is about 10 days higher when compared to the EVI. The results for the AG and DL fitting functions generally show similar values when compared per index, however the SG function shows earlier SOS by several days, more noticeable in the EVI-SG experiment.

For the SD, the minimum boundary is fairly consistent across experiments ranging from 4.92 to 5.04 days, however we have higher deviations in the max boundary with 95.03 for EVI- AG to 147 for NDVI-SG. We observe much higher SD values for the NDVI ($30 - 40$ days) than for the EVI. We also observe similar AG and DL results across VIs, and higher SD values for SG when compared with the rest of the fitting functions.

Figure 1 shows a histogram of the SOS mean value distribution for each of the experiments. Results are fairly consistent, with most of the values clustering around the 100^{th} day. The histograms are skewed to the left indicating a higher percentage of values greater than 100. In the $150 - 200$ days range, the biggest differences are observed between VIs, where the NDVI experiments show higher SOS values.

To better understand the differences between each VI and fitting function, the average values were separated per ecological region. An eco-region is a recurring pattern associated with characteristic combinations of soil and landform that characterize a given region. Desserts, forests and corn plains are examples of eco-regions. By observing Figure 2, we can better judge the spatial difference among the experiments. In the central and Midwest regions, minimal differ-

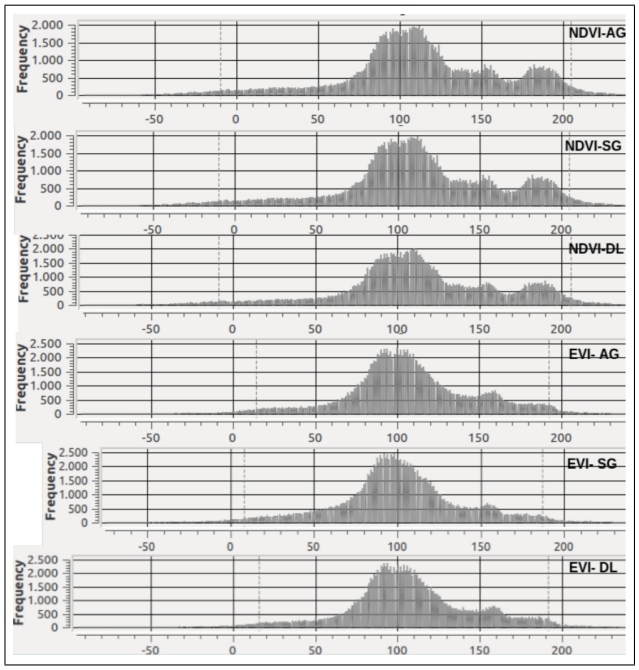


Fig. 1: Histogram with the mean difference in each experiment

ences are observed across VIs and functions $(0-10)$. Larger variations start to show across other regions. For example in Florida in an eco-region classified as Eastern Temperate Forests, the biggest differences are observed between the VIs. In this eco-region the NDVI-based SOS values are much later (by 30 to 70 days). Another interesting region is located in the West Coast, where the difference is between 0 and 20 days. However, the NDVI-based SOS leads to earlier predictions in the deserts $(0-30$ days), although later predictions were found in the Chihuahuan desert.

3. COMPUTATIONAL PLATFORM

Determining the "right" fitting function and set of parameters for SOS long-term studies at high spatial resolution and at continental scales is a computational challenge that traditionally could not be tackled by single researchers. In this work we use a cloud-based solution based on Apache Spark [2] to perform our analysis. With the data stored in the original file formats, such as GeoTiff and HDF, users are able analyze the data through Jupyter notebooks running either Python, R or Scala. These notebooks are not only used to share results among scientists but also as a provenance method for scientific results.

3.1. The platform's architecture

Our computation platform extends the one presented in [1]. Here we briefly recap its architecture organized in three layers: storage layer, processing layer and JupyterHub services for user-interaction.The storage layer offers two flavors of storage, file-base by Hadoop Distributed File System (HDFS), and object-based by Amazon S3 service. For local environments we use Minio, an open source object storage server with Amazon S3 compatible API, to avoid application rewrite when moving to a cloud provider. HDFS is used by Apache Spark to exploit data locality and to store intermediates to avoid re-computations. The object storage is used to store the phenology data products and other remote sensing data products.

[2]https://spark.apache.org

3.2. Distributed SOS computation

Several software packages allow calculating phenological metrics from regular time series of RS images. For instance, TimeSat [7], Spirits [8], and Sen2Agri [9]. These packages need a certain degree of integration and customization to be deployed on a cluster environment. TimeSat, a software package for analyzing time-series of vegetation index derived from satellite sensors, was our choice because it was easy to separate the module for SOS generation from the rest of the functionalities. In addition, it has a way to define the processing area as a subset of the whole area of interest, allowing us to use spatial partitioning to distribute load in Spark. Also, it offers several fitting functions and options that we evaluated during our phenology analysis.

TimeSat requires POSIX file system to read input data and output data. With the input being generated by Spark stored in S3, and the TimeSat output to be used by Spark, we simply mounted S3 buckets locally on each machine. With the data partitioned over several tiles, the extent and meta-data provided as a "settings file" for TimeSat, the computation of SOS was executed in parallel over a series of VMs in the cloud.

3.3. Data partitioning

Spatial partitioning was the chosen strategy since TimeSat requires the time-series to have minimum length 3 years and it is set to find $n-1$ seasons, where n is the number of years. Hence, a temporal partitioning schema would drastically impact on our partition granularity. To assess the optimum number of tiles and tune the load distribution in the cluster, we ran a set of experiments where the number of tiles were roughly doubled on each experiment. We choose to keep the dimensionality ratio between height and width of a tile for the cost of not precisely doubling the number of tiles for each subsequent experiment.

The experiments were run on 19 years of NDVI data with the TimeSat AG fitting function. The files were cached on each of the nodes. The test setup was 4 nodes with 4 cores each having 1 worker and 1 executor per node. Each node has 4 cores resulting in a total of 16 cores in the cluster. The nodes memory was set in $3:1$ ration; 12 GB for Spark jobs and 4 GB left for TimeSat processing. The results in Table 2 show that the best number of tiles is 1024.

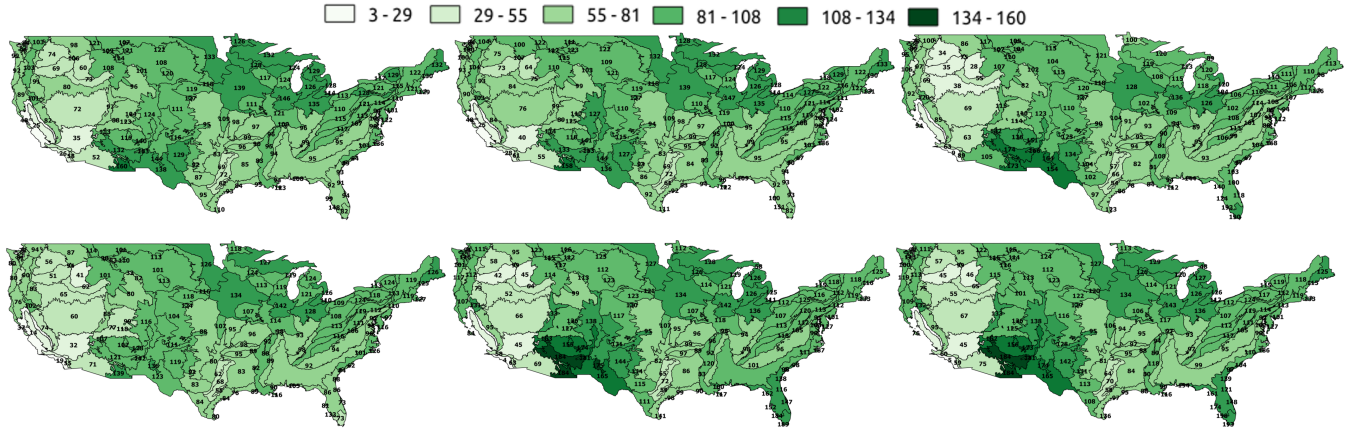


Fig. 2: Average SOS over the $1998 - 2017$ Period on Ecological Regions. Row-wise **EVI NDVI**; Column-wise **AG SG DL**.

Table 2: Duration for different tiling sizes

Tiles	30	64	121	256	529	1024	2025
h:mm	14:0	7:30	6:43	6:50	6:22	6:20	6:46

3.4. Platform Scalability

Due to high degree of data parallelism we spatially partition continental USA and then used Spark to distribute the work over several instances of TimeSat. To achieve that we created an RDD holding the input parameter for the TimeSat executable, that is the path to a "settings file" that defines the area on which to produce results. To invoke the execution of the external TimeSat process and imitate how Timesat is run from the command line we used the Scala system process call for each RDD record. With the data loaded into Spark's memory-based structures, distributed task scheduling and fault-tolerance was then handled by Spark.

Preliminary scalability performance tests show that the platform is able to scale horizontally (add more nodes) and vertically (add more cores per node). The computation times for SOS using NDVI and as VI and AG as fitting function are in Table 3. The fact that the scaling is not linear is due to amount of work per tile and due to the non-uniform geographic shape of USA (i.e.some tiles contain less valid data than others). In future work we plan to design a dynamic partitioning solution based on the size of the dataset and the available resources.

Table 3: Platform scaling experiment

Nodes x Cores	4x4	4x8	4x12	8x4	12x4
Time (h:mm)	7:11	3:40	2:33	3:43	2:36

4. SUMMARY AND FUTURE WORK

This works presents a computational platform to study the validity and coherence of SOS products derived from two common VIS and using 3 popular fitting functions. The different products were compared by calculating the average and standard deviation SOS values for all years and by analyzing differences per Ecological region.

Future work will focus on including additional RS data sets and on performing a more detailed study of various other methods to extract land surface phenology metrics. We think that this solution will allow individual scientists to efficiently derive phenological metrics on their own while benefiting from the increasing availability of global RS data at very high spatial resolutions (10 to 30m).

5. REFERENCES

[1] R. Zurita-Milla, R. Goncalves, and et al., "Exploring vegetation phenology at continental scales : Linking temperature-based indices and land surface phenological metrics," pp. 63–66, 2017.

[2] M. A White and et al., "Intercomparison, interpretation, and assessment of spring phenology in North America estimated from remote sensing for 1982–2006," *Global Change Biology*, 2009.

[3] L. Hongjun and et al., "Comparison of NDVI and EVI based on EOS/MODIS data," *Progress in Geography*, p. 26, 2007.

[4] J. M. Omernik and et al., "Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework," *Environmental management*, vol. 54, no. 6, 2014.

[5] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Rem. Sensing of Env.*, 1979.

[6] A. Verhegghen, S. Bontemps, and P. Defourny, "A global NDVI and EVI reference data set for land-surface phenology using 13 years of daily SPOT-VEGETATION observations," *International Journal of Remote Sensing*, vol. 35, no. 7, 2014.

[7] P. Jönsson and L. Eklundh, "TIMESAT- a program for analyzing time-series of satellite sensor data," *Computers & Geosciences*, vol. 30, no. 8, 2004.

[8] F. Rembold and et al., "Remote sensing time series analysis for crop monitoring with the SPIRITS software: new functionalities and use examples," *Frontiers in Env. Sci.*, vol. 3, pp. 46, 2015.

[9] S. Valero and et al., "Production of a dynamic cropland mask by processing remote sensing image series at high temporal and spatial resolutions," *Remote Sensing*, vol. 8, no. 1, pp. 55, 2016.

PRODUCTION OF COPERNICUS HIGH RESOLUTION LAYERS 2018 – A LARGE-SCALE LAND COVER MAPPING ENVIRONMENT ON MUNDI (COPERNICUS DIAS)

Marcus Sindram[1], Gernot Ramminger[1], Martin Ickerott[1], Carolin Sommer[1], Anna Homolka[1], Christoff Fourie[1], Christoph Rieke[1], Cornelia Storch[1], Benjamin Mack[2]

GAF AG[1], MunichRE[2]

ABSTRACT

GAF AG supported by its partner's e-GEOS, GeoVille and SIRS has developed an end-to-end cloud-based processing environment for the production of the Copernicus HRLs 2018, a large-scale land cover mapping activity covering an area of ~5.8 Mio km². It is partially based on developments undertaken in the H2020 project ECoLaSS, targeting evolutions in the CLMS, specifically the High Resolution Layers. The processing environment is currently deployed on the DIAS Mundi platform, which offers a scalable, high-performance cloud infrastructure in the Open Telekom Cloud (OTC) with direct access to various earth observation data archives. Fast access to EO data provides CUBEO, e-GEOS scalable pre-processing and Data Cube platform developed in cooperation with MEEO. The workflow is based on a highly automated, multi-stage land cover classification using dense time series of Sentinel-1 and -2 data. It is characterized by the use of spatial-temporal features, bio-geographical regions for product harmonization as well as iterative optimization techniques, e.g. active learning for sample refinement. The main challenges consist of the efficient handling of pre-processing and classification of the vast data volumes as well as ensuring the consistent product quality at European scale despite strong regional differences in geographical settings.

Index Terms— Big Data, Data Cube, Land Cover Mapping, Large-Scale, Cloud, Mundi, Active Learning, Sentinel, Copernicus, Land Monitoring Service, ECoLaSS, High Resolution Layers

1. INTRODUCTION

Ever increasing volumes of earth observation and in-situ data reinforce the need for cloud-based data processing solutions and the development of novel approaches to large-scale land cover classification. Just a few years back, approaches that focused on classifying a single "best scene" or ideally a few, optimally cloud-free satellite images per image tile, were an established option. This was despite the inherent problems of the approach in regions with frequent cloud cover and extensive manual effort for scene selection and post-processing at scenes tile borders. Combined with the potential of today's readily available, dense time series data, these approaches seem outdated. However, optimal exploitation of the time series information for the classification of large areas spawns new challenges, most notably related to efficient processing and data handling, the required technical infrastructure, as well as sensor data fusion and spatially consistent product quality.

In December 2018, GAF AG, GeoVille, e-GEOS and SIRS were awarded by the European Environment Agency (EEA) to produce the Copernicus High Resolution Grassland & Forest Layers 2018. Based on high to very high resolution satellite imagery, including full time series of ESA's Sentinel-1 and -2 satellite sensors, the team will perform within 12 months an update for the year 2018 and change mapping between the years 2015 & 2018 for the EEA-39 countries covering an area of approx. 5.8 Mio. km². All thematic layers provide dedicated information on current environmental conditions at 10m and major change trends at 20m spatial resolution. Final products are freely available by EEA via the Copernicus Land Monitoring Service portal (http://land.copernicus.eu).

The workflow is partly based on developments achieved in the H2020 project ECoLaSS (www.ecolass.eu) and currently updated and implemented on the Copernicus Data Access and Information Services (DIAS) platform Mundi (www.mundiwebservices.com). The infrastructure of Mundi is based on the Open Telekom Cloud (OTC), a secure and scalable Infrastructure as a Service (IaaS) solution based on OpenStack and hosted by T-Systems in Germany. Because of the short production period given by the HRL project, the workflow has a strong focus on automation & efficient processing in a fully cloud-based environment together with iterative classification optimization. Desired land cover layers are, among others, the dominant leaf type and grassland status layers, land cover class-specific density layers, and various land cover change layers.

The following chapters describe the technical infrastructure and implementation of the processing environment on the Mundi cloud platform, the input data and employed data pre-processing, the data cube technology as well as the thematic classification methodology and optimization techniques.

2. DIAS MUNDI CLOUD PLATFORM

For the production of EO based Land Cover maps based on dense time series data and sophisticated methodologies, a scalable and performant, cloud-based data-management and processing environment is required. Due to the powerful

compute instances and the standardized access to the complete Sentinel-1 and -2 archives, the Mundi cloud platform was selected as the technical infrastructure for this project (Fig. 1).

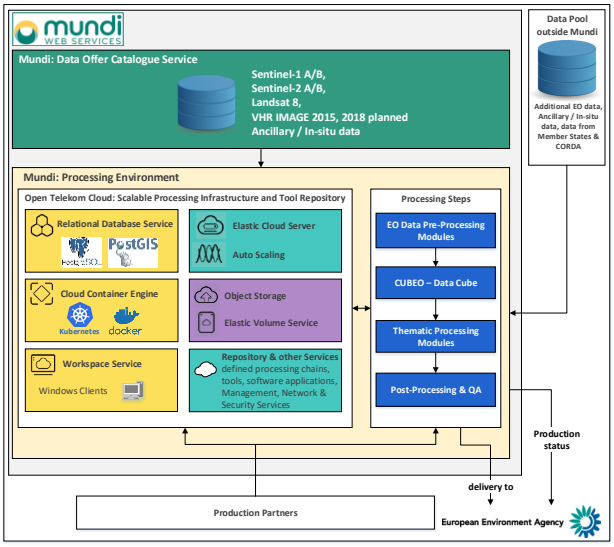


Fig. 1: Data Management & Processing Setup on Mundi

The platform, developed by ATOS is established on the infrastructure of the Open Telekom Cloud (OTC), which is hosted by T-Systems. The Mundi Data Catalogue Service provides access to Copernicus and other remote sensing and ancillary datasets, and enables customizable queries through OpenSearch and OGC CSW protocols. Data can be processed and analyzed on-demand with a broad range of tools (e.g. relational database services, container engine, workspace services, data cubes etc.) and scalable cloud services (e.g. elastic cloud server, object storage, elastic volume service, repository support). The technical implementation of the production system of this project includes a centralized data storage and relational database using PostgreSQL/PostGIS for file and process management. The database contains all relevant metadata, production status, reference and training samples and other data management & production relevant information. To enable sufficient scalability of the pre-processing, thematic classification and post-processing chains, the workflows steps are modularized based on Docker. The deployment, scaling and management of the micro services are orchestrated via the Kubernetes orchestration framework.

The Mundi cloud platform also enables a fully centralized processing approach, where all production partners will be able to use the same processing chains and software tools and have access to all virtual machines, databases as well as Elastic Cloud Services (ECS) and object storage systems. This allows convenient sharing of production tasks between the partners, enabling the exploitation of regional thematic expertise. The cloud-based approach thus supports the achievement of the project requirements regarding time schedules, product consistency, reproducibility and long-term sustainability of the desired land cover products.

3. METHODOLOGY

3.1. Input data

The main input data consists of Sentinel-1 and -2 time series data (with Landsat 8 as fall back option for gap filling). Sentinel-1 imagery will especially benefit the production in areas where too few cloud-free optical high-resolution images are available. The extensive imagery archive is filtered by temporal and minimum cloud cover criteria. The temporal subsets account for the vegetation period (growing season) which is most relevant for the successful delineation of the land cover products. The Sentinel datasets are complemented by additional very-high resolution aerial imagery and ancillary datasets, e.g. DEMs the 2012 and 2015 Copernicus High Resolution Layers (HRL) as well as various in-situ datasets, e.g. land parcel information system (LPIS) data.

3.2. Pre-processing

The pre-processing is targeted at ensuring the spatial and temporal homogeneity of the time series data. It aims at removing inhomogeneous image effects from different atmospheric, illumination and topography-related conditions, establishing a more direct linkage between the data and the biophysical phenomena on the ground. Due to the large-area and high spatial and temporal resolution imagery, the processed data volume requires an automated, scalable and efficient pre-processing chain for the generation of analysis-ready data. Sentinel-2 Level-1C and Level-2A time series data as well as Sentinel-1 Single Look Complex (SLC) data, generated by ESA and available through MUNDI will be the main input for the production process. Figure 2 provides an overview of the optical and SAR satellite data pre-processing and provision of data within the CUBEO Data Cube platform.

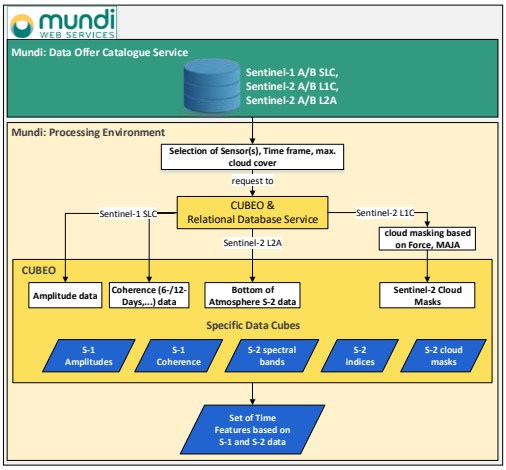


Figure 2: Optical & SAR satellite data pre-processing workflow

Unfortunately, the quality of cloud- and cloud shadow masks provided via L2A data from ESA is usually not sufficient for further thematic processing steps. Therefore, separate cloud masks are processed based on an enhanced Fmask algorithm implemented in Force (Frantz et al. 2018) or MAJA (Hagolle et al. 2010) using Level 1C data. For further thematic processing, spectral information of atmospherically corrected (BoA) Sentinel-2 L2A data is ingested in the processing framework as well as Sentinel-1 SLC data to derive reliable intermediate products such as the Radar Cross-Section backscatter (RCS) and the interferometric coherence of the vegetation period time series. All satellite data, cloud masks and derived indices are stored in different Data Cubes for performant access via WCS. The Data Cubes are included in e-GEOS scalable pre-processing platform called CUBEO (Corsi et al. 2018), which was developed in cooperation with MEEO from Italy. CUBEO has been developed using a cloud-oriented approach (containers, process orchestration, auto-scaling) to maximize the horizontal and vertical scalability of the data preparation pipelines and of the WCS data access endpoints.

3.3. Spatio-temporal features

From the spectral bands and derived indices, several spatio-temporal or "time" features and are computed. They describe distinct spectral, temporal and phenological properties that summarize the spectral evolution of each pixel over time. This allows the characterization and differentiation of distinct patterns of seasonal, short-term and long-term changes such as induced by phenological cycles, extreme events or human activity (Valero et al., 2016) and helps to differentiate and classify the desired land cover classes. Compared to single-scene classification approaches, time-features are preferable since they do not require manual scene selection and offer greater robustness against thematically irrelevant effects (e.g. seasonal changes in illumination conditions, undetected clouds and cloud shadows).
The range of potentially valuable time-features comprises well-known statistical parameters such as the min, max, mean and percentiles. Also, for more complex time-features, multiple sliding windows are computed and used to iteratively update the desired feature over the course of the time series. Time-features can be computed for different temporal subsets of the time-series allowing a flexible adaptation to changes in land cover characteristics and phenological cycles in different bio-geographical regions.

3.4. Bio-geographical strata

Instead of classifying on a fixed satellite tile grid system, the project workflow proposes a regional stratification strategy based on bio-geographical regions with similar landscape characteristics (Fig. 3). If required, these production units (level-1) can be further divided into sub-production units (level-2). This stratification approach helps improve the overall thematic accuracy by application of specifically tailored classification models and allows for flexible enhancements of the classification by further subdivisions.
It also helps limits discrepancies and time-consuming subsequent post-processing corrections at grid unit borders by encouraging classifier decision fusion in overlap regions.

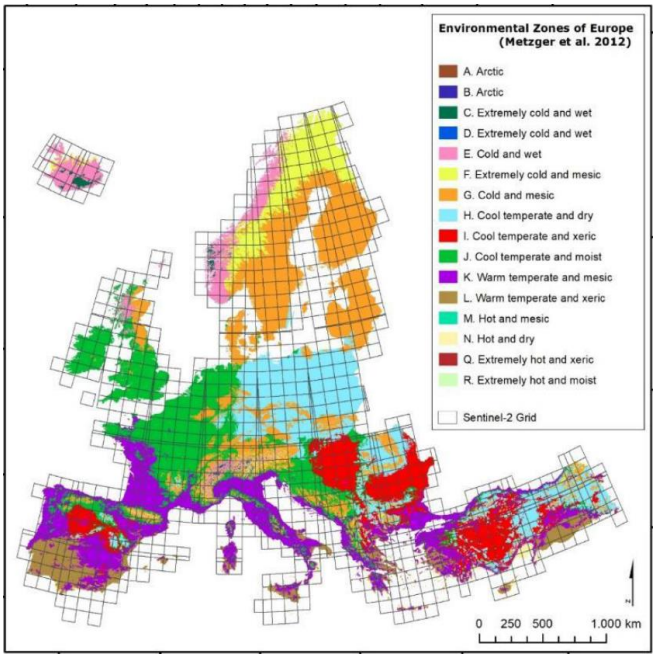


Fig. 3: Environmental Zones of Europe as proposed by (Metzger et al., 2012) and the Sentinel-2 tiling grid

3.5. Classification workflow

The main inputs to the actual classification workflow are the training and test samples, the spatio-temporal features computed from the Sentinel-2 and Sentinel-1 time series, as well as the bio-geographical strata (Fig. 4).
The sampling process is initially based on stratification via existing auxiliary land cover datasets. The sample distribution within the class area is then refined through a secondary stratification via a feature space clustering based on the derived time features. This reduces the chance of missing spectrally distinct but low-area land cover types. The workflow uses the non-parametric random forest classifier which equals or outperforms the accuracy of other methods for a wide range of classification problems (Fernández-Delgado et al., 2014).
Although the project workflow is highly automated, it features dedicated break points, which will allow adapting and optimizing the classification according to regional diversity and local phenomena. The initial classifications on the test samples, respectively full raster extent yield automatically generated classification reports with various metrics on accuracy and regional variations as well as operator guidelines. This is followed by the operator either accepting the classification as the final result or triggering a

multi-stage pipeline of optimization, retraining and evaluation until satisfactory results are achieved for the respective region. The operator can deploy multiple optimization methods that may include active learning for iterative sample optimization, splitting the regional strata into further sub-strata, renewed feature selection or model hyper-parameter tuning. **Active learning** generally enables the use of a much lower number of samples to achieve similar or better classification accuracies. The assumption for employing the active learning framework is that the sampling should be focused on areas for which the initial model is particularly uncertain. Since already a few of these difficult samples (often from underrepresented classes) can improve the accuracy of the classification significantly, the effort for time-intensive manual labelling process can be significantly reduced. Which samples are selected for labelling depends on the distribution of the class probabilities (as predicted by the machine learning model) which are summarized with a specific uncertainty / confidence metric such as breaking ties or entropy (Settles, 2012).

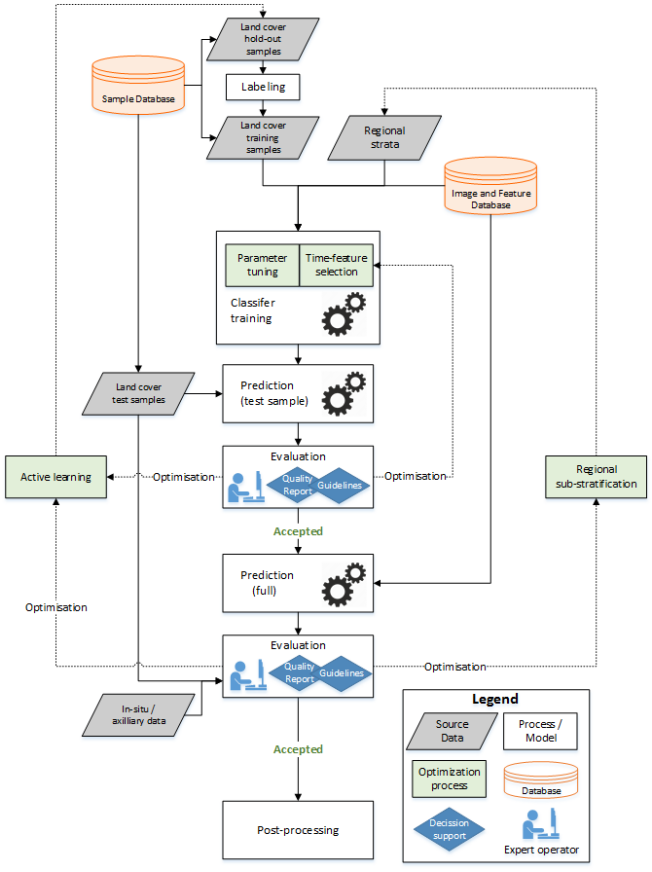


Fig. 3: Classification including the quality evaluation, optimization steps and decision mechanisms.

Additional **post-processing** for the classification result can include automatic and manual corrections, e.g. the relative calibration of the full raster classification results from adjacent and partially overlapping bio-geographical regions and sub-regions by **model decision fusion**. Horizontal model stacking is the combination of multiple models from neighboring regional strata at the same level. This is particular useful to harmonize the transition in overlapping areas of adjacent production units and select for each pixel the result from the more reliable model. Vertical model stacking is the combination of multiple models from different levels of the same regional strata. For example, a classifier that is optimized for a specific level-2 region (for pixels of classes that are only minorities considering the full level-1 region) typically yields a better accuracy for most pixels but not necessarily for all. A fusion of the most reliable classifications from different levels can hence be used to harmonize and improve the results.

4. ACKNOWLEDMENT

The research leading to these results has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement no 730008.

5. REFERENCES

[1] Corsi, M., Grandoni, D., Biscardi, M.A., Volpe, F., Pistillo, P., Mantovani, S., Cavicchi, M. Ferraresi, S. & Barboni, D., 2018: CUBEO: a scalable pre-processing and Data Cube platform for Geoinformation application services. Presentation @ESA Phi-Week, 15.11.2018.

[2] Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research 15, 3133–3181.

[3] Frantz, D., Haß, E., Uhl, A. Stoffels, J. & Hill, J., 2018: Improvement of the Fmask algorithm for Sentinel-2 images: Separating clouds from bright surfaces based on parallax effects. Remote Sensing of Environment, Volume 2015, 15. September 2018, 471-481.

[4] Hagolle, O., Huc, M., Villa Pascual, D. & Dedieu, G., 2010: A multi-temporal method for cloud detection, applied to Formosat-2, Venμs, Landsat and Sentinel-2 images. Remote Sensing of Environment, Volume 114, Issue 8, 16.08.2010, 1747-1755.

[5] Metzger, M.J., Bunce, R.G., Jongman, R.H., Sayre, R., Trabucco, A., Zomer, R., 2012. A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring. Global Ecology and Biogeography 22, 630–638.

[6] Settles, B., 2012. Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 6, 1–114.

[7] Valero, S., Morin, D., Inglada, J., Sepulcre, G., Arias, M., Hagolle, O., Dedieu, G., Bontemps, S., Defourny, P., Koetz, B., 2016. Production of a dynamic cropland mask by processing remote sensing image series at high temporal and spatial resolutions. Remote Sensing 8, 55.

CLOUD COMPUTING CASE STUDIES AND APPLICATIONS FOR THE SPACE AND SECURITY DOMAIN

Anca Popescu, Adrian Luna, Sergio Albani, Vasileios Kalogirou, Jean-Philippe Robin

European Union Satellite Centre, Apdo de Correos 511, 28850 Torrejón de Ardoz, Spain

ABSTRACT

The Space and Security domain is heavily relying on effective processing of big geospatial data to provide value added products for decision making. Traditional information technology approaches have important limitations in areas such as infrastructure, data and user management. With the prevalence of commercial and public cloud computing solutions for EO applications, there is an opportunity for the adoption of these technologies in the Security domain. This paper revises the typical geospatial intelligence production workflow, identifying the main areas that can be positively impacted by the use of cloud solutions and showcasing two implementations at different TRL within one of EU key operational agencies, the European Union Satellite Centre (SatCen).

Index Terms— Cloud Computing, Space and Security, Change Detection, MTC, NextGEOSS, RTDI

1. INTRODUCTION

The Space and Security domain is an integral part of the European Space Policy, and includes Security from Space and Security of Space. Driven by a positive landscape shaped through European policies [1], including the EU Cloud Initiative, EOSC, Digital Single Market [2], EU Comm(2012) 529, EU Space Strategy [3], in the last years there has been a growing tendency in the democratization of the use of space, with a greater participation of industry and academia and an enhanced institutional support, shaping a complex ecosystem of sensors, platforms, applications, and tools.

Triggered by the evolution of web technologies, high performance computing and advances in data science, EO has become a ubiquitous resource, with a plethora of downstream services being proposed on top of the data (e.g. Sinergise, AWS, Google, Planet, and others). One of the main innovations is in the area of data provisioning and in how users (including Space and Security stakeholders like Member States, missions and operations, international organizations) can access relevant information. The new paradigm for knowledge discovery from geospatial products is user-centric, where dedicated applications can trigger complex processing performed at the data location (often distributed and heterogeneous), with minimal expert user intervention. This is enabled by free and open data and tools provided by programmes such as Copernicus, and the availability of cloud computing solutions. For the Security domain, this context poses both opportunities and challenges. The growing need to process heterogeneous EO and collateral big data and share information faster and globally, as well as the need for more effective EO logistic chains, are main drivers for institutions in the geospatial service provision field to adopt cloud computing approaches to their production chains. Challenges are posed by legacy ITC systems, difficulties in implementing official changes, and growing cyber security concerns. This paper discusses the opportunities and challenges of adopting cloud-based solutions for the Security domain and showcases two successful scenarios implemented at the European Union Satellite Centre – one of EU essential operational assets, contributing to the security of countries. Section 2 discusses the advantages for GEOINT providers of moving to cloud, section 3 presents the typical technical requirements to be considered by security institutions looking to implement this infrastructure change, while section 4 details the use cases: an implementation in the operational production chain to demonstrate geospatial information product generation and an implementation at lower Technology Readiness Level (TRL) to demonstrate cloud service provision.

2. IMPACT OF CLOUD ON GEOINT WORKFLOWS

To understand where cloud computing could improve the provision of products and services resulting from the exploitation of space assets and collateral data to support decision making and actions in the field of security, Figure 1 presents a brief overview of a typical GEOINT operational chain. Following the tasking process, the GEOINT provider performs the feasibility analysis of the task, identifies data sources and requests necessary data from the data providers. Then, upon data reception, performs the analysis using dedicated tools and finally the product is delivered through dedicated channels to the requesting entity. EO satellites deliver daily massive amounts of spatio-temporal observations [4]. During the Feasibility Analysis, the GEOINT provider decides which data type is most suitable to the scope of the application (e.g. field support, planning or decision making), accounting for constraints including availability, price, provider agreements, licensing, data access and delivery.

For the stage of Data Request (including the whole chain from discovery to order placement and retrieval), the main advantages of using cloud-based solutions versus classical IT

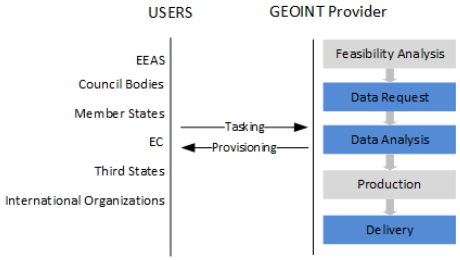


Figure 1. Workflow for the provision of GEOINT products. In blue the stages for which cloud solutions could have most impact.

approaches are: ensuring interoperability and promoting use of common standards to enable a unitary treatment of all data types and formats, having proper resource management to enable accessing archives as well as triggering new acquisitions and large volumes, maintaining replicable and reproducible procedures and approaches to minimize response time and optimize analysis and production times (e.g. re-use of data, systematic data pre-processing) [5].

Furthermore, by providing an agile alternative to on-premises systems, cloud computing can impact the Data Analysis and Production stages especially through built-in capabilities for scalability and elastic provisioning of IT infrastructure, which mitigates complex procurement and integration processes, as well as maximizes focus on added-value tasks for specialized staff, since updates and monitoring tasks are applied transparently. Bringing processing close to the data and scalable storage capacity are important advantages with respect to the execution of routine analyses (e.g. generation of high level products) and recurrent processing (e.g. for long-term monitoring tasks).

The main impact on the product Delivery stage is on the need for secure and global data sharing. Cloud computing can facilitate the secure transfer of information between relevant parties (through controlled access protocols) even if they are in different parts of the globe, boosting the ability for faster response and enhancing interoperability.

3. TECHNICAL REQUIREMENTS FOR CLOUD ADOPTION IN SPACE AND SECURITY

Several institutions and industry players in the Space and Security domain are already provisioning EO Cloud services or are on the way to provisioning them. At EU level, the biggest effort to provide seamless access to geospatial and collateral data, computing power and applications, is put on the five Copernicus DIAS (Data and Information Access Services) which are expected to boost the exploitation of EO products, increase the industrial participation and expand the cross-sectorial and cross-boundary impact of geospatial data. Functionally, the DIAS architecture consists of a back office (a scalable computing environment in which users can build and operate their own services, with unlimited, free and complete access to Copernicus data and information, and any other data that may be offered by the DIAS provider), the DIAS integration services (orchestration of interactions, development tools and services), and the front office (service provision by third parties)[6]. Furthermore, the European Space Agency (ESA), through its programmes, supports the development of Thematic Exploitation Platforms (TEPs) that are using a broad range of EO data, including Copernicus, and allow the processing of (not only) satellite data for specific themes.

In the Space and Security domain, the move from legacy applications to cloud solutions, relies on a shared responsibility scenario between the application owner (operational security actor) and the cloud provider, with established means for information security controls. The main challenge comes when it is required to deal with Classified Information (CI), which in EU terms refers to “any information or material designated by an EU security classification, which, if disclosed without authorization, could affect the interests of the EU or of one or more EU countries”. The rules for protecting EU Classified Information (EUCI) are set in the Council Decision 2013/48/EU [7] and refer to the means for information assurance in the field of communication and information systems that handle EUCI. For a system, information assurance is the confidence that it will protect the information it handles and will function as it needs to, when it needs to, under the control of legitimate users. Effective information assurance must ensure appropriate levels of confidentiality, integrity, availability, authenticity and non-repudiation. Confidentiality requires ensuring the security of personnel, any areas in which EUCI is stored or handled, including information systems, as well as appropriate measures to deter, detect and recover from deliberate or accidental compromise or loss of information. The confidentiality of EUCI shall be protected by approved cryptographic products corresponding to the level of classification. Integrity refers to the accuracy and completeness of information. Preventive, contingency and recovery plans for the protection of EUCI in emergency situations are mandatory under [7]. Availability is the property of being accessible and usable upon request by an authorised entity. Elimination of single points of failure, redundancy and automatic failover mechanisms increase availability of information. Information authenticity (i.e. is genuine and from bona fide sources) shall be ensured through system and communication level strategies. Finally, non-repudiation mechanisms (e.g. Identity management, authentication, signatures, and use of encryption keys) shall ensure that any occurrence of actions or events is associated with a particular trusted individual or system and thus, it is beyond reasonable doubt.

4. USE CASES

This section details two use cases exploring the integration of Cloud Computing solutions at SatCen. The first one

doi:10.2760/848593

illustrates the advantages of cloud in terms of computational scalability and elasticity required by intensive processing of High-Resolution Single Look Complex SAR imagery in an operational context. The second use case is a demonstrator pilot developed in the frame of the H2020 project NextGEOSS, where the whole data lifecycle is handled in the cloud, from discovery, to processing and publication of results.

4.1. PaaS for High Complexity Processing for GEOINT

A wide range of Security applications require the detection of changes occurring on the surface of the Earth within a predefined (typically short) interval, by analysing satellite observations. Depending on the geographical position, time of observation and the type of activity to be detected, imaging from space the area of interest can pose a number of difficulties, including presence of clouds, night time, or the presence of given activity outside of imaging intervals. For such cases, it is typically recommended the use of coherent SAR acquisitions, due to the independence on weather and external illumination and the possibility to use the phase variation information to detect target movement. One of the analysis methods most employed in SatCen - and deeply investigated in its Research, Technology Development and Innovation (RTDI) Unit - is addressing the coherence variation from successive coherent SAR acquisitions, also known as MTC (multi-temporal coherence) analysis.

While the information retrieved from the analysis is particularly useful to detect very small variations of assumed coherent objects, it comes with high and non-linear complexity costs, MTC analysis requiring the processing of the complex data with typically 16 bit per pixel for the in phase and in quadrature components. The most demanding stages of MTC product generation are sub-pixel image registration and computation of coherence, since they require the execution of convolution operations both in spatial and spectral domains. This imposes an overhead for processing such complex-nature images requiring an extensive processing infrastructure. To tackle the varying and exhaustive processing needs, a Cloud-based Workflow Orchestration Framework for SatCen Multi-temporal Change Detection Service based on Sentinel-1 SLC data has been put in place based on the ASB solution[1]. The system, depicted in Figure 2, is implemented as a cloud PaaS (Platform as a Service) offering scalability, automated generation of workflows, platform-agnostic orchestration, monitoring, control and algorithm integration. From a functional perspective, the framework permits to:

- Deploy, configure and execute custom algorithms as processing chains using workflows;
- Deploy processes in a distributed environment;
- Parallelize the execution of processes;
- Monitor the executions;
- Access execution reports and generated products;
- Access and manage everything through a web-based UI.

The platform is based on open source components like Apache Mesos, Marathon and Chronos to provide horizontal scalability and Apache Zookeeper for cluster coordination. This stack allows tasks to be submitted to the cluster,

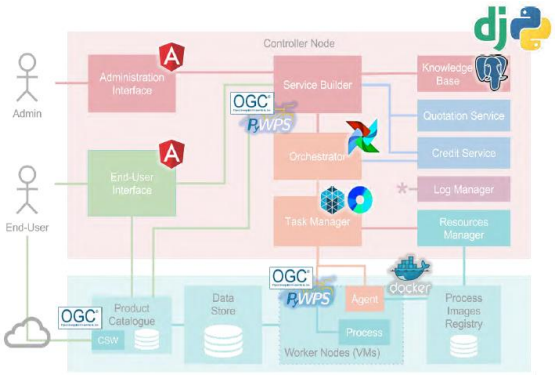


Figure 2. Cloud-based Workflow Orchestration Framework for SatCen Multi-temporal Change Detection Service[1]

orchestrated by the workflow engine, packaging user algorithms, deploying and executing them on user-selected platforms. Docker containers are used to provide the proper runtime for each task. Processing chains are designed as workflows in form of directed acyclic graphs (DAGs). Apache Airflow is used to describe those workflows and to orchestrate their execution. Data access and data processing is done through OGC standards (CSW and WPS) to enhance interoperability and reusability. Finally, quota and credit services, user management, logging and UI interfaces are integrated within the platform in order to provide a user friendly interface for users, platform operators and resources management.

4.2. Systematic Change Detection on the cloud – the NextGEOSS Space and Security Pilot

NextGEOSS [8] is the European Data hub and platform providing a sustainable approach for EO data distribution and exploitation, with the aim to increase access to Earth observations from Europe, supporting decision making. NextGEOSS includes a set of demonstrative research and business-oriented pilot activities, showcasing the system capabilities. The innovation pilot Space and Security, led and implemented by SatCen in the frame of its RTDI activities, is addressing the needs of stakeholders involved in the EO data exploitation for security purposes (e.g. EU decision makers and Member States). The pilot has a cross-domain focus, by implementing and integrating tools for change detection and characterization based on SAR and multispectral imagery (Sentinel-1 and 2). The fundamental principle of NextGEOSS

[1] https://www.spaceapplications.com/products/automated-service-builder-asb/

doi:10.2760/848593

is to optimize the connectivity of the European and global data centres with new discovery and processing methods. Applying this concept to the Security domain has the advantages of effective data management, leveraging Web and Cloud technologies, offering seamless access to all relevant data repositories (including open source and complementary information) as well as efficient operations (search, retrieval, processing, visualization and analysis), for example to extract and distribute single parameters, or to combine products on user demand from federated infrastructures. Furthermore, the concept revolves around providing the data and resources to the user community, so that the high level products generated from running the processing services are back-fed into the platform, becoming discoverable and accessible by the interested parties.

Figure 3 provides the Architectural view of the NextGEOSS services and their interoperability.

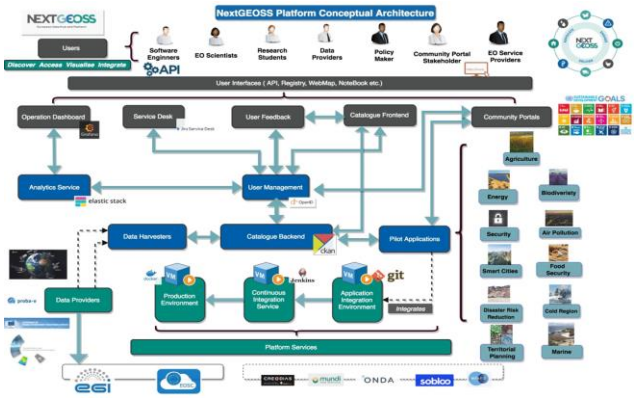


Figure 3. NextGEOSS Service Architecture - using Terradue Cloud Platform the processing service is integrated and deployed on production servers via the EGI.eu federated Cloud

The boxes denote Services and the Arrows indicate how the services are integrated. Data Harvesters aggregate metadata from several EO-data providers (e.g. Copernicus, Proba-V) into the Catalogue Backend. Applications are developed autonomously by pilot integrators using an Application Integration Environment. Changes to be integrated are simply committed and pushed by developers to the central repository and a Continuous Integration Service triggers the deployment to the Production Environment. User Management and Access Control is transversal to the whole Platform and uses OpenID Connect protocol. Another transversal component is the Analytics Service. Finally at the User Interface level, several subcomponents are built for operations and data exploitation: Operations Dashboard, Service Desk, User Feedback, Catalogue Frontend and Community Portals.

The SatCen pilot on Space and Security demonstrates the usage of the NextGEOSS Data Hub and Platform to discover and access Copernicus Data, exploit it for systematic Change Detection processing over very large Areas of Interest (Figure 4), and make available the high level and information rich products to the community by enabling the results to be harvested by the platform.

Figure 4. Space and Security Pilot AOI

The main benefit of using cloud processing is the execution of complex processing chains over large datasets in an automatic and systematic manner. Sentinel Application Platform (SNAP) is the main technology behind the implementation of the Change Detection Workflow [9].

5. CONCLUSIONS

This paper revised the main challenges and advantages of adopting cloud based solutions to address the requirements of the Space and Security domain. Two main use cases exemplify different implementation scenarios in the SatCen operational environment, pointing out the potential benefits of cloud computing in terms of infrastructure scalability, data management and information delivery for decision making.

6. REFERENCES

[1] European Commission, Big Data in Earth Observation, *Digital Transformation Monitor*, July 2017, (available online)

[2] Directorate-General for Communications Networks, Content and Technology, *Open innovation 2.0 yearbook 2016*, ISBN 978-92-79-53366-2

[3] Space Strategy For Europe, Com(2016) 705 Final, *Communication From The Commission To The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions*, 26 October 2016

[4] P. Soille, A. Burger, D. Rodriguez, V. Syrris, V. Vasilev, Towards A JRC Earth Observation Data And Processing Platform *Proc. of the 2016 conference on Big Data from Space (BiDS'16)*S.

[5] D. Müller; S. R. Holm; J. Søndergaard, Benefits of Cloud Computing: Literature Review in a Maturity Model Perspective, *Communications of the Association for Information Systems*: Vol. 37, Article 42, 2015. http://aisel.aisnet.org/cais/vol37/iss1/42

[6] Functional Requirements for the Copernicus Distribution Services and the Data and Information Access Services (DIAS), European Commission, Ref. Ares(2016)6929597 - 13/12/2016C.

[7] Council Decision of 23 September 2013 on the security rules for protecting EU classified information 2013/48/EU

[8] H2020 Project NextGEOSS https://nextgeoss.eu/

[9] S. Albani, M. Lazzarini, P. Nunes, E. Angiuli, "A Platform for Management and Exploitation of Big Geospatial Data in the Space and Security Domain", *Proc. of 2017 Big Data from Space Conference, (BiDS'17)*

CLOUD BURSTING EXPERIMENT AT CNES

Erwann Poupart[1], Denis Caromel[2], Paraita Wohler[2], Philippe Pham Minh[3]

CNES, 18 avenue Edouard Belin, 31400 Toulouse, France(1)

ActiveEon, 2000 route des Lucioles, 06560 Sophia Antipolis, France(2)

Microsoft France, 39 quai du Président Roosevelt

92130 Issy-les-Moulineaux, France(3)

ABSTRACT

In the context of PEPS platform (https://peps.cnes.fr), a cloud bursting experiment has started this year to evaluate some possible use cases taking benefits from cloud capacities. Sentinel-1 ortho-rectification processing chain, currently available on PEPS platform, successfully burst from CNES datacenter into Microsoft Azure public cloud.
Exactly the same processing workflow, using ActiveEon portable solution, could burst to 600 Azure compute cores. This was done in a "cloud agnostic way", ortho-rectification processing chain doesn't know where it runs. This paper will present the architecture of this experiment, its first results and current limitations observed. It shall be noted that this experiment is still ongoing to get some first evaluation of costs.

***Index Terms*—** Cloud, Microsoft Azure, Burst, Meta-scheduler, Docker, Image data processing, ActiveEon workflows

1. INTRODUCTION

Is it possible to burst Earth Observation data processing chain from CNES datacenter to a public cloud without major change on CNES datacenter infrastructure? What about performances in the cloud compare to the ones in CNES data center (HPC or dedicated nodes)? Does it scale? Is it possible to have a better idea of the costs for a precise use case? This cloud bursting experiment on PEPS processing platform tried to answer to those questions and identified some other ones.

2. PEPS PROCESSING ARCHITECTURE

Since 2018, the following processing chains are available on demand on PEPS platform:

- Sentinel-2 false or true color composition and radiometric indices (NDVI, LAI) computation.
- Sentinel-2 Atmospheric correction and cloud detection using MAJA processing chain (see http://www.cesbio.ups-tlse.fr/multitemp/?p=6203)
- Sentinel-1 ortho-rectification to produce ortho-rectified tiles at 10 meters of resolution using Sentinel-2 MGRS grid to be able to superpose Sentinel-1 and Sentinel-2 pixels (see http://www.cesbio.ups-tlse.fr/multitemp/?cat=38)
- It is planned to add some other processing chains like soil moisture computation using both Sentinel-1 and Sentinel-2 products, ortho-rectification with multi-temporal filtering of the speckle noise, etc.

The main components of PEPS processing architecture are (see figure 1):

- WPS (OGC Web Processing Service, see http://www.opengeospatial.org/standards/wps) to ease external access to PEPS processing infrastructure.
- ActiveEon workflows (cf. [1], [2] see http://activeeon.com/workflows-scheduling) to interconnect CNES HPC nodes : 300 Tflops, 8000 cores, PEPS dedicated nodes and external nodes (e.g. public cloud).
- Docker (http://docker.com) to ease process deployment on nodes with different OS and package distribution.
- HPSS (High Performance Storage System : http://www.hpss-collaboration.org/) storage to handle PetaBytes of data coming from Copernicus data hub relays.

Processing requests are sent via the Web Processing Service which routes them through Activeeon workflow scheduler to compute on resources allocated using its resource manager. Currently those computing resources are dedicated Linux servers and computing resources from CNES HPC cluster with 300Tflops, 380 batch servers and 8400cores, 6,2 PB

GPFS, 100GBs bandwidth and Infiniband low latency network.
HPSS stores all the Sentinel-1,2 and 3 images coming from Copernicus data hub relays since the beginning of the mission which started at the end of 2014. We currently have around 9 Peta-bytes of data and 10 millions of images.

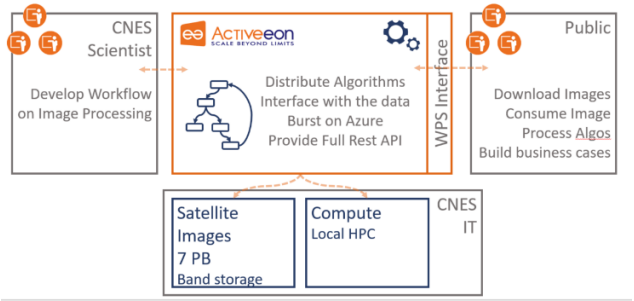


Figure 1. PEPS processing architecture

3. AGNOSTIC CLOUD DEPLOYMENT

3.1. Sentinel-1 ortho-rectification processing chain

Processing chain used for this experiment was Sentinel-1 ortho-rectification which takes GRD Sentinel-1 level 1 products as input and produces ortho-rectified tiles at 10 meters of resolution using Sentinel-2 MGRS grid to be able to superpose Sentinel-1 and Sentinel-2 pixels (see figure 2). In figure 2 Monteverdi tool from Orfeo Toolbox shows both Sentinel-2 and Sentinel1 ortho-rectified images using a chess display. We can observe that circles in the resulting image superposes well. It provides analysis ready data and the ability to combine both Sentinel-1 and Sentinel-2 to build other processing chains.
All this processing is done thanks to Orfeo ToolBox library (http://www.orfeo-toolbox.org) maintained and developed at CNES. To ease processing deployment in the cloud and in different on premise resources, this library has been dockerized.

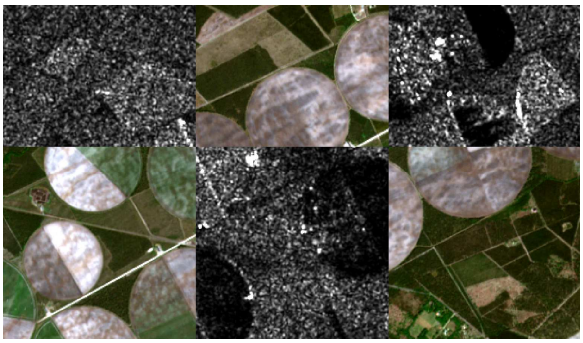


Figure 2. Sentinel-1 ortho-rectification result

There are two main steps in the workflow (see figure 3), the first step consists in collecting image and SRTM data necessary for the processing, the second step processes the data.
Result is provided in Cloud Optimized Geotiff (COG : http://www.cogeo.org/) format to improve data access efficiency.
One characteristic of this processing is that it is highly parallelizable using spatial dimension because each Sentine1 image that don't overlap can be processed in parallel.

Figure 3. Sentinel-1 image processing workflow

3.2. Processing deployment in the cloud

An important point is the cloud datacenter choice, Microsoft Azure provides many datacenters in the world. For this experiment, we choose the "FranceCentral" datacenter which opened in 2018 and offered free access at the beginning of this year.

The first step was to allocate cloud resources. This was done transparently thanks to ActiveEon resource manager. It provides several connectors to on premise resources (SSH connectors for on premise nodes, PBS connectors for HPC nodes, etc.) and several connectors to different cloud providers like AWS, Azure and so on. ActiveEon scheduler has a meta-scheduler capability.
ActiveEon resource manager provides a single generic access to compute nodes. An important consequence is that processing workflow is "node agnostic". The same workflow task runs either on "on premise" nodes or on public cloud nodes. Cloud bursting is then easier to use.
Cloud knowledge is important to optimize its use and costs, but its configuration is completely separate from workflow design.

We tried different type of image size like STANDARD_D2_V3 from D series or STANDARD_E32S_V3 and STANDARD_E64-32S_V3 from E series provided by Azure.
Finally, STANDARD_E64-32S_V3 image size with 32 vCPU and SSD disk at 8 GB/sec was the model that provided the closest processing response time (between 20 minutes and

one hour for one Sentinel-1 image depending on its size) compare to CNES on premise CPU's (HPC cluster with 24 Intel(R) Xeon(R) CPU @ 2.20GHz and GPFS at 9GB/s or dedicated nodes with 32 Intel(R) Xeon(R) CPU @ 3.30GHz and ATA disk at 0.4GB/sec.

This model costs approximately 3.9€ per hour and it is the main element of cost in the process. Consequently, processing one Sentinel-1 image would cost for this experiment approximately between 0.5 and 1.5 €. It shall be noted that this is only an evaluation based on a first deployment that is not optimized. Any optimization that enables to use a cheaper azure virtual machine model with equivalent response time will reduce the costs.
To give first elements of comparison, the costs on CNES cluster are of 0,05€ for one-hour CPU including computation, local storage and support. To be able to compare with Azure, we must multiply this cost by 32, which gives 1.6€ to compare with 3.9€ for the Azure STANDARD_E64-32S_V3 virtual machine.
However, Azure VM configuration can be optimized and costs on CNES cluster makes the assumption that cores are fully occupied. For highly variable processing requests the costs of using physical servers can increase significantly. We started this year a second experiment to evaluate more precisely those costs.

We used Activeeon Azure connector to allocate up to 10 VM's with 64 cores each (see figure 4). One Sentine1 image processing requires 6 cores, so tens of images have been processed in parallel using those resources. It shall be noted that 640 cores are not a limit for Azure or Activeeon scheduler. Some tests done by Activeeon succeeded to allocate 20 000 Azure VM cores.

Cloud transparency is not enough to facilitate application process deployment; the use of Docker images is also a key point to deploy application process packages on any Linux operating system including Docker in its distribution.
We only had to load the processing Docker images into a virtual machine boot script so that everything needed to run the application is available as soon as the virtual machine is available. A second boot script was necessary to start the Activeeon node service to make azure nodes communicate with the ActiveEon resource manager installed in the CNES datacenter.

For the network connection required for processing interactions between CNES datacenter and Azure, we only needed two SSH access. one SSH access in a two-ways mode for the interactions with Activeeon resource manager, and one in a SSH access to a diagnostic VM to monitor deployment and processing in the Azure scaleset.
This was sufficient to use public cloud in a hybrid way.
It shall be noted that it is recommended to have an additional direct link with the cloud datacenter for operational use cases. This direct link provides a better quality of service and may reduce data transfer costs.

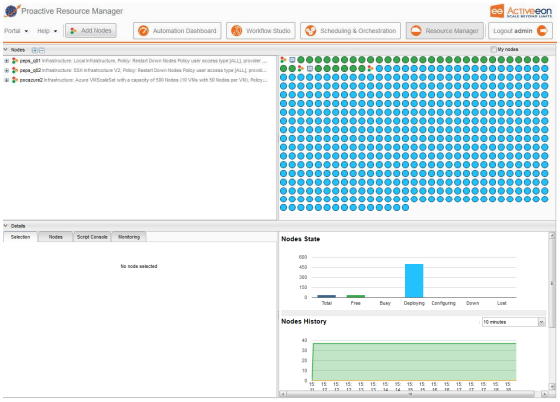


Figure 4. processing deployment in the cloud

3.3. Data deployment in the cloud

Transparent access to cloud and on premise data facilitates cloud bursting. It provides more architectural choices for data location. This data access transparency was not the main focus of this first experiment, we simplified this problem using logical paths.

To fetch Sentinel-1 image data required for the processing, we used the same http process that we use internally. One difference was in the time necessary to fetch image data; due to external bandwidth, it took a few minutes in the cloud compared to an internal fetch which took less than 20 seconds to download a few Gigabytes of data.

CNES datacenter external bandwidth through firewall is limited to 2 Gbit/sec. It is not sufficient to download efficiently hundreds or thousands Sentinel-1 images occupying each several gigabytes.
This really is a key point, for high data volume cloud bursting, a high bandwidth capacity is required between cloud provider and on premise datacenter. A dedicated link between CNES datacenter and public is necessary.
For the auxiliary data (around 500 Go of SRTM data), we transferred it on an Azure file share resource that was accessible from the computing VM's.

A shared working directory was needed to share Sentinel-1 repository data between VM's. We used Azure file share resource mounted using cifs.

To reach good performances for the data processing we had to copy image data to be processed from Azure file share resource to the VM SSD local disk. The reason for this is that Azure file share is an object storage resource not designed for fast I/O processing, consequently execution time is an order

of magnitude higher. Another option could be to have a file share resource with better I/O capability (eg GPFS). We didn't try this option which is more expensive, it can be evaluated if it's significantly improves execution time.

4. CONCLUSION

Sentinel-1 ortho-rectification processing chain, currently available on PEPS platform, successfully burst from CNES datacenter into Microsoft Azure public cloud. Although there are still many challenges to be addressed, such as network bandwidth, data access transparency, cost optimization, data privacy and confidentiality, and agile development processes. This paves the way for new opportunities for a data processing architecture that makes the best of planned and optimized on premise resources and flexible on-demand cloud resources.

A second experiment was launched this year to more accurately assess costs, test elastic cloud bursting (see figure 5), improve data transparency and optimize cloud configuration.

Figure 5. elastic cloud bursting

5. ACKNOWLEDGEMENTS

Many thanks to Microsoft Azure and ActiveEon for their support to this first experiment.

6. REFERENCES

[1] Asynchronous and Deterministic Objects:
Denis Caromel, Ludovic Henrio, Bernard Serpette,
POPL'04, Proceedings of the 31st ACM Symposium on Principles of Programming Languages",
2004, 123—134 https://dblp.org/rec/html/conf/popl/CaromelHS04

[2] Interactive and Descriptor-based Deployment of Object-Oriented Grid Applications
F. Baude, D. Caromel, F. Huet, L. Mestre and J. Vayssiere
pp. 93-102, in HPDC-11, Edinburgh, Scotland, July 2002. https://dblp.org/rec/conf/hpdc/BaudeCHMV02

USE CASES FOR THE ESAC SCIENCE EXPLOITATION AND PRESERVATION PLATFORM

Christophe Arviset[1], Vicente Navarro[1], Rubén Alvarez[1], Bruno Altieri[1], Deborah Baines[2], Carlos Gabriel[1], Rocio Guerra[1], Aitor Ibarra[2], Marcos López- Caniego[3], Anthony Marston[1], Bruno Merin[1], Fernando Perez[4]

[1]ESA-ESAC, Madrid, Spain

[2]QUASAR for ESA-ESAC, Madrid, Spain

[3]AURORA for ESA-ESAC, Madrid, Spain

[4]RHEA for ESA-ESAC, Madrid, Spain

ABSTRACT

ESA space science missions are continuously producing data which is being hosted at the ESAC Science Data Centre and made available via archive services to the astronomical, planetary and heliophysics science community. The GNSS Navigation Science Office has been conceived with the mission to foster the consolidation of a world-wide reference centre for the GNSS Scientific Community, maximizing possibilities to perform GNSS Science activities and utilization of European GNSS Infrastructures (Galileo & EGNOS). The increased volume and complexity of such science data now calls for a paradigm shift in data analysis approach, enabling science users to bring their code to the data, rather than bringing the data to the users. ESAC Science Exploitation and Preservation Platform (SEPP) project plans to develop a multi-missions and multi-disciplines data analysis platform in an open and collaborative environment. This paper describes the main science use cases to be addressed by this platform.

Index Terms— Exploitation Platform, Archives, Space Science, Galileo, Code to the Data

1. INTRODUCTION

Through an internal workshop, science use cases have been collected from the missions at ESAC to ensure the Science Exploitation and Preservation Platform (SEPP) implements their new needs about data analysis services.

The main initial features of the platform are planned to be:

- Interactive and collaborative "in-situ" data analysis through Jupyter Lab,
- Execution of user's custom pipeline on archive science data,
- Web based instantiation and access to legacy systems (i.e. data processing, mission planning, etc, …),
- Storage space in the platform for user's to bring their data and processing code close to the archives,
- Crowdsourcing pipelines allowing for processing of massive, highly distributed datafeeds (i.e. IoT datafeeds),
- Publication of discovery of science products and processors through a "Science App Store".

2. SEPP AS AN ARCHIVE DATA EXPLOITATION AND COLLABORATIVE RESEARCH PLATFORM

2.1. ESDC Collaborative Research Lab

The ESAC Science Data Centre (ESDC [1]) is responsible for developing, maintaining and operating all ESA Space Science astronomy, planetary and heliophysics science archives. The ESDC Collaborative Research Lab will take advantage of the SEPP infrastructure for testing and dissemination of software libraries and example workflows that demonstrate the potential of real time querying, visualization and analysis of data from ESA's science archives.

- For astronomy missions, this is done producing example Jupyter notebooks that combine powerful queries to the astronomy science archives using ESDC contributed modules to astropy (for example, the GAIA TAP+; the ESASky [2]; and the general TAP/TAP+ astroquery modules; and the prototype Hubble Space Telescope TAP+AIO (Archive Inter-Operability subsystem) astroquery module) with advanced visualization tools using the Jupyter widget for ESASky (pyESASky) and other community provided libraries.
- For planetary and heliophysics missions, example Jupyter notebooks are produced querying the Planetary Science Archive (PSA) using TAP[3], and the Heliophysics archives with TAP+AIO, and exploring

doi:10.2760/848593

and visualizing the results using external libraries like SunPy and PyPDS.

Moreover, the ESDC Collaborative Research Lab will also take advantage of the SEPP infrastructure to test innovative collaborative research tools that allow real-time data analysis and visualization of astronomy and planetary science archives data products using scalable computing resources.

2.2. Planck Legacy Archive Collaborative Research Area

The Planck mission developed a precursor project called PLAAVI (Planck Legacy Archive Added Value Interface), enabling the execution of pre-defined data analysis pipelines directly on the science products stored into the Planck Legacy Archive (PLA). SEPP will expand these facilities by providing a computing and storage environment connected to the PLA for scientists to be able to run on-the fly tools and pipelines to explore and analyze more easily Planck data through:

- Collaborative tools: JupyterLab, Apache Zeppelin, others.
- High-level programming languages and interpreters: Fortran, C, C++, Java, Python, R, Octave and GDL (or Matlab and IDL, license permitting).
- Access to the Planck repository or to a subset of it in secure environment.
- Libraries needed to handle Planck products, in particular fits tables and fits maps in Healpix format (cfitsio, Healpix).
- Private storage areas to host the scripts, notebooks and results from the users, local or remote (VOSpace [4], Dropbox, Google Drive).
- Public storage areas to host scripts, notebooks and results that can be shared with a restricted group of people or the general public.
- FAQ and Examples Library of read-only notebooks that the users can use to learn how to explore the Planck repository and do simple processing of Planck data.
- Embedded TOPCAT, Aladin, ESASky, GLUE, and other visualization tools available
- Users Forum. Support to the community for questions, doubts.

2.3. GSSC Collaborative Research Area

The GNSS Science Support Centre (GSSC) aims at consolidating a world-wide reference GNSS environment for scientific communities. GSSC promotes collaboration and innovative research integrating GNNS Data, Products, Information Services and Resources from multiple sources in a single repository.

Primary actors for the GSSC Collaborative Research Area will be the GSSC collaborating organizations and institutions with knowledge of GSSC data, products and information services with access to the full GSSC repository and have privileges to run complex pipelines in the system on GSSC's computing and storage resources. Secondary actors will be the general public, with or without prior knowledge of the GSSC products and their complexity, willing to explore and analyze GSSC data, products and information services on a limited range of products with limited computing and storage resources. These users could become collaborators.

The GSSC Collaborative Research Area will provide the users with tools to explore and analyze GSSC repository. These tools will include:

- Collaboration based on JupyterLab. Examples of Jupyter Notebooks:
- High-level programming languages and interpreters: python, C++
- Libraries needed to handle GNSS data and products (RINEX ...) with various programming languages (python, Fortran, C, C++ and MATLAB).
- Private storage areas to host the scripts, notebooks and results from the users.
- Public storage areas to host scripts, notebooks and results that the users feel should be shared among a restricted group of people or the general public.

2.4. Euclid Collaborative Research Area

Euclid, to be launched in 2021, will generate PB of data residing in the Euclid Archive to be located at ESAC. Analysis of Euclid data will definitely require brining the code to the data as the data will be too big to be transferred to the end user's location. Members of the Euclid consortium and later any scientist interested in processing Euclid data will have to do this in-situ, close to the archive. The SEPP will allow to explore the Euclid repository and do simple processing of Euclid data, modifying and saving the resulting processing elements in public, shared or private user areas.

The Euclid Collaborative Research Area will provide the users with tools to explore and analyze Euclid repository, such as:

- Collaboration based on Jupyter python notebooks close to the Euclid archive for interactive access and light processing related to image cut-out service, user-defined source extraction in a given area, user-defined spectra extraction, re-running the SIR pipeline on a given set of sources and apertures.
- Execution of some systematic pipeline (post-) processing on level 1 or level 2 (most probably) Euclid maps for solar system object detection, strong lenses detection (would require accessing machines with GPUs)

The Euclid Science Operations Centre is significantly involved in the execution of pipelines. However at this stage it is not entirely clear if these pipelines should be run and integrated in the SGS consortium or could be run ESAC as soon as data are available.

2.5. JWST Workspaces

The main goal of JWST for workspaces is to allow specialized data reduction on large amounts of data that can be processed via a dedicated JWST SEPP Application Packages (SAP) running on the SEPP at ESAC, on a set of servers or cloud services. This allows:

- Faster archive data access – useful for projects with large amounts of data.
- Collaboratively, groups do not need to download multiple copies of the data, but can bring elements into their workspace and share it among each other's.
- The workspace always has the environment setup with the latest pipeline software, JWST data analysis software and reference data releases – automatically available to users on setup of a JWST SEPP Application Package (SAP).
- Selectable set(s) of reference data or user can incorporate her/his own. This is done with the association table, nothing to "select"
- Allow incorporation or replacement of pipeline elements – specialized pipeline
- Selectable saving of results (all, only some)
- Possible tools for creating user/group database(s) of results.

2.6. Science App Store

Once being used by various users and communities, the SEPP objective is to allow these users to publish their pipelines, scripts, tools and software through a "Science App Store". This would represent a two ways benefit: from the application developer to publish their work and get recognition for it and from the end users to have access to a wider range of applications made easily accessible through the platform.

3. SEPP AS A DATA PIPELINE PLATFORM

3.1. BepiColombo Instrument Pipelines Scheduling and Execution

Members of the BepiColombo Science Ground Segment (SGS) operations team need to be able to generate, schedule the execution and run the BepiColombo Instrument and Auxiliary pipelines. Developed by the Instrument Teams, these pipelines could be run either at their premises on directly on the SEPP at ESAC, offering a common and unified platform for all SGS partners [5].

The BepiColombo instrument pipelines involves the integration of pipelines coding in different languages and operating systems [6]:

- Coding languages supported: Matlab, IDL, C++, Java and python
- Operating systems supported: Linux OS, MS Windows

The pipeline source code is developed by the SGS and/or Instrument teams and maintained under configuration control (GIT) by the SGS [3]. SEPP shall provide support for the creation, scheduling and execution of all pipelines and also to the execution of the pipelines by the users in their specific SEPP user areas.

4. SEPP AS A LEGACY SOFTWARE PRESERVATION PLATFORM

4.1. On the fly Instantiation of Legacy Systems

When missions enter in Legacy phase, their raw data and their scientific products are preserved into their science archive. However, there can also be the need to preserve the associated data analysis software so users can re-analyze these data in the future. Similarly, preserving the capability to run again mission planning tools of old missions might also be required when preparing future similar missions.

Maintaining these in the long term represents a real challenge that can be solved more easily by virtualizing and "dockerizing" such legacy software to make them available through one-click instantiation within the platform.

The following legacy systems have already been initially identified for this preservation use case:

- EXOSAT interactive analysis software (from the mid-80s)
- ISO PHOT instrument interactive analysis (from the mid 2000)
- Herschel Interactive Processing Environment (from 2009)
- Rosetta Mission Planning system (from 2016)
- Lisa Path Finder Data Analysis Scope (from 2017)

4.2. XMM-Newton Legacy Science Processing Capability Area

XMM-Newton ESA spacecraft mission produces scientific raw data that has to be analyzed in order to produce high quality science data. To do this, XMM-Newton provides users with a software package called Science Analysis System (SAS) [7], to process raw data up to high-level scientific products. XMM-Newton Science Operations Centre also provides ready-to-use scientific products produced by an analysis pipeline (based on SAS tasks) and accessible through the XMM-Newton Science Archive (XSA). In the second scenario, both the raw data and high-level products produced by XMM-Newton are stored in ESA Science Mission Archives where the users can access them at any time. This way, the long-term preservation of the data, both in raw and high-level format, is ensured. The first scenario becomes more cumbersome, as preserving software processing capabilities beyond the operational phase of the mission, implies that a user has to have access to the data processing software, in such a condition that the user can run it regardless of their operating system and third-party software, like compilers and libraries.

During the mission and post-mission operations phase, the SAS software is under constant development, is regularly maintained or improved and new technologies are investigated and implemented. However, once the post-operations phase is finished, the data processing software is no longer maintained, which means that in the long term, it will most likely become obsolete. We understand by long term that, technology, programming languages and operating system changes, and even an evolving user community, should be a concern for software preservation.

There are however, several scenarios which will benefit from having access to a working version of the SAS XMM-Newton data processing software. There are now ways to ensure the delivery of a working version of the SAS package to the science community that would allow software to be of use well beyond the post-operations phase of a mission.

A scientific user might want to reanalyze the data long after the spacecraft mission is well past its post-operation phase under the following scenarios:

- Time series taken from a set of energy ranges and time intervals
- Spectra taken from a set of time ranges
- Spectra taken between a set of count rates (fluxes)
- Spectra taken at certain phases of a periodic signal.
- Images taken over a certain spectral and/or time range
- Spectra taken from a given extraction region (circle, box, annulus, user-defined)
- Background spectra or light curves taken from a given spatial region
- Combination of two or more of the above

As part of the process of keeping XMM-Newton SAS processing software available in the long term, XMM-Newton team started the development of the Remote Interface to Science Analysis (RISA [8]) software to explore the possibility of offering SAS functionalities encapsulated within web technologies and grid/cloud infrastructures. RISA is a Java web Client/Server application, which makes use of grid technologies to run all SAS tasks offering all SAS functionalities over the network, without having to install SAS, any associated third-party libraries, or grid credentials.

Since XSAv9.4, the XMM-Newton Science Archive offers the possibility to process on-the-fly XMM-Newton through RISA Restful services, fulfilling some of the scientific cases mentioned above.

5. CONCLUSIONS

The increased data volume for space science missions call for new methods to provide access and analysis tools to fully exploit science data. This paradigm shift can be described by "bring the users to the data", replacing the previous model "bring the data to the users". The implementation for this new approach can be done through science exploitation platforms, which will also offer collaborative research environment for the scientific community. With the maturity of IT technology like virtual machines, docker containers and more recently Jupyter Notebooks, these platform can serve not only upcoming space science missions but also represent opportunities to preserve missions' legacy software, and make them more easily available to the end users. All these various and complementary use cases define the baseline set of requirements for the multi-disciplinary Science Exploitation and Preservation Platform to be built at ESAC in support to the Operations for science and GNSS missions.

6. REFERENCES

[1] C. Arviset et al, "From ISO to Gaia: a 20-years journey through data archives management," *ADASS XXVII proceedings*, ASP Conf. Series, 2016, in press.

[2] F. Giordano et al, "ESASky: A science-driven discovery portal for space-based astronomy missions", *Astronomy and Computing*, Volume 24, p. 97-103, 2018.

[3] P. Dowler et al, "Table Access Protocol", *IVOA Standards*, http://www.ivoa.net/documents/TAP/, 2018.

[4] M. Graham, B. Major et at, "VOSpace Service specifications", *IVOA Standards*, http://www.ivoa.net/documents/VOSpace/, 2018.

[5] F. Pérez-López et al, "Framework for the integration of multi-instrument pipelines in the BepiColombo Science Operations Control System", ADASS XXIV proceedings, ASP Conf. Series, vol. 491, page 273, 2015.

[6] J.C Vallejo et al, "Flexible and Modular Design for the BepiColombo Science Operations Control System", *ADASS XXIV proceedings*, ASP Conf. Series, vol. 491, page 277, 2015.

[7] C. Gabriel et al, "XMM-Newton Science Analysis System (SAS): medium and long term strategy", *The X-ray Universe 2017 Proceedings*, page 84, 2017.

[8] A. Ibarra et al, "On-the-fly Data Reprocessing and Analysis Capabilities from the XMM-Newton Archive", *The X-ray Universe 2017 Proceedings*, page 281, 2017.

JASMIN: MANAGING VARIETY IN A CLIMATE DATA COMMUNITY PLATFORM

Victoria L. Bennett[1,2,3], P.J. Kershaw[1,2], R.D. Smith[1], B.N. Lawrence[3,4]

Science and Technology Facilities Council, UK [1], National Centre for Earth Observation, UK [2], National Centre for Atmospheric Science, UK [3], University of Reading, UK [4]

ABSTRACT

The JASMIN data analysis infrastructure has been in operation for 7 years now, supporting the UK climate and environmental science community with a diverse range of datasets, compute capabilities and storage. A number of innovative solutions have been developed and applied to cope with not only large data volumes, but particularly the variety in types and sources of data, the range of user skill levels and the heterogeneity in the hardware deployed in JASMIN's upgrades and expansions.

JASMIN provides a high performance and flexible compute environment alongside 10 Petabytes of curated data (with over 5,500 different datasets in the Earth Observation and Climate domain). This paper describes the data and infrastructure management approaches that have been implemented to support nearly 2,000 users of this system.

Index Terms— EO data archive, analysis platform, climate data, data exploitation platform, data storage

1. INTRODUCTION

The JASMIN infrastructure at the UK Centre for Environmental Data Analysis (CEDA), has been developed primarily to support the UK environmental research community with high-performance analysis and compute capability alongside a big data storage environment [1]. The facility has nearly 2,000 active users, from the UK but also from elsewhere as part of international collaborative projects. The latest hardware expansion has increased the available (disk) storage to 44 Petabytes, of which 6 Petabytes holds the curated EO and climate on-line data archive and the remainder is primarily for user and project workspaces. The compute has been increased to 12,000 cores, deployed in different configurations: hosted processing is offered via the HPC-like cluster, Lotus, and a community cloud enables users to provision their own bespoke application environments.

As the system has evolved over the last 7 years, and available technologies and costs have changed, different choices have been made over how best to store and manage the data.

The curated data archive on JASMIN holds many different datasets, including Earth Observation (EO) data, climate model simulations, meteorological data, reanalyses, in situ observations from ground and airborne platforms and other experimental data. Given the complexity in navigating this heterogeneous resource, it is essential to provide a range of search and browse options for our users, supported by comprehensive metadata.

JASMIN users span a range of technical capability, from highly IT-literate computational scientists, to students taking their first steps in data processing and analysis. To support this diversity of requirements we offer a number of different ways to take advantage of the JASMIN processing capacity and data resources.

This paper describes JASMIN's approaches to managing highly diverse data, storage and compute offerings.

JASMIN is hosted at the Science and Technology Facilities Council (STFC), Harwell, UK, and funded by the Natural Environment Research Council (NERC), both are part of UK Research and Innovation (UKRI).

2. JASMIN CONCEPT

JASMIN is modelled on the paradigm of "bring the compute to the data" as a means to overcome the challenges of Big Data. Data volumes in the Earth Observation and climate domain are so great that in many cases it becomes impracticable to move data to a user's computer or to computing resources at their home institution. Instead, we keep the data in one place and provide access to computing resources to analyse in situ.

The CEDA data archive, hosted on JASMIN, holds over 10 Petabytes of Earth Observation, atmospheric and climate data, including the the ESA Climate Change Initiative's data archive of Essential Climate Variables (ECVs).

This large and heterogeneous collection of data (over 5,000 datasets, 180 million files as of October 2018) is curated - organised and catalogued in such a way as to make it easy for users to discover it and access it. This overcomes a major overhead for many researchers of doing the necessary conditioning to get the data into a structure in which it can processed and analysed to achieve the given research goal.

Currently our biggest datasets come from the Sentinel missions, where we hold global data for most sensors on the active Sentinels 1A, 1B, 2A, 2B, 3A, 3B and 5P (see Figure 1)

doi:10.2760/848593

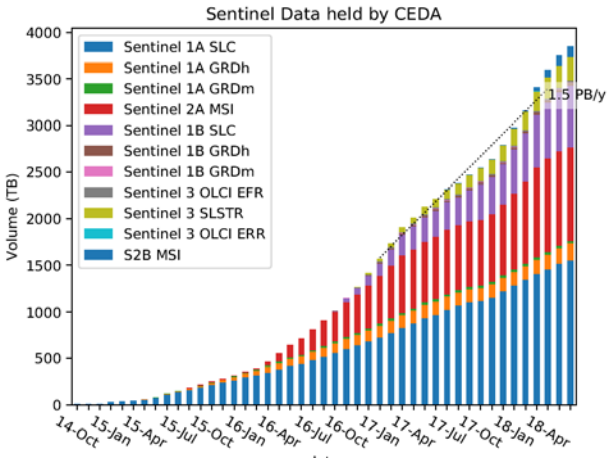


Figure 1: Sentinel data growth 2014-2018 in CEDA archive on JASMIN

As well as the curated data archive, users, groups and projects can be allocated portions of storage to store their own data – 'Group Workspaces'. Group workspace data takes up considerably more space on JASMIN than the data archive.

A range of technologies have been at our disposal during the evolution of JASMIN to implement the "bring the compute to the data paradigm". One of the challenges has been how to combine these capabilities in the best way so that users can take advantage of the strengths of each.

3. COMPUTATION, ANALYSIS AND SHARING OF DATA

In order to support a wide range of user workflows, a number of different interfaces and applications have been developed for access to JASMIN's data, storage and compute:

- Lotus batch computing: a cluster of physical hosts together with software to manage parallel processing tasks and manage fair share of the cluster between multiple users
- Virtualisation: a system for provisioning virtual machines
- Data Transfer Zone: a dedicated area of the network outside the site firewall to enable high bandwidth data transfer of data in and out of JASMIN
- Community Cloud: provides the ability for users to provision their own virtual machines and other computing resources

These are summarised graphically in figure 2. The "External cloud" (Infrastructure-as-a-Service, IaaS) sits outside the site firewall, allowing more flexible usage by its tenants, but with reduced data access performance. The "Managed cloud" has direct POSIX access to the full data archive and group workspaces, but is operated as Platform as a Service (PaaS) with no root privileges for users and virtual machines based on fixed a fixed catalogue of templates.

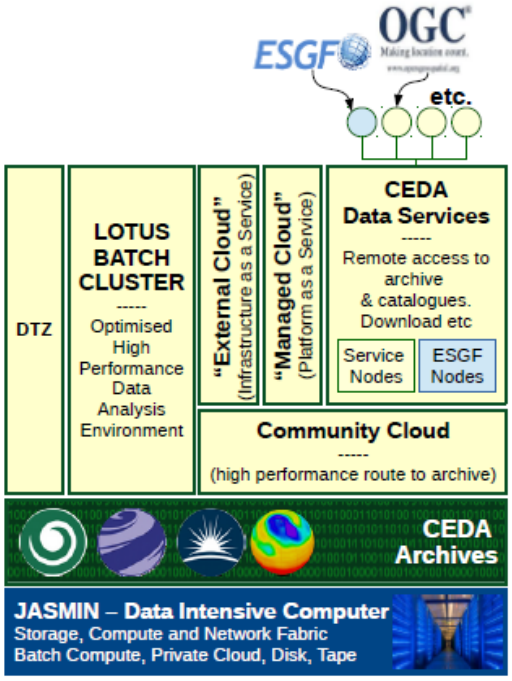


Figure 2: Schematic of JASMIN architecture

Much of the JASMIN success has been driven by the batch computing cluster, Lotus, providing a traditional means to provide computing. In the EO domain, Lotus is extensively used for global climate data processing, applying scientific algorithms to multi-mission satellite data, combined with relevant auxiliary data, to derive ECVs – climate quality products for project such as ESA Climate Change initiative and Copernicus Climate Change Service.

The batch-computing offer has been enhanced with support for Singularity container technology, and the provision of a set of standard libraries and packages: the JASMIN Analysis Platform (JAP), tailored to the needs of EO, climate and atmospheric science community.

The JASMIN cloud system provides additional dynamism enabling users to provision their own OPeNDAP service to publish their group workspace data externally and provision their own VMs based on the Scientific analysis VM (Virtual Machine) template. This is configured with the JAP packages. Access to the cloud management interface is provided through a simple custom web portal. Direct access to the OpenStack API is granted to advanced users.

These capabilities can be interpreted in the context of the "Pets and Cattle" analogy[1], with Lotus providing the 'cattle'

[1] https://cloudscaling.com/blog/cloud-computing/the-history-of-pets-vs-cattle/

compute. To date the cloud has for the most part been employed for the 'pets' use cases with various user groups employing it to deploy long run hosts which provide important services. In our new phase of expansion for JASMIN, the use of the cloud will be expanded to a more flexible and dynamic approach providing an additional class of cattle compute. This will foster an automated approach to provisioning and management of infrastructure.

Importantly, expanding the use of the cloud in a cattle scenario could allow more effective exploitation of parallelism. Running processes in parallel will deliver more effective use of the available computing resources and is essential to demonstrate increased performance over the laptop or desktop experience. However, parallel programming is not easy for users, particularly those who are not familiar with more advanced programming techniques. Tools like Spark, Dask and Slurm can help abstracting to some extent the parallel programming from the user making it easier for users to take advantage of parallelism.

Virtual Research Environments (VREs) have an important additional role to play in making these tools more accessible. The NERC DataLab [2] has been an important prototype project hosted on JASMIN over the last twelve months. The project has developed a complete system for managing analysis interfaces with Jupyter and Zeppelin Notebooks and R Studio, integrated with Spark and Dask, together with the management of virtual storage space. The system is underpinned with the Kubernetes container technology to make it resilient, easy to scale and portable.

Experience has shown that for projects like the NERC DataLab to flourish we need to help developers by providing recipes to create services so that they can deploy them in their cloud-hosted apps. We have developed a Cluster-as-as-Service concept to which resources have been allocated in a project in the coming months. This will build prefabricated components for developers to easily deploy in the cloud including Slurm and Kubernetes clusters.

However, more extensive use of cloud will also be dependent on effective integration between cloud and the traditional storage interfaces (the parallel file system).

4. STORING THE DATA: EVOLUTION OF STORAGE HARDWARE

A core architectural challenge is the interface between storage and computing and analysis capabilities. JASMIN was first launched in 2012, and has had four significant upgrade phases since then, the most recent in 2018 (Phase 4). Different storage hardware has been procured and deployed in JASMIN's expansion phases, partly due to the rapidly growing data volumes, making it economically unfeasible to keep up with on-line disk storage.

Historically, JASMIN has made extensive use of the Panasas parallel file system. Together with the high-performance network architecture, it gives high-performance I/O (input/output), critical for a data-intensive system. Object storage provides an alternative way to store and access data in which rather than a hierarchical file system, files – or rather objects – are stored in a flat structure of key-value pairs. JASMIN has not previously deployed any object storage but this kind of storage has been included in the Phase 4 upgrade. Although cost efficiency was a driver to the selection of object storage, the primary reason was to address some of the limitations of traditional file systems (in terms of flexibility in our cloud environment, performance, and scalability).

Tape is increasingly important as the sheer volumes of data outstrip the disk capacity that it is practicable to purchase for JASMIN. Figure 3 shows the volume of data moved from disk to tape, in total (black line) and per month (red columns). The process is currently dominated by Sentinel data: most of the Sentinel data archive (typically all data more than 3 months old) is routinely moved onto the tape Near Line Archive (NLA).

To date approximately 4 PB of archive data is in the near line tape archive, compared to 6 PB always on disk, with the proportion on tape growing rapidly.

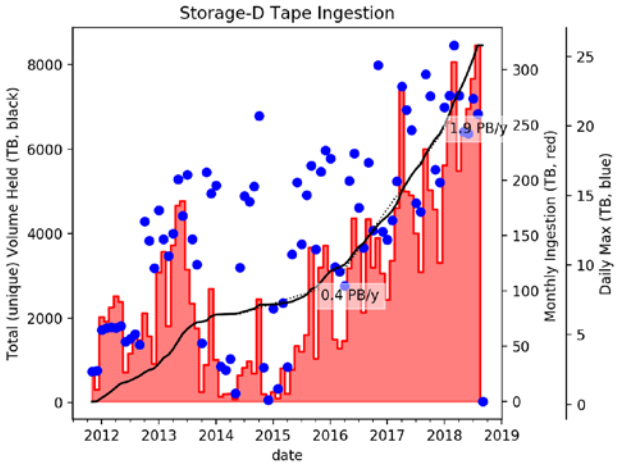


Figure 3: Data ingestion from CEDA archive to tape, 2012-2019 : Dominated by Sentinel data

Suitable interfaces are needed to enable the user to pull data back from tape, without needing to know the details of where and how it's stored. A dedicated tool, the NLA client[2], has been developed to allow users to find and restore files to disk for processing and analysis.

[2] https://github.com/cedadev/nla_client

5. FINDING THE DATA: INDEXING AND CATALOGUING

Aside from services for processing and analytics and data access, data discovery is fundamental to JASMIN's remit or more especially the CEDA curated archive.

We not only index the geospatial and temporal coverage, but also capture contextual information about the source, purpose and usage of the data. This necessitates a metadata system that both populates a filesystem index "bottom-up" by scanning all the files and extracting key features, but also a human-based approach where catalogue records are crafted per dataset with all the necessary descriptive information to aid our users.

A project underway to address the "bottom-up" approach is File-based Search (FBS) [3]. It provides for the first time an index of metadata at the granularity of individual files. This involved writing parsers for all the major file formats in the CEDA archive and running large indexing processing jobs on the Lotus batch compute system. These processes extract key metadata from all the data files and add them to an ElasticSearch database. 180 million files have been indexed, using 10 different parsers. The number of parsers gives an indication how many different file formats are in the archive: in fact we have chosen the top 9 major formats to index fully; this accounts for 42% of the CEDA archive. The 10th parser has indexed the remaining files, of miscellaneous formats, in a minimal way, with only very basic information added to the index (e.g. filename and file size, housekeeping information).

The File-based index is being used to develop new services and improve existing ones. For example, the CEDA satellite data finder[3] allows geo-temporal search for satellite data products in the archive including the Sentinel missions and Landsat.

The CEDA Data Catalogue (MOLES) [4] has a different but complementary role to FBS. It provides a catalogue of datasets (collections of file corresponding to a given project, campaign, instrument or experiment). The data in this catalogue is compiled by human input. There are over 5,000 datasets now in the catalogue. The catalogue has been augmented to include information about dataset variables sourced from FBS. This allows users to search for a specific geophysical parameter and match to a given datasets or set of datasets.

The relationship between FBS and MOLES records is discussed in [4] and [5] – being "Archive" and "Browse/Discovery" metadata in their notation.

FBS is key for the future as we move to using new storage interfaces such as S3 with object storage. This is because with object storage we will no longer have a hierarchical file system and data directories. The data directories contain metadata themselves by virtue of their names, for example, instrument name, processing level, acquisition date. This information has been indexed into the FBS database so that even though data directories may no longer exist, users can discover the data they need by querying for these terms using FBS.

6. CONCLUSIONS AND FORWARD LOOK

JASMIN is continuing to grow, both in size and capabilities. The most recent expansion (Phase 4) has seen a significant expansion to the storage and compute, and also to the cloud service.

The increasing complexity of the system, and the heterogeneity of the data and workflows requires constant innovation and development of services to support users: addressing diverse computing environments, increasing storage requirements and data management challenges.

Looking ahead, a number of specific activities are underway to support greater exploitation of the cloud, in particular we will be developing a suite of preconfigured packages for analysis (Cluster-as-a-Service) and providing new interfaces to access data from the cloud. In particular, high performance OPeNDAP access and Object Store with S3 interface, and NetCDF/HDF data over S3[4] are under development.

REFERENCES

[1] Lawrence, B. N., Bennett, V. L., Churchill, J., Juckes, M., Kershaw, P., Pascoe, S., Stephens, A. (2013). Storing and Manipulating Environmental Big Data with JASMIN. In 2013 IEEE International Conference on Big Data (pp. 68–75). https://doi.org/10.1109/BigData.2013.6691556

[2] Downing, J. J. Foster, G. Lloyd, The NERC DataLabs Initiative, proceedings of JASMIN2018 Conference, June 2018, Harwell http://www.jasmin.ac.uk/jasmin2018/

[3] Smith, R., W. Garland, P.J. Kershaw, G.A.Parton, A. Stephens, Elastic Search to Geo Search: Delivering web-based search tools for large volume, heterogeneous airborne, in-situ and satellite-based observations, ESA EO Open Science Conference, September 2017, DOI: 10.13140/RG.2.2.20474.59843

[4] Graham A Parton, Steven Donegan, Stephen Pascoe, Ag Stephens, Spiros Ventouras, Bryan N Lawrence, MOLES3: Implementing an ISO standards driven data catalogue; International Journal of Digital Curation, Vol 10 No 1 (2015); DOI: https://doi.org/10.2218/ijdc.v10i1.365

[5] Lawrence, B.N., R. Lowry, P. Miller, H. Snaith, A. Woolf, Information in Environmental Data Grids, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1890), 1003–1014., DOI: https://doi.org/10.1098/rsta.2008.0237

[3] http://geo-search.ceda.ac.uk/

[4] https://github.com/cedadev/SemSL

COPERNICUS GLOBAL LAND MAPPING FROM PRIVATE TO PUBLIC CLOUD

Bruno Smets, Marcel Buchhorn, Dirk Daems

VITO, Boeretang 200, 2400 Mol, Belgium, http://www.vito.be

ABSTRACT

VITO and partners are responsible for the global land cover mapping within the Copernicus Global Land service [1]. The land cover maps are initially derived from the 100m PROBA-V [2] time-series and were generated through a full reprocess of the PROBA-V archive to improve accuracy at high latitudes, to align with Sentinel 2 UTM tiling grid, and to generate a PROBA-V Analysis Ready Data (ARD) archive. The workflow has been developed and executed on the PROBA-V Mission Exploitation Platform (MEP) [3] in an operational context providing the necessary means to check the processing status and outputs of handling millions of files. The workflow supports the ingestion of Sentinel datasets, hence enabling global land mapping at 20m or even 10m resolution. VITO has developed a solution to deploy its workflow from its private MEP cloud onto a public cloud, i.e. DIAS or EODC, and hence deal with the required exponential expansion of the required resources without substantial altering the available workflow.

Index Terms— Copernicus, MEP Mission Exploitation Platform, dynamic land cover mapping, DIAS, ARD Analysis Ready Data, data analytics on time series, virtual resource environment

1. INTRODUCTION

The Copernicus Global Land Service (CGLS) is a component of the Land service to operate "a multi-purpose core service component". It targets to monitor the status and evolution of the land surface at global scale. The service provides next to a series of bio-geophysical products, a dynamic land cover map at 100m spatial resolution. Next to discrete mapping, a continuous classification scheme, aka known as continuous cover layers, is introduced to depict areas of heterogeneous land cover better than a standard classification scheme (Figure 1). This advanced classification scheme allows the user to tailor the land cover product to his application and needs (e.g. forest monitoring, crop monitoring, biodiversity and conservation, monitoring environment and security, etc.).

The first Land Cover map (version 1) was provided for the African continent for the 2015 reference year and based on the available PROBA-V Collection 1 archive [4]. The second edition of the Land Cover maps (version 2) is under preparation for the entire globe covering the 2015 reference year with yearly updates. To avoid distortions in the northern hemisphere and to be prepared to ingest Sentinel data, the entire workflow has been revised to support S2-UTM tiles as well as the entire PROBA-V archive has been reprocessed.

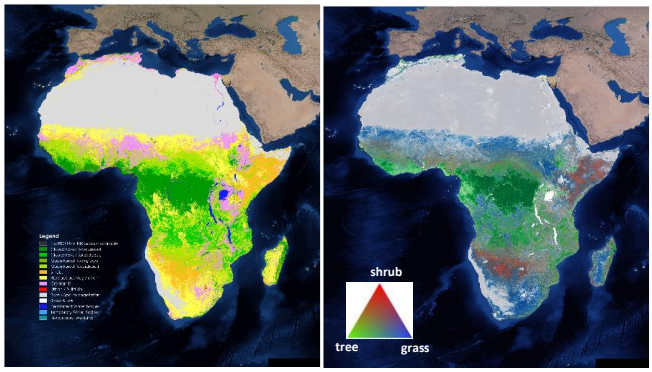


FIGURE 1 : COPERNICUS LAND COVER V1, LEFT DISCRETE MAP (18 CLASS), RIGHT CONTINUOUS COVERS (0-100%)

To realize this land cover mapping and its pre-processing, the PROBA-V MEP platform was used [5]. It provides scalable processing facilities with access to the complete PROBA-V data archive and a rich set of processing algorithms and open source processing libraries/toolboxes. The land cover map workflow was developed on this platform as well as executed. The workflow was also tested, on a limited area, through using Sentinel-1 and Sentinel-2 data and is being prepared to be deployed on a public cloud for a large scale test.

2. GLOBAL LAND COVER WORKFLOW

2.1. Land Cover workflow

The workflow consists of three major parts, as shown in Figure 2:

- EO data pre-processing
- LC pre-processing
- LC classification & post-processing

First, the Earth Observation input data needs to be prepared. With PROBA-V, the non-projected L1C segments are projected using the Sentinel-2 UTM tiling grid before the atmospheric correction is performed to generate L2B products. The final step is to generate single day composites by selecting the best quality pixel of the geometric overlapping segments to create an Analysis Ready Data

doi:10.2760/848593

(ARD) stack to start the land cover mapping. With Sentinels a similar ARD stack is created. The Sentinel-1 Synthetic Aperture Radar (SAR) data is first transformed into gamma0 and coherence information to be used as ARD.

Then the ARD information is further cleaned, composited into 5-daily median composites and in case of PROBA-V fused through a Kalman filter [6] with the daily 300m timeseries to fill gaps in the 100m dataset, before the actual metrics are calculated. The metrics ARD stack consists of several hundred metrics originating from the spectral signal, the spatial texture, statistical descriptions, topographic information and some phenological metrics.

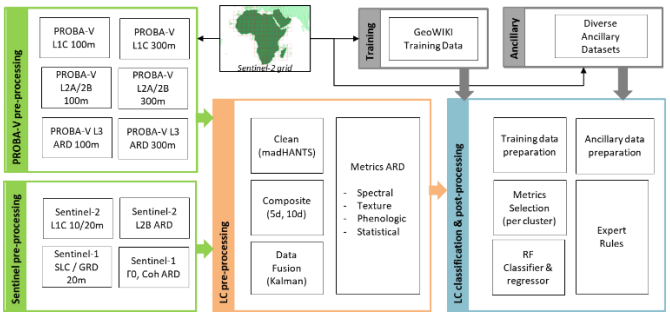


FIGURE 2 : COPERNICUS GLC WORKFLOW

Training data, more than 120,000 points at 100m resolution, with description at 10x10m, are gathered through the Geo-WIKI platform [7]. This training data is screened for outliers and prepared for classification. The training of the Random Forest (RF) classifier and regressor is performed through a 5-folded Cross-Validation to select the best metrics and hyper-parameters as well as is executed per biome cluster zone. In a final step, ancillary data is prepared and fused into a decision tree to incorporate areas of agreement of existing Global Land Cover maps and to imprint specific classes from ancillary datasets (i.e. urban, glaciers, etc.).

2.2. MEP Platform

The Mission Exploitation Platform (MEP) is embedded into VITO Remote Sensing Data Center. This data center contains at writing this paper a tiered storage of 7 PB disk storage and 5PB tape archive, about 650 physical servers and 300 virtual servers with a network capacity of 10GB internal and external through Géant.

The MEP platform is a scalable private cloud platform used to execute the land cover workflow and at writing this paper contained 90 active nodes providing 2,800 virtual cores and 10 TB memory or in average 3.5 GB available memory per core (Figure 3).

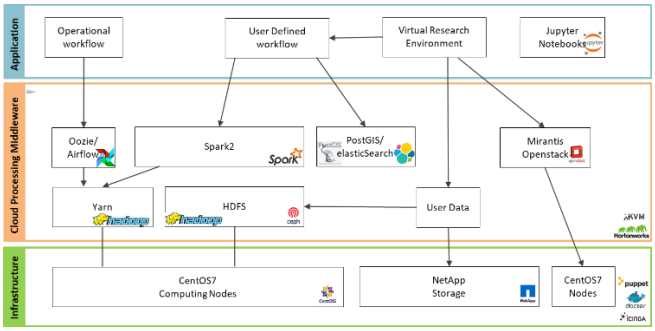


FIGURE 3 : MEP PRIVATE CLOUD SYSTEM

MEP platform users can request access to a Virtual Research Environment (VM) or a Jupyter notebook, both containing a rich set of pre-installed image processing tools and libraries with direct access to the complete data archive. The virtual machines are deployed on a privately hosted OpenStack cloud and all have direct access to the complete data archive. After a prototyping and testing phase, processing can be scaled out by making use of the shared resources of the Hadoop platform. While this transition often takes a lot of time on other platforms, it is painless on the MEP platform due to the use of Spark and alignment of library versions on both the virtual machines and the Hadoop processing nodes.

Hadoop, a software framework for data-intensive distributed applications, is designed to process large amounts of data by separating the data into smaller chunks and performing large numbers of small parallel operations on the data. Yarn is used as resource manager and enables to share resources between multiple applications. Spark is used intensively on the MEP to allow analytics on large time series of data. The Hadoop ecosystem provides a rich and still growing set of tools which are used to give fast access to the data in a format needed by the specific application. The platform also provides tools to operationally schedule workflows (i.e. Airflow) and perform quality control (i.e. PostGIS and ElasticSearch).

All EO raster data is accessible via NFS and possibly uploaded to the Hadoop Distributed Filesystem (HDFS) if beneficial.

3. DEPLOY WORKFLOWS IN PUBLIC CLOUDS

VITO has prepared a system to deploy its Spark cluster, as explained in the previous paragraph, in a public cloud environment (Figure 4). This solution enables to setup new processing clusters with relaxed security, resource management and high-availability requirements as they are flexibly instantiated on a per use basis and do not require any sharing. Such Spark cluster can be deployed on any cloud infrastructure, as provided by the DIAS or other clouds hosting the Sentinel-1, Sentinel-2 and Landsat-8 datasets.

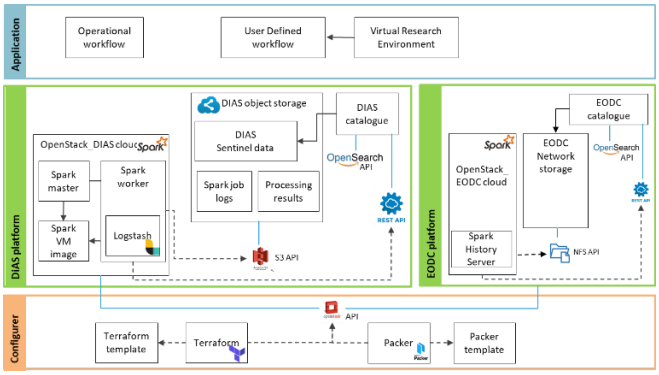


FIGURE 4 : SYSTEM FOR CLOUD DEPLOYMENT

Although some cloud providers offer Hadoop or Spark as a service (SaaS), this is not yet the case on the DIAS clouds. Furthermore it is considered best practice to provision the cloud resources through an abstraction layer. By using open-source Terraform software, we support automated deployment of a Spark processing cluster on both the well-known public cloud providers and the DIAS (OpenStack-based) cloud providers. Although some cloud providers (e.g. Sobloo / Orange cloud) also provide Terraform resources for their proprietary cloud API's, the system is entirely built on the generic OpenStack API's where possible.

As of Spark version 2.3.0, Kubernetes is also supported as cluster manager backend. Although we believe that support for Kubernetes will become stable in future Spark versions, it is currently (version 2.4.0) experimental. Also, first experiments on DIAS cloud providers showed that the DIAS managed Kubernetes support is not yet stable. Therefore, our solution sets up a Spark cluster using Spark standalone mode, deployed on virtual machines which were provisioned in an automated way using Packer and Terraform.

When a processing cluster is spawned, the workflow can be started. To query the Sentinel datasets, a DIAS cloud typically exposes one or more catalogue interfaces (e.g. REST, OpenSearch) which can be used to query the Sentinel products available on object storage. These products, available in the native SAFE format, can be accessed through the Simple Storage Service (S3) API.

Due to the transient nature of cloud processing resources, the logs generated by the processing chain are pushed to a dedicated S3 bucket using Logstash or the Spark history server, which allows them to be consulted after the processing cluster was destroyed.

4. RESULTS

4.1. PROBA-V land cover

As explained earlier, the Land Cover workflow was completely revised. Version 1 was based on the available PROBA-V L3 archive with 400 tiled images globally per day. Version 2 has reprocessed over 50 million single acquisitions to generate a global PROBA-V UTM ARD, fully aligned to the Sentinel-2 tiling grid with nearly 16,000 land tiles, as shown in Figure 5. A PostGIS/PostgreSQL server is used to keep track on the processing status of the 50 million images.

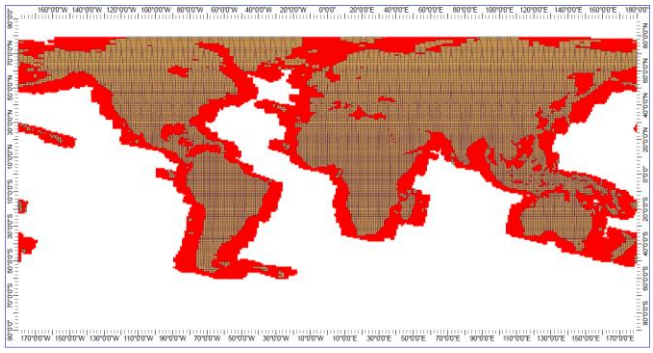


FIGURE 5 : PROBA-V (S2) UTM TILES FOR LAND MASSES (PROBA-V IMAGING AREA IN RED)

The workflow consists of ~45,000 lines of code and was highly optimized to use less memory resources to further increase parallelism if required. Version 1 required 1,200 executors with 5 TB memory, while Version 2 processes the same job even faster with 500 executors only using 750 GB memory.

Overall the entire EO archive pre-processing of 4 years of data, in total 45 Million files times 4 steps or 175 Terra-bytes of data, required 400 hours or less than two weeks of processing using no more than 500 executors or 18% of the MEP platform. The actual land cover pre-processing and classification runs in less than 70 hours while version 1 took about 160 hours. Nearly 75% of the time is needed for the LC pre-processing (additional data cleaning and metrics generation), however about 400 metrics are calculated. An example (Red = Sum HUE, Green = 10^{th} percentile NDVI, Blue = RED band median) is shown in the Figure 6 below.

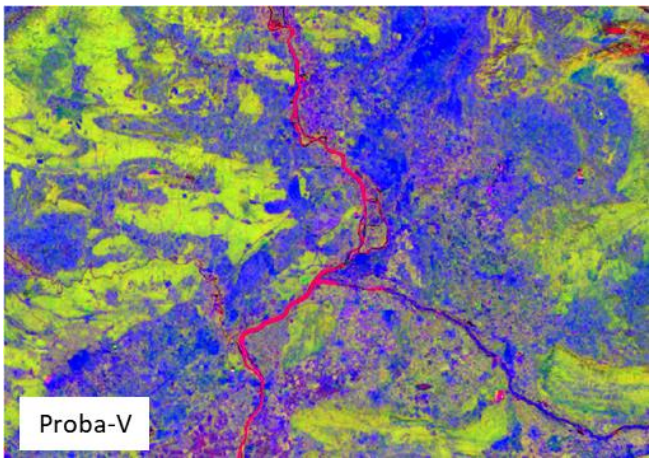


FIGURE 6 : PROBA-V 100M METRICS, AVIGNON

doi:10.2760/848593

Every tile consists of about 1.2 million pixels and the processing is performed in parallel per UTM zone and tiles. As such a continent as Africa could be run simultaneously by reserving more than 3,500 executors, if available.

The LC workflow is highly agile, iterative such that in no more than 2 days, new training data can be collected, a new classification can be performed, and the validation results are available. The Africa continent is used as a reference to further improve the classification accuracy and in the meantime already 16 classification runs (scenarios) were performed, as well as 1 to 2 runs per other continents. Since the V2 workflow uses less resources, multiple continental classifications are run in parallel.
The amount of intermediate results kept in the platform is optimized to minimize the costs, hence only the results of the three major workflow steps is kept online, see Figure 2.

4.2. Sentinel land cover

The Land Cover workflow has been used to perform a first classification based on Sentinel data. To test the workflow, the Sentinel data over 32 UTM tiles was downloaded to the MEP platform, atmospherically corrected through the iCOR toolbox, and a first execution has been performed at 10m and 20m spatial resolution. The workflow could be re-used, adding some metrics through making use of the higher spectral detail of the input data. However, the amount of memory required to perform the processing at 10m spatial resolution, compared to the 100m, is higher and hence reduces the number of available executors. A solution currently in progress is to further chunk the tiles into smaller blocks for processing in order to keep the memory within the same limits as the 100m processing.

An example of the Sentinel metrics ARD can be found in Figure 7 showing Sentinel1 (Red = 12 day Mean Coherence, Green = Gamma0 VH, Blue = 12 day Standard Deviation Coherence) and Sentinel2 (Red = Sum HUE, Green = 10^{th} percentile NDVI, Blue = RED band median) composites.

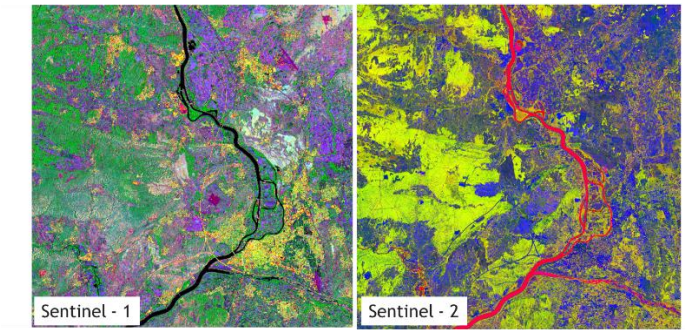


FIGURE 7 : SENTINEL METRICS 10M, AVIGNON

To prepare the deployment of Sentinel Land Cover workflow to the public cloud, a simple workflow has been created. This workflow calculates the Normalized Difference Vegetation Index (NDVI) from two Sentinel-2 bands which was developed on the MEP and then tested on the Sobloo DIAS [8] using VITOs stand-alone Spark solution with a limited number (4) of executors. To test interoperability of the solution, a second test has been performed on the public Earth Observation Data Centre (EODC) [9]. The terraform templates make it easy to configure the Spark stand-alone environment to the public cloud platform, however the interface to access the data is not standardized and requires some small adaptations in the workflow.

5. CONCLUSIONS & FURTHER WORK

The amount of data in remote sensing has increased exponentially the last two years. This paper has shown that workflows can be developed to support multiple sensors and be executed in a big data environment coping with the volume, velocity, variety and veracity. The value of our solution is that it is easy to scale up the processing and deploy the workflow on a public cloud that hosts the required big datasets and offers the best service.

The next steps are to deploy the full land cover workflow on the public cloud and perform a test at large scale, extending the algorithm to further make use of the richness of the Sentinel datasets and to investigate the use of more standardized interfaces, as already explored in the OpenEO project [10].

Acknowledgement.
This work was supported by the European Commission, DG-JRC through the Copernicus program of DG-GROW (Copernicus Global Land Service); and by ESA through the ESA Earthwatch program (Mission Exploitation Platform, a Virtual Research Environment) and by ESA through DIAS program (Copernicus DIAS Airbus).
The global land cover maps are developed in partnership with IIASA and Wageningen University.

6. REFERENCES

[1] http://land.copernicus.eu/global/
[2] http://proba-v.vgt.vito.be/
[3] https://proba-v-mep.esa.int/
[4] Buchhorn et.al, Algorithm Description Copernicus Global Land Cover V1 https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/ CGLOPS1_ATBD_LC100m-V1_I1.00.pdf
[5] Proba-V Mission Exploitation Platform, Remote Sensing Journal, Technical Note, 2 July 2016, http://www.mdpi.com/2072-4292/8/7/564/pdf.
[6] Kempeneers et.al, Data Assimilation of PROBA-V 100 and 300m, IEEE Geoscience & Remote Sensing Society, DOI 10.1109/JSTARS.2016.2527922
[7] https://www.geo-wiki.org/
[8] https://sobloo.eu/
[9] https://www.eodc.eu/
[10] https://openeo.org/

EVER-EST: THE PLATFORM ALLOWING SCIENTISTS TO CROSS-FERTILIZE AND CROSS-VALIDATE DATA

Mirko Albani[1], Cristiano Silvagni[2], Rosemarie Leone[1], Fulvio Marelli[3], Sergio Albani[4], Michele Lazzarini[4], Anca Popescu[4], Federica Foglini[5], Francesco De Leo[5], Valentina Grande[5], Stefano Salvi[6], Elisa Trasatti[6], Hazel Napier[7], Tim Aldridge[8], Steven Cole[9], Robert Moore[9], Iolanda Maggio[1]

[1] European Space Agency ESA-ESRIN, Largo Galileo Galilei 1, 00044, Frascati, Italy
[2] European Space Agency ESA-ESAC,Camino bajo de Castillo S/N
[3] Terradue, Via Giovanni Amendola 46, 00185, Rome, Italy
[4] European Union Satellite Centre, Apdo de Correos 511, 28850, Torrejón de Ardoz, Spain
[5] CNR ISMAR via Gobetti 101, 40129 Bologna, Italy
[6] INGV-ONT, via di Vigna Murata 605, 00143, Roma, Italy
[7] British Geological Survey, Nicker Hill, Keyworth, Nottingham, UK, NG12 2DL
[8] Health and Safety Executive, Harpur Hill, Buxton, UK, SK17 9JN
[9] Centre for Ecology & Hydrology, Wallingford, Oxon, UK, OX10 8BB

ABSTRACT

Over recent decades large amounts of data (Big Data time period) about our Planet have become available. If this information could be easily discoverable, accessible and properly exploited, preserved and shared, it would potentially represent a wealth of information for a whole spectrum of stakeholders: from scientists and researchers to the highest level of decision and policy makers. By creating a Virtual Research Environment (VRE) tailored to the needs of Earth Science (ES) communities, the EVER-EST (http://ever-est.eu) project provides a range of both generic and domain specific data analysis and management services to support a dynamic approach to collaborative research.

Index Terms— Virtual Research Environment, Remote Sensing, Research Object, Cross-fertilization, Data Analysis, Earth Science, Education.

1. INTRODUCTION

EVER-EST is funded by the European Commission H2020 programme for three years starting in October 2015. The project is led by the European Space Agency (ESA) and involves some of the major European Earth Science data providers/users including NERC, DLR, INGV, CNR and SatCEN. The paper presents specific aspects of this collaboration platform in terms of infrastructure and implemented paradigms. Some case studies on cross-fertilization analysis are documented in order to show the process for creating knowledge and new data starting from collected data from different sources (e.g. from remote and social sensing). For each use case the outcomes are presented.

2. EVER-EST PLATFORM

EVER-EST is a research and development platform that offers a framework based on advanced services to support each phase of the Earth Science Research and Information Lifecycle. The project follows a user-centric approach which have produced a wealth of innovative and state-of-the-art technologies, systems and tools for e-collaboration, e-learning, e-research, big data management and long term data preservation.

3. RESEARCH OBJECT

Central to the EVER-EST approach is the concept of the Research Object (RO), which provides a semantically rich mechanism to aggregate related resources about a scientific investigation so that they can be shared together using a single unique identifier. The original definition of RO is available in Bachhofen et al. [2]. Although several e-laboratories are incorporating the research object concept in their infrastructure, the work done with research objects during EVER-EST, is a novel effort done to adapt the RO model to Earth Science and support automatic generation of research object content-based metadata as presented at the 2017 IEEE 13th International Conference on e-Science [1]. The EVER-EST VRE is the first infrastructure to leverage the concept of Research Objects and their application in observational rather than experimental disciplines. Research objects aim to account, describe and share everything about your research, including how those things are related. A Research Object (RO) is defined as a

semantically rich aggregation of resources that bundles together essential information relating to experiments and investigations. This information is not limited merely to the data used and the methods employed to produce and analyse that data, but it may also include the people involved in the investigation as well as other important metadata that describe the characteristics, inter-dependencies, context and dynamics of the aggregated resources. As such, a research object can encapsulate scientific knowledge, workflows and provide a mechanism for sharing and discovering assets of reusable research and scientific knowledge within and across relevant communities, and in a way that supports reliability and reproducibility of investigation results [4].

More specifically, by encapsulating workflows, using Apache Taverna, into research objects and accompanying them with the necessary data and metadata needed for their execution and understanding, one makes the latter more (re-)usable and preservable. This metadata can include, among others, details like authors, versions, citations, etc., and links to other resources, such as the provenance of the results obtained by executing the workflow or datasets used as input. Such additional information enables a comprehensive view of the scientific investigation, encourages inspection of its different elements, and provides the scientist with a clearer picture of the investigation's strengths and weaknesses with respect to decay, adaptability and stability.

The research object recommendation system that shall be used as a basis in this project is again derived from the WF4Ever project and consists of two components:

- The Research Object Recommendation Service API that combines a variety of recommender algorithms.
- The Collaboration Spheres Web Application, that implements a novel visual metaphor for the more intuitive interaction of the user.

In turn, the visual metaphor implemented by the Collaboration Spheres web application is based on a set of concentric spheres centred around a central point that represents the user. These spheres represent different types of similarity metrics between the context of interest and the results obtained by the recommenders. The closer to the center, the more specific the recommendation result will be with respect to the user and the current context of interest.

3. VIRTUAL RESEARCH ENVIRONMENT

The EVER-EST e-infrastructure is validated by four virtual research communities (VRC) covering different multidisciplinary Earth Science domains including: ocean monitoring, natural hazards, land monitoring and risk management (volcanoes and seismicity).

- Land Monitoring: Monitoring of urban, built-up and natural environments to identify certain features or changes over areas of interest.
- Sea Monitoring: The Sea Monitoring VRC focuses on finding new ways to measure the quality of the maritime environment and it is quite wide and heterogeneous, consisting of multi-disciplinary scientists such as biologists, geologists, oceanographers and GIS experts, as well as agencies and authorities.
- Geohazard Supersites and Natural Laboratories: is a collaborative initiative supported by GEO (Group on Earth Observations) within the Disasters Resilience Benefit Area. The goal of GSNL is to facilitate a global collaboration between Geohazard monitoring agencies, satellite data providers and the Geohazard scientific community to improve scientific understanding of the processes causing geological disasters and better estimate geological hazards.
- Natural Hazards Partnership: is a group of 17 collaborating public sector organisations comprising government departments, agencies and research organisations. The NHP provides a mechanism for providing co- ordinated advice to government and those agencies responsible for civil contingency and emergency response during natural hazard events.

5. CASE STUDY: LAND MONITORING

- CHANGE DETECTION: The Change Detection service allows to select a pair of Sentinel-1 GRD images, within a timeframe, and to identify changes through suitable algorithms. The service has been deployed on the T2 Sandbox and it can be initiated via the EVER-EST Land Monitoring Portal through a Web Processing Service (WPS). It represents a pre-operational use of the EVER-EST infrastructure for the targeted community. The service launches a set of chained processing modules based on the Sentinel Application Platform (SNAP): Thermal Noise Removal, Orbit-Based Correction, Calibration, Terrain Flattening and Terrain Correction. Successively, the images are co-registered and a Change Detection algorithm identifies the areas with changes. The output of the Change Detection is a raster product containing the pixels where changes have been detected. The Land Monitoring use case provided a concrete case of cross-disciplinary interaction between Earth Scientists and Institutional entities to transfer knowledge, best practices and tools.

6. CASES STUDY: SEA MONITORING

- EVALUATE HOW HUMAN ACTIVITIES CAN CAUSE POSIDONIA MEADOWS REGRESSION: Coastal anthropogenic activities increased worldwide in the last half century, amplifying the pressures on marine coastal ecosystems. The management of those

doi:10.2760/848593

multiple and simultaneous threats requires reliable and precise data on the distribution of the pressures and of the most sensitive ecosystems. In this case study, starting from historical remote sensing data of Posidonia meadows distribution, the Sea Monitoring (SM) VRC detected Posidonia regression areas off shore the Apulia region in Italy and compared their distribution with the different human activities identified by the Change Detection WPS developed by Land Monitoring (LM) VRC. LM run the change detection WPS using the EVER-EST VRE service in the Apulia Region and created a RO encapsulating the Taverna workflow and the results as .shp file. In parallel SM run runs a workflow implemented to detect Posidonia regression using the EVER-EST VRE Virtual Machine and created a RO with data, results, and workflows. Overlaying through the EVER-EST VRE globe the results from the LM and SM research object it was possible to visually identify a correlation visual between the human activities detected by LM and the Posidonia regression off shore Gallipoli detected by SM. Among the various types of human activities, the mechanical damages resulting from boats anchoring in shallow coastal waters appear to be responsible for localized regressions of Posidonia oceanic meadows.

- CORRELATION BETWEEN ENVIRONMENT SATELLITE VARIABLES AND JELLYFISH OUTBREAKS: Cross-Fertilisation study in synergy between University of Tor Vergata in Rome and UniSalento biological researchers group located in Lecce. The group is specialized on the quantification of deterministic and stochastic components of environmental change that lead to outbreaks of maritime species: in this specific case, the jellyfish. The Research Objects created by UniSalento have been cross-fertilized with the RO on “Mediterranean Sea Anomalies detection” developed during the Master. This can be considered as a good example of joint work between two communities – Earth Observation researchers and Maritime Biologist – which could be not necessarily strictly linked in their everyday activities and that was de facto facilitated by the common use of RO’s and the adoption of the EVER-EST infrastructure as working environment. The analysis led to the successful identification of correlations between the two phenomena over specific areas of the Adriatic Sea. Partial results were collected in terms of light correlations with temperature, chlorophyll and particulate.
- HABITAT SUITABILITY MODEL - BARI CANYON: Habitat Suitability Model of the Cold Water Corals (CWCs) in the Bari Canyon (Apulia, Italy). In this RO we derive the MSFD indicator 1.5 (Habitat area) to assess the biological diversity descriptor. To do this in deep sea environment, the scientist (user) needs to implement a habitat suitability model.
- JELLYFISH SPECIES DISTRIBUTION ALONG ITALIAN COAST: Starting by sightings from citizen science campaign "Occhio alla medusa"; CNR wants to fully exploit within the EVER-EST initiative the database potential to generate meaningful indicators (species distribution) in MSFD perspective.
- TREND IN THE EVOLUTION OF NON INDIGENOUS JELLYFISH SPECIES: Starting from Jellyfish sightings, we elaborate data to produce explicit geographical information concerning trend about the evolution and distribution of alien species according with MSF directive descriptors “Abundance and state characterisation of non-indigenous species (NIS), in particular invasive species (IAS)”.
- DIGITALIZATION OF HISTORICAL VENICE LAGOON MAPS: Historical maps comprise a lot of inherent information on natural environmental and anthropogenic changes. They are commonly the most important database for various spatial analyses of the land use as well as historical landscapes, urban development, influences of the economy development, toponyms changes, etc.
- MULTIPLE AND PERVASIVE HUMAN IMPACTS IN COASTAL LAGOONS LITERATURE REVIEW: Coastal wetlands are among the most studied, most vulnerable, and economically most important ecosystems on Earth; nevertheless, little attention has been paid, so far, to their sea-floor integrity and the human footprint on their deepest reaches.
- POSIDONIA REGRESSION ALONG APULIAN COAST CROSS-FERTILISE LAND MONITORING VRC: In our study case, starting from historical data of posidonia meadows distribution, we try to individuate regression area and to compare their distribution with the different human activities that can determinate change in the Land/Sea use detecting by WPS developed by Sat Cen VRC.
- RMS FROM BATHYMETRY: This RO calculates the roughness of the seafloor, as RMS, starting from Multibeam Bathymetry. It was applied with bathymetry files of the Venice Lagoon.

7. CASES STUDY: SUPERSITES

- VOLCANIC PLUME RETRIEVALS PROCEDURES: During eruptions, volcanoes emit large quantities of particles and gases into the atmosphere. The Volcanic Plume Retrieval procedure has the capability, simultaneously and in real time, to estimate physical parameters of volcanic ash and SO2 clouds from multispectral MODIS data in the Thermal InfraRed (TIR) spectral range. Plume altitude and temperature are the only two input parameters required to run the procedure. By linearly interpolating the radiances

doi:10.2760/848593

surrounding a detected volcanic plume, the VPR procedure computes the radiances that would have been measured by the sensor in the absence of a plume, and reconstructs a new image without plume. The new image and the original one allows computation of plume transmittance in the TIR-MODIS bands 29, 31, and 32 (8.6, 11.0 and 12.0 μm) by applying a simplified model consisting of a uniform plume at a fixed altitude and temperature. The transmittances are then refined using a polynomial relationship obtained by means of MODTRAN simulations adapted for the geographical region, ash type, and atmospheric profiles.

- VOLCANIC GEODETIC DATA INVERSION: The RO was created to invert 2004-2006 ground deformation data for the Campi Flegrei volcano. The inverted datasets were ascending and descending Line of Sight ground displacements from COSMO-SkyMed InSAR time series. The data were modelled with a spherical magma chamber. At the end of his inversion procedure, Elisa created a RO containing the input data, the inversion workflow, and the output results, then added some descriptive information and finally archived the RO with a DOI to ensure authorship of the research.
- INSAR PROCESSING WITH SARSCAPETM ON A WINDOWS VIRTUAL MACHINE: This use case shows how to download Sentinel 1 SAR image data from the EVER-EST VRE interface, and launch the SARscape SAR processing software in a Windows Virtual Machine to carry out Interferometric SAR processing.

8. CASES STUDY: NATURAL HAZARDS

- SURFACE WATER FLOODING: The Surface Water Flooding Hazard Impact Model (SWF HIM) is a well-developed Hazard Impact Model approaching operational deployment with on-going work focussed on validation of impacts through chosen case studies. This involves running a countrywide (1km grid, 15 min time-step) Grid-to-Grid (G2G) hydrological runoff and routing model (CEH) using rainfall inputs (Met Office), and linking its surface runoffs to potential impacts (HSE) and verifying these against observed impacts.
- DAILY HAZARD ASSESSMENT (DHA): The DHA is a summary of forecasted hazards released on a daily basis to the responder community, local government and national agencies. It is based on information provided by various partner organisations including the FGS, the NSWWS and the DLHA. Each piece of evidence is linked by date, however if any of the evidence is updated due to a change in the hazard forecast, then an updated piece of evidence is submitted for inclusion in the DHA. The VRC decided to test the storage of each DHA and its contributing evidence in a bibliographic Research Object.

9. CONCLUSIONS

During the three-year project, the EVER-EST consortium developed a VRE for Earth Sciences, where the requirements of four communities were addressed. The VRE has been recognized as a successful solution to boost open science and innovation by enabling research life cycle management, long term data preservation, EO data exploitation and capacity building. A sustainability plan has been presented to maintain the findings after the end of the project: in addition, further efforts will be focused on make the platform fully operational (e.g. services improvement, architecture optimisation and user support), to improve the services model and to engage new communities.

10. REFERENCES

[1] Gomez-Perez, J.M., Palma, R., Garcia-Silva, A.: Towards a human-machine scientific partnership based on semantically rich research objects. In: 2017 IEEE 13th International Conference on e-Science (e-Science). pp. 266–275 (Oct 2017)

[2] S Bechhofer, I Buchan, D De Roure, P Missier, J Ainsworth, J Bhagat, P Couch, D Cruickshank, M Delderfield, I Dunlop, M Gamble, D Michaelides, S Owen, D Newman, S Sufi, and C Goble. Why linked data is not enough for scientists. Future Generation Computer Systems, 29(2):599 – 611, 2013. Special section: Recent advances in e-Science.

[3] K Belhajjame, O Corcho, D Garijo, J Zhao, P Missier, DR Newman, R Palma, S Bechhofer, E Garcia-Cuesta, JM Gomez-Perez, G Klyne, K Page, M Roos, JE Ruiz, S Soiland-Reyes, L Verdes-Montenegro, D De Roure, and C Goble. Workflow-centric research objects: A first class citizen in the scholarly discourse. In 2nd Workshop on Semantic Publishing (SePublica), number 903 in CEUR Workshop Proceedings, pages 1–12, Aachen, 2012.

[4] R Palma, P Hołubowicz, O Corcho, JM Gomez-Perez, and C Mazurek. Rohub—a digital library of research objects supporting scientists towards reproducible science. In Semantic Web Evaluation Challenge, pages 77–82. Springer, 2014.

[5] ESA, NERC, INGV, ISMAR, SatCen, “Use Cases Description and User Needs” , EVER-EST DEL WP3-D3.1

[6] ESA, NERC, INGV, ISMAR, SatCen, “Workflows and Research Objects in Earth Science - Concepts and Definitions”, EVER-EST DEL WP4-D4.1

[7] ESA, NERC, INGV, ISMAR, SatCen, “VRE Architecture and Interfaces Definition” , EVER-EST DEL WP5-D5.1

[8] Lisandro Benedetti-Cecchi, Antonio Canepa, Veronica Fuentes, Laura Tamburello, Jennifer E. Purcell, Stefano Piraino, Jason Roberts, Ferdinando Boero, Patrick Halpin, “Deterministic Factors Overwhelm Stochastic Environmental Fluctuations as Drivers of Jellyfish Outbreaks”, doi:10.1371/journal.pone.0141060

[9] EVER-EST: A VIRTUAL RESEARCH ENVIRONMENT FOR EARTH SCIENCES Paper in BiDS2019 conference

Online Data Access and Big Data Processing in the German Copernicus Data and Exploitation Environment (CODE-DE)

Christoph Reck [1], *Tobias Storch* [1], *Stefanie Holzwarth* [1], *Michael Schmidt* [2]

[1] DLR EOC, Münchner Str. 20, 82234 Weßling, Germany
[2] DLR Space Administration, Königswinterer Str. 522-524, 53227 Bonn, Germany

{christoph.reck, tobias.storch, stefanie.holzwarth, michael.schmidt}@dlr.de

ABSTRACT

We present architecture and various capabilities of CODE-DE (Copernicus Data and Exploitation Platform – Deutschland, www.code-de.org) which is the German operational environment for accessing and processing Copernicus Sentinel products, a so-called Copernicus collaborative ground segment activity. Since March 2017 the element for online data access to Sentinel-1 and Sentinel-2 products is operational and tapped by over 1200 registered users until September 2018. During this period more than 100,000 products were downloaded and the global catalogue is continuously updated with all Sentinel product metadata and references a data volume of 800 TByte accessible on a rolling archive. Since November 2017 the element for big data processing is operational, where registered users automatically process and analyze data using various methodologies into value-added products. Special features enhance the user experience, like the full resolution browsing at 10 meter resolution of Sentinel-2 products giving interactive insight to the catalog contents. In 2018 the full spectrum of available Sentinel-3 products also are offered online.

Index Terms: Copernicus, Sentinel, CODE-DE, Online Data Accessing, Big Data Processing

1. INTRODUCTION

The fleet of Copernicus Sentinel satellites provides unprecedented opportunities for global environmental monitoring. However, the capability to effectively and efficiently access, manage, process, and analyze the mass data streams from the Sentinels, but also from other big data missions such as the Landsat program, still poses major conceptual and technical challenges. The German Aerospace Center (DLR) works towards bridging the gap between the immense data volumes collected by modern Earth Observation missions and their application-driven, on-demand exploration through geo-information services [2].

CODE-DE (Copernicus Data and Exploitation Platform – Deutschland, www.code-de.org) which is realized by the Earth Observation Center (EOC) of DLR is the German entry point to the EU Copernicus Sentinel Satellite Systems under the framework of the Copernicus Collaborative Ground Segments. It provides their data products and the products of the Copernicus Services [4] with a focus on fulfilling national needs.

2. METHODOLOGY AND ARCHITECUTRE

The client, which is the DLR Space Administration on behalf of the German Federal Ministry of Transport and Digital Infrastructure (BMVI), provided 156 User Requirements (REQ) for a complete and consistent description of CODE-DE, covering the national needs on data access and the capabilities to process data.

For the implementation and to fully exploit the possibilities of the continuous data stream of free, full, and open Copernicus Sentinel products the development of the CODE-DE system was conducted with work packages for project management, product assurance, and systems engineering. The platform was designed based on several subsystems: Infrastructure, Portal, Ingestion and Archive, Search and Access, User Management Service, Processing Environment, Value Added Products, Monitoring and Reporting, Help Desk and Operations. The architecture is illustrated in Figure 1.

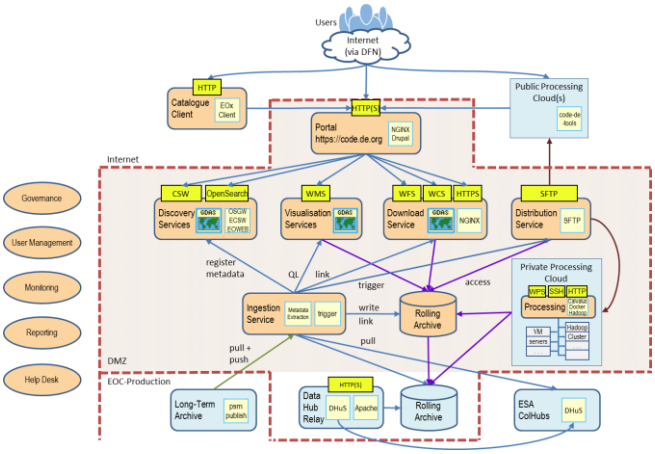


Figure 1. CODE-DE Architecture

Each subsystem lies within the responsibility of a contracted supplier. These subsystems are further broken down to 42 components – where each component provided a

doi:10.2760/848593

specific functionality on its own. A configuration control is performed for the items at the level of hard- and software. Items of type software are deployed in Virtual Machines (VMs) which are assigned to a server, namely an item of type hardware.

The infrastructure consists of a dedicated GPFS (general parallel file system) with 1,444 TByte disk storage and nineteen VMs deployed on six service hosts with 244 cores and 1,536 GByte RAM, located in Frankfurt, Germany. The User Management System as well as the Monitoring and Reporting modules rely on external computing environments and services which are shared with other projects. The hosted processing environment currently consists of four nodes with 112 cores and 512 GByte RAM. The CODE-DE system is linked to the internet via a 5 GBit/s connection.

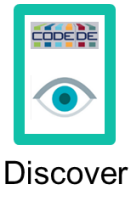
The CODE-DE system went operational in three major iterations:

- Release 1.0 on 2017-03-09, focused on online data access of Sentinel-1 and Sentinel-2 data products
- Release 2.0 on 2017-11-30 focused on big data processing (see Section 5) for registered users to automatically process and analyze data applying various methodologies to value-added products. Improved visualization of the catalogue of available Sentinel-2 products with the catalogue client at full 10 meter resolution (see Figure 2)
- Release 2.1 on 2018-04-26 completed the client requirements; makes Sentinel-3 products accessible, and is now taking onboard the applications.

In order to guarantee the full functionality and high system availability of the CODE-DE services, intense test activities are conducted. Failed test case execution is linked with a corresponding observation. These test activities are complemented by daily manual operation tests, automatic monitoring and reporting activities, and regression tests, when the system is modified.

3. COPERNICUS SENTINEL PRODUCTS

Copernicus Sentinel-1, Sentinel-2 and Sentinel-3 products each with respective satellites A and B are continuously ingested and archived in CODE-DE. The data are provided via data hubs to the EU member states, namely the Copernicus Collaborative Ground Segments hubs. These require to be connected to the internet with at least 1 GBit/s in order to handle the annual data volume of about 4,000 TByte. As the archive is a rolling archive also the evictions of data is required, e.g., Sentinel-2 products are never removed from the archive for Germany and earliest after 1 month global and earliest after 12 months for Europe. However, global Sentinel-2 Level 1C products are long-term archived in the German Satellite Data Archive (D-SDA) of EOC [1]. A historical data reload mechanism enables access to these products.



Sentinel-5P products data will become available next in CODE-DE.

4. ACCESSING

The portal (www.code-de.org) provides information on all available Copernicus Sentinel products and services together with links to available tools in this context; Sentinel-1, Sentinel-2 and Sentinel-3 products are searchable and accessible via catalogue client. It provides an enhanced user-friendly solution to discover, view, and download available Earth Observation (EO) data. For a user-friendly search the time, spatial, additional filters, e.g., polarization mode or cloud cover, are applied in combination with the various layers, e.g., Sentinel-2 Level 1C and overlays.

The catalog client features a full-resolution browsing displaying the Sentinel-2 products in full 10 m spatial resolution as illustrated in Figure 2 [3].

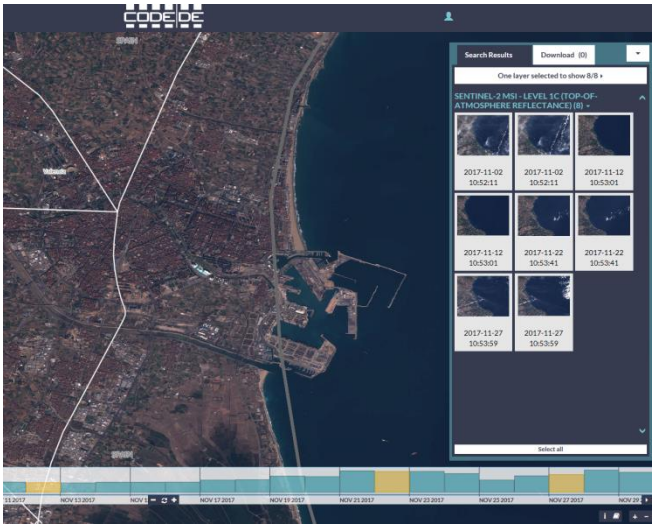


Figure 2. Illustration of Sentinel-2 Full-Resolution Browsing for Valencia, Spain

Based on the search results products can be selected and downloaded either as single products, via a metalink or URL listing.

Automatic access to data of the rolling archive is provided via download service. A subscription service to products based on the user needs, e.g., specified areas, is available to automatically transfer files of interest to a remote location (external cloud).

CODE-DE provides a powerful, yet simple download service, as defined by the OGC 13-043 best practice Download Service for Earth Observation Products using the

HTTPS (Hypertext Transfer Protocol Secure). The features range from simple directory browsing, direct download, single sign on, and quota handling limiting the parallel downloads per user and throttling the bandwidth depending on the groups the user belongs to. Single downloads are allowed up to 80 Mbytes/s, where 10 parallel downloads can surpass the available 5 Gbit/s internet connection. Of note, due to the latency of the internet connection to the remote users, normal download rate often is less than 10 MBytes/s, allowing these users to obtain a typical Sentinel-2 Level 1C scene of 600 Mbytes size in 60 seconds. Yet internal access is 30 times more performant!

The data is organized in directory structure in the form of `mission/year/month/day/<files>`.

5. PROCESSING

The portal (www.code-de.org) marketplace provides links to the available processors, processing chains and the processing environment user interface. Registered users process the data themselves with the help of various methods and can then download the generated value-added products. The processing environment is based on the Calvalus [5] software and uses an Apache Hadoop cluster as back-end. This processing environment allows the individual selection of algorithms and a spatial search for the remote sensing data to which the selected methodology is to be applied. The current status of the processing can be monitored in a processing tab.

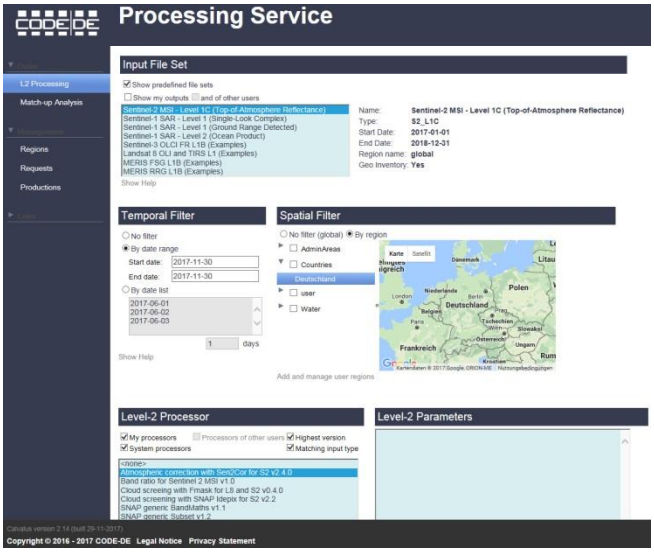


Figure 3. Illustration of Processing Sentinel-2 Products by Atmospheric Compensation

The methods currently available are excerpts from Sentinel tool boxes such as Sen2Cor for the atmospheric compensation of Sentinel-2 Level 1C products as illustrated in Figure 3.

Earth observation data processors in CODE-DE are considered software items that can be executed to transform Earth observation data into an output product. This determines the way results are produced, i.e. by applying processors to data that in turn may be further processed to higher level outputs repeatedly until the desired result is produced. Processors can be executed concurrently on different input products.

A processor installation package (processor bundle) is a set of files, e.g. .tar.gz or .zip files, containing the runtime software of one or several processors. An optional descriptor file (bundle descriptor) identifies data processors, their input product type, parameters, output product type, and bands of its output product. A processor installation package may contain a Docker image, i.e. a software item with a stack of layers (libraries, processor software) stored in a local .tar.gz file that can be loaded by the Docker daemon and instantiated as Docker container. Each call to the container applies the processor to one input product.

In addition to the Web GUI, CODE-DE offers to qualified users a possibility get access to a dedicated service host within the processing environment – a project VM. These users can deploy their own processors, submit processing requests directly to the processing cluster, to access the file system and processing results, to automate processing and set-up own services.

Project VMs are either dedicated VMs or receive a Docker container, run on the hosted processing environment. While a project VM is not necessarily a powerful machine itself, it can use the processing cluster as powerful computing facility. Project VMs can use a command line client to submit processing requests. They can implement workflows to automate the generation and submission of jobs for complex processing tasks and they may be used as host for services to be linked from the CODE-DE web portal to provide data or processing services to the external world. Project VMs have access to the storage with EO data and processing results.

In order to work with the CODE-DE processing environment the following steps need to be performed:

- login to the project VM
- submit a processing request
- install a processor bundle
- run the processor
- wait for the results
- access or download the results

Figure 4 depicts the role and the interfaces of a project VM.

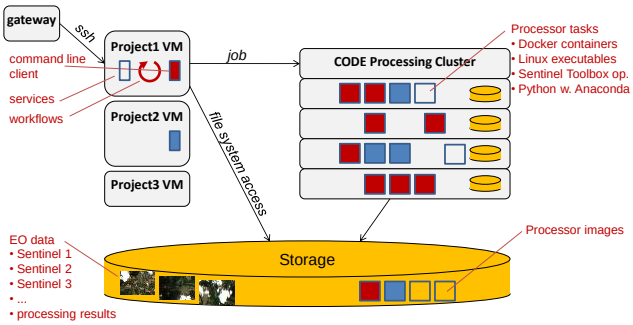


Figure 4. Project VMs with access to data and cluster computing resources.

In general the following use cases are possible:

- scripting for bulk processing, i.e. using Docker
- data driven processing

Besides the dedicated project VMs, the processing cluster is a shared facility. One queue per project on this cluster in combination with fair scheduling ensures that each project at least gets its share of the cluster computing and memory resources. A project gets more whenever not all projects are processing at a certain time, up to the full cluster capacity, with dynamic adaptation as new request are submitted.

Storage furthermore provides the location for software, in particular data processor packages, pre-installed ones and those provided by the project. Several conventions are available for processor packages, among them Docker images, Linux executables (for CentOS), Sentinel Toolbox operators (as jar files), and processors implemented in Python using Anaconda as runtime environment. The storage is also the location for input data access and value added product output.

Any registered user hosted in his remote home environment or in an external cloud, can also invoke processors to work directly on the CODE-DE data offerings, by submitting their processing requests using the code-de-tools [6].

6. VALUE ADDED PRODUCTS

In addition to the original Sentinel Data, CODE-DE processes and provides a set of value added products as a basis for information extraction and demonstrating the processing capabilities. Currently the following datasets are offered for download:

- Maritime Products
- Temporal Feature Extraction
- Cloud-filtered mosaic of Germany

The CODE-DE marketplace enables users to publish and visualize other geo-service data offerings; currently it includes the RapidEye Science Archive (RESA) and MODIS Germany mosaics, and SRTM X-Band DEM datasets.

7. WHAT'S NEXT?

CODE-DE as a cloud based platform for Sentinel data access and processing is foreseen to enter a second phase in late 2019. Continuity of operations for at least four more years with some upgrades is envisaged. This is with a particular focus to assist users in national public institutions with access to analysis ready Sentinel data and Copernicus service products, as well as with on-boarding their processing algorithms. It is envisaged that this CODE-DE transition will make a best possible use of synergies with the European level Copernicus data information and access services (DIAS).

8. CONCLUSIONS

The methodology, architecture, and various functionalities of CODE-DE (Copernicus Data and Exploitation Platform – Deutschland, www.code-de.org) are presented and analyzed to obtain a high-quality system. The focus is on the major challenges of user-friendly online data access and high-performance big data processing. In the next years European initiatives such as Copernicus Data and Information Access Services Operations (DIAS) will complement the national initiatives.

9. REFERENCES

[1] Kiemle, S., K. Molch, S. Schropp, N. Weiland, and E. Mikusch, *Big data management in Earth Observation: the German Satellite Data Archive at DLR*, *Proceedings of the 2014 Conference on Big Data from Space*, Frascati, Italy, pp. 46-49, 2014.

[2] Reck, C., G. Campuzano, K. Dengler, T. Heinen, and M. Winkler, *German Copernicus Data Access and Exploitation Platform*, *Proceedings of 2016 Conference on Big Data from Space*, Auditorio de Tenerife, Spain, pp. 1-4, 2016.

[3] Storch, Tobias and Reck, Christoph and Holzwarth, Stefanie and Keuck, Vanessa (2018) *CODE-DE – The Germany Operational Environment for Accessing and Processing Copernicus Sentinel Products.* In: 2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IGARSS 2018, Valencia, Spain

[4] Storch, T., M. Habermeyer, S. Eberle, H. Mühle, and R. Müller, *Towards a Critical Design of an Operational Ground Segment for an Earth Observation Mission*, *Journal of Applied Remote Sensing*, 7(1), pp. 1-12, 2013.

[5] http://www.brockmann-consult.de/calvalus/, 2018

[6] https://github.com/dlr-eoc/code-de-tools/ released with CODE-DE version 2.1, 2018

QUERY PLANET - DEMOCRATISING INSIGHTS FROM EO BIG DATA

Grega Milcinski[1], Devis Peressutti[1], Matej Batic[1], Anze Zupanc[1], Matej Aleksandrov[1], Matic Lubej[1], Drew Bollinger[2], Olaf Veerman[2], Pierre-Philippe Mathieu[3]
(1) Sinergise, Ljubljana, Slovenia
(2) Development Seed, Lisbon, Portugal
(3) European Space Agency, Frascati, Italy

ABSTRACT

Recent implementations of machine learning tools running on large scale data have been enabled by vast improvements in computing capability and parallelization, theoretical advances in machine learning algorithms (in particular in the field of deep learning), and most importantly, an explosion of available data – the fuel of any machine learning engine. Copernicus' unparalleled volume and quality of earth observation data is contributing strongly to the latter.

The earth observation (EO) field, however, has not yet seen significant uptake of machine learning methods, mainly due to a lack of annotated data, a lack of tools to handle spatio-temporal EO data, and the inherent infrastructure complexity of handling vast amounts of EO data.

In this paper, we outline how the Query Planet project seeks to address these existing shortcomings, providing open-source tools to facilitate development of machine learning models to exploit EO big data.

1. INTRODUCTION

The Sentinel missions are collecting more data than any existing earth observation programme, opening up unprecedented opportunities for understanding and exploiting EO data. In particular, recent advances in machine learning (ML) play a pivotal role in analysing and processing large amounts of data, allowing to extract actionable information from complex spatio-temporal data.

The following challenges, however, are currently hindering the uptake of ML methods in EO:

- Many ML efforts in EO have been limited to a very specific geographic area. In order to scale these results to a global scale, tools and infrastructures capable of handling EO data in an efficient manner are necessary;
- Most available ML frameworks require special approaches to accept EO data. The most challenging aspect is handling time-series data, which is critical in many EO applications, ranging from vegetation classification, to change detection.
- There are only a few high-quality sources of annotated data currently available for training supervised ML algorithms on satellite imagery. Accurate annotated data are critical to the advancement of ML research and development of new EO applications.

The Query Planet project, running within ESA's Phi-lab, seeks to tackle each of the above challenges, by employing cloud-based technologies to seamlessly access EO data, and by developing open-source tools to process and annotate spatio-temporal EO imagery. All software tools and datasets developed under the project are released under open-source licenses, providing a unique opportunity for research and development of novel machine learning applications. The following Sections provide details on the solutions developed to tackle the above-mentioned challenges.

2. ACCESSING EO DATA

The first challenge on how to exploit the wealth of information contained in the EO big data is of technical nature – how to provide an access to the data in a manageable way. This was effectively addressed by Sentinel Hub [1], which provides streamlined access to many satellite missions using the Amazon Web Services infrastructure. However, new challenges soon surface - how to analyze a vast volume of dense data - there are simply not enough eyes in the world to check every image acquired. Modeling of time-series EO data has received a lot of attention recently as it can significantly contribute to traditional EO interpretation methodologies. It is not just that there are greater amounts of data available, or that we need to model periodic (yearly) data sequencing - the problem is much more rooted. Since EO data reveals inherent yearly cycling, we need models that can adapt to these cycles, considering that at specific points in time the cycle is broken due to changes on the ground - converting arable land into urban area, flooding, etc. Therefore, tools to efficiently process and analyse spatio-temporal data are required to extract meaningful information.

3. BRIDGING THE GAP BETWEEN EO AND MACHINE LEARNING

With the availability of massive volumes of data, machine learning (ML) has become an important tool for analysis of EO data. ML models used in EO to date range from random forest algorithms, to more complex convolutional neural networks. These ML models need to cope with the peculiarities of satellite imagery - clouds, atmospheric effects and inaccurate geolocation are distorting the data, missing or cloudy scenes create gaps, etc. These artefacts make it

doi:10.2760/848593

difficult to directly use on the data well-known ML frameworks such as TensorFlow, MXNet, etc. In addition, lack of ground truth for training and validation of supervised ML models is a major challenge preventing from efficient use of ML tools.

The Query Planet project seeks to address these challenges by developing the following open-source tools:

- eo-learn [2], a Python based package acting as a bridge between EO data and existing ML and computer vision tools;
- classification application and label maker enhancements [3];
- repository for ground truth data.

Importantly, the above-mentioned elements are integrated and demonstrated on a pair of use-cases - land cover classification and global water monitoring. These use-cases are available in an open-source manner, including training sets and ML models, which will make them extremely convenient for anyone to use, modify, and extend.

3.1. eo-learn

eo-learn is a collection of open source Python packages that have been developed to seamlessly access and process spatio-temporal image sequences acquired by any satellite fleet in a timely and automatic manner. eo-learn is easy to use, its design is modular, and encourages collaboration – sharing and reusing of specific tasks in a typical EO-value-extraction workflows, such as cloud masking, image co-registration, feature extraction, classification, etc. Everyone is free to use any of the available tasks and is encouraged to improve them, develop new ones and share them with the rest of the community.

eo-learn makes extraction of valuable information from satellite imagery as easy as defining a sequence of operations to be performed. Figure 1 below illustrates a processing chain that executes automatic classification of land cover in a user specified region of interest.

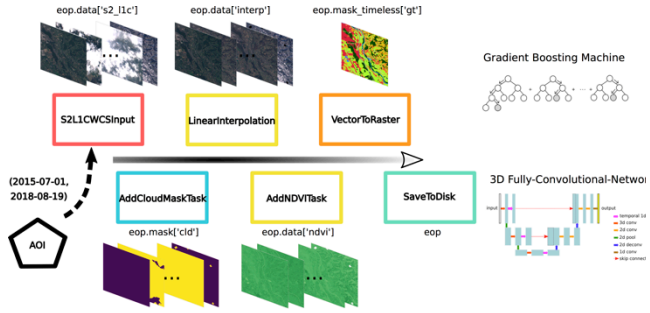


Fig 1. Example workflow for land cover classification

The eo-learn library acts as a bridge between the EO/remote sensing field and the Python ecosystem for data science and machine learning. The library is written in Python and uses NumPy arrays and Shapely polygons to store and handle remote sensing data. Its aim is to make entry easier for non-experts to the field of remote sensing on one hand, and bring state-of-the-art tools for computer vision, machine learning, and deep learning existing in Python ecosystem to remote sensing experts.

The design of the eo-learn library follows the dataflow programming paradigm and consists of three building blocks:

- EOPatch - common data-object for spatio-temporal EO and non-EO data, and their derivatives; it contains multi-temporal remotely sensed data of a single patch (area) of Earth's surface typically defined by a bounding box in specific coordinate reference system, both in raster and vector format. The size and shape of the EOPatch can vary based on specific needs and available resources (large patches will require more memory). The EOPatch object can also be used as a placeholder for all quantities, either derived from the satellite imagery or from some other external source, for example biophysical indices, ground truth reference data, weather data, etc. EOPatch is completely sensor-agnostic, meaning that imagery from different sensors (satellites) or sensor types (optical, synthetic-aperture radar, etc.) can be added to an EOPatch.
- EOTask - a single, well-defined operation being performed on input EOPatch(es) and which returns a modified EOPatch. EOTasks are the heart of the eo-learn library. They define in what way the available satellite imagery can be manipulated in order to extract valuable information. Typical users will most often be interested in what kind of tasks are already implemented but can also write custom EOTasks, as shown in Figure 2, if their desired functionality doesn't yet exist.
- EOWorkflow is a collection of EOTasks that together represent an EO-value-adding-processing chain or EO-value-extraction pipeline by chaining or connecting a sequence of EOTasks. The EOWorkflow takes care that the EOTasks are executed in the correct order and with correct parameters. EOWorkflow is executed on a single EOPatch at a time, but the same EOWorkflow can be executed on multiple parallel processes. Under the hood the EOWorkflow builds a directed acyclic graph. There is no limitation on the number of nodes (EOTasks with inputs) or the graph topology. The EOWorkflow first names the input tasks that persist over executions, determines the ordering of the tasks, executes the task in that order, and finally returns the results of tasks which represent terminal nodes of the graph. Reports and logs on execution are automatically provided to ease monitoring and debugging.

There are several existing sub-packages, covering common EO analysis steps:

- eo-learn-core, the main sub-package which implements basic building blocks (EOPatch, EOTask and EOWorkflow) and commonly used functionalities.

doi:10.2760/848593

- eo-learn-coregistration, dealing with image co-registration to correct geolocation errors.
- eo-learn-features is a collection of utilities for extracting data properties and feature manipulation.
- eo-learn-geometry is used for geometric transformation and conversion between vector and raster data.
- eo-learn-io, input/output sub-package that deals with obtaining data from various data source services or saving and loading data locally. It provides seamless access to global archive of Sentinel-1 GRD, Sentinel-2 (L1C and L2A), Sentinel-3 OLCI, Sentinel-5P, Landsat-8, MODIS, Envisat MERIS and ESA archive of Landsat-5 and -7. Open-source libraries sat-utils[4] are used to work with locally stored or remotely accessible GeoTiff files and OpenStreetMap data.
- eo-learn-mask, used for masking of data and calculation of cloud masks.
- eo-learn-ml-tools - set of tools that can be used before or after the machine learning process.

The eo-learn package can be easily integrated with other Python packages, e.g. within an EOTask node. Jupyter Notebook is used as IDE [5].

```
class FooTask(EOTask):
    def __init__(self, foo_param):
        self.foo_param = foo_param

    def execute(self, eopatch, *, patch_specific_param):
        # do what foo does on input eopatch and return it
        return eopatch
```

Fig 2. Code snippet showing how developers can extend the package

3.2. Ground-truth labels

The lack of ground truth data required for supervised ML training is addressed in two ways: by identifying openly available regional and global datasets of proper quality, which can be used as an input, and by creating a classification app, which can be used by experts or crowds to collect annotations. OpenStreetMap, SpaceNet, Corine land-cover, various official register datasets (buildings, roads, farm parcels) and similar can be efficiently used to create training data. Label maker [3] is used to create training data patches from these sources and package them as NumPy arrays for easy integration with machine learning libraries.

The classification app is a web-based tool, that was designed to allow users to easily set-up a new "campaign" and start collecting the data. It was first put in place for development of s2Cloudless, where it was used to classify clouds[6]. The tool requires authentication of the user, so that it is possible to associate individual records with a specific user (and flag all user's entries if low quality input is detected). Users are then presented a satellite image (e.g. Sentinel-2 or other data-source, depending on the use-case) and asked to annotate an area of a complete randomly defined patch (e.g. 64x64 px or 512x512 px), as shown in Figure 3. Completeness of the labelling is required in cases where one wants to avoid vaguely defined data - e.g. border of the clouds in the cloud examples. In other cases, where we are looking only for specific elements (e.g. built-up areas), completeness is not enforced. Users are able to explore the area around the dedicated tile, and check various band combinations (e.g. NDVI, false colour, NDWI, custom option). The tool is configurable to address various use-cases (label options, area limitations, patch size, satellite imagery sources, supporting datasets). The open-source nature of the tool allows further customization. Classified data can be exported using a dedicated API (integrated with eo-learn) or exported in standard formats (e.g. SHP, GeoTiff, GeoJSON).

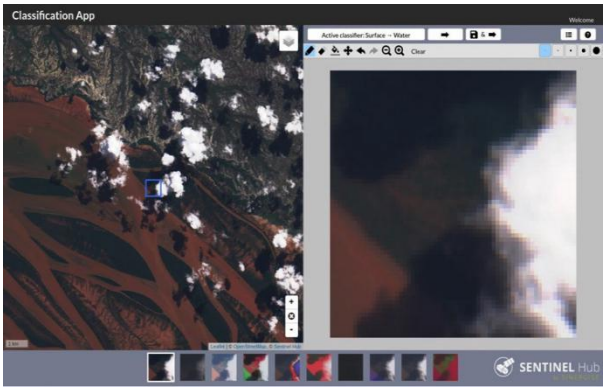


Fig 3. Snapshot of the classification application user interface. On the left, users can visualise an imaging source over a given area-of-interest, while on the right they can annotate the selected patch according to the labelling guidelines

4. USE CASES

Two use-cases can already be demonstrated in a start-to-end fashion, making it easier for other developers to take on existing work and modify it for their own case.

4.1. Automatic land cover classification

Land cover classification is using a combination of inputs for training data - land-use data from several European countries, where these are regularly updated for the purpose of Common Agriculture Policy, Corine land cover and OSM data in other parts of the world, already existing crowd-sourced data as well as newly collected data using the classification tool developed within the project. Detailed description of the land cover classification workflow is available in a parallel paper "Multi-Temporal Land Cover Classification Using Sentinel Data and the eo-learn Open-Source Python Project" [7]. An example of classification output is shown in Figure 4.

doi:10.2760/848593

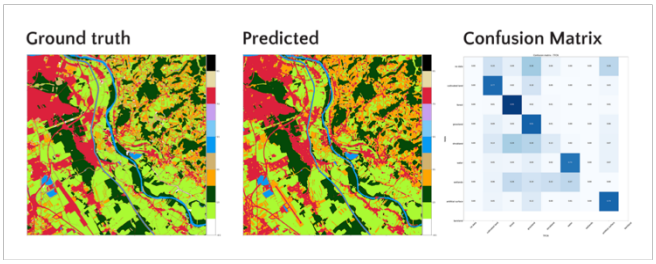


Fig. 4 Standard outputs of automatic land cover classification

4.2. Global monitoring of water in lakes and reservoirs

Being enlightened by JRC's Global Surface Water project [8] we have set out to build a service, which does not only show historic data but is also up-to-date. Copernicus Sentinel mission, with its global coverage and short revisit time, combined with an efficient use of cloud infrastructure resources makes it feasible to do a global scale project with limited resources. The Blue Dot Water Observatory [9] is a showcase of this approach, both as an operational service as well as a set of open-source tools, which can be modified to fit specific needs. The dashboard of the Blue Dot Observatory is shown in Figure 5.

The Blue Dot Water Observatory provides reliable and timely information about surface water levels of water bodies across the globe. All observations are provided and can be explored interactively via the Water Observatory Dashboard or via RESTful API. The key benefit of the service is the accumulation of current and historic surface water level data in one place, presented in a clear and interactive way, free of charge. The Water Observatory provides a valuable service to local authorities, governmental agencies, natural parks and reserves, agricultural ministries and agencies, stakeholders in food and energy production, and citizens alike.

With this service, we are also demonstrating how global monitoring of the environment using EO data can be done efficiently and orders of magnitude cheaper than before, if done in an intelligent way. We are sharing as an open-source the code, water detection algorithms as well as details on how to put service like this in production. We hope this will inspire others to build on top of it and develop similar services for other use cases.

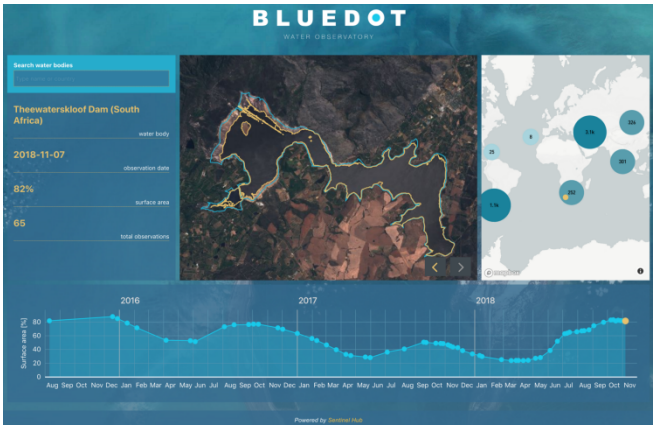


Fig. 5 - Blue Dot Water Observatory displaying the Theewaterskloof reservoir near Cape Town, South Africa

5. CONCLUSION

Volume, availability and quality of open earth observation data have reached the level where machine learning methods are not just a meaningful option, but a necessity. However, where there are several well established ML options available for imagery in general, not many of these are supporting EO data and their complexity.

Query Planet is addressing this challenge by introducing eo-learn, to bridge the gap between EO and standard ML tools, by developing classification tools to create labels required for supervised learning, and by publishing actual start-to-end use-cases, which make it easier for researchers to start with the process and customize it for their needs.

This announcement, part way through the project, is meant to call for cooperation with other researchers in the field, so that we can produce results fitting their requirements if possible.

6. REFERENCES

[1] Milcinski et al, Integration of Web World Wind and Sentinel Hub – a Global 4D Big Data Exploration and Collaboration Platform, Proceedings of the 2017 conference on Big Data from Space

[2] https://github.com/sentinel-hub/eo-learn

[3] https://github.com/developmentseed/label-maker

[4] https://github.com/sat-utils

[5] https://github.com/sentinel-hub/example-notebooks

[6] https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13

[7] M. Lubej et al, Multi-Temporal Land Cover Classification Using Sentinel Data and the eo-learn Open-Source Python Project, Big Data from Space 2019

[8] Jean-Francois Pekel, Andrew Cottam, Noel Gorelick, Alan S. Belward, High-resolution mapping of global surface water and its long-term changes. Nature 540, 418-422 (2016). (doi:10.1038/nature20584)

[9] BlueDot Water Observatory, http://water.blue-dot-observatory.com

EXPLORATION OF NATURAL LANGUAGE PROCESSING TECHNIQUES TO LINK SCIENTIFIC PUBLICATIONS WITH OBSERVATIONAL DATA

Giannakis, Omiros [(1)], Akylas, Athanassios [(1)], Ruiz, Angel [(1)], Demiros, Iason [(2)], Antonopoulos, Vassilios [(2)], Voutas, Michalis [(2)], De Marchi, Guido [(3)], Arviset, Christophe [(4)]

[1]National Observatory of Athens, IAASARS, Greece
[2]QUALIA, Greece
[3]European Space Research and Technology Centre, European Space Agency, the Netherlands
[4]European Space Astronomy Center, European Space Agency, Spain

ABSTRACT

We present a study to search and classify semantic relations among the entities in the Rosetta mission papers, in order to link them to the existing Rosetta ESA Science Archive database for further processing. The type of text and the variety of the linguistic phenomena lead us to apply techniques from the domains of natural language processing and machine learning for addressing the problem. While similar activities are already carried out manually as part of the science operations of some of the ESA science missions, the ultimate goal is to devise a semi-automated system able to apply activity to all the science missions, to facilitate the work of the missions librarians and to allow effective cross-comparisons. Our dataset consists of 40 papers covering the following missions: HST, MEx, PROBA-2, ROSETTA.

Index Terms— Planetary Science Archive (PSA), Natural Language Processing, Machine Learning, Information Retrieval, Semantic Analysis

1. INTRODUCTION

Of major interest to the ESA Science Archives is to establish precise links between the scientific publications and the specific science instruments and experiments carried by the space science missions, at a high level of detail, so as to highlight which instruments, data products and modes of operation lead to scientific papers. However, since the number of publications is large and growing, manual annotation and link creation by human experts cannot scale well. Moreover, although the move towards formal citation of data is growing, there is a lot of work to be done until data holders provide mechanisms and guidance for scholars to cite datasets.

A way of approaching this problem is processing the cross-references and the text content of scientific publications through supervised machine learning techniques. This method allows the classification of publications on different categories based on, for example, the data products of a particular mission, using a training sample of manually classified publications. We are currently developing a test-case application of this technique for the Planck mission. However, this approach can only obtain a gross classification. For a finer linking of publications and mission data, other methods like Natural Language Processing (NLP) are needed.

NLP is a key technology that enables the scientists to access the scientific information. Extracting information from scientific papers can contribute to the development of rich scientific knowledge bases that can support intelligent knowledge access and decision making.

In this project, our objectives are to:

- Find potential links between publications and the science instruments and experiments carried by the space science missions (observational data).
- Monitor the scientific productivity of a space mission and how it evolves over time.
- Support the decision making regarding future missions: role of detectors, areas of the sky studied, wavelength bands of major interest.
- Identify which instruments and modes of operation are more effective at leading to scientific papers.
- Allow effective cross-comparisons between different missions and agencies

To this end, we have conducted a small-scale experiment on 40 papers equally distributed among four ESA missions (HST, MEx, PROBA-2, ROSETTA). The purpose of the pilot was to inspect the data and to understand the linguistic phenomena that occur in the texts, in order to devise a methodological framework that can be applied and meet the project's objectives. At a later stage and in the main phase of the project, more missions will be added, and the total set of publications of each mission will be processed.

2. DATA INSPECTION

In the initial phase of the pilot we conducted an analysis of the characteristics of the text in order to decide on the type of NLP that we have to apply. Our analysis has demonstrated that there are no direct links from the papers' time expressions to the Observation ID of the Planetary Science Archive. Moreover, general references to data, instruments,

doi:10.2760/848593

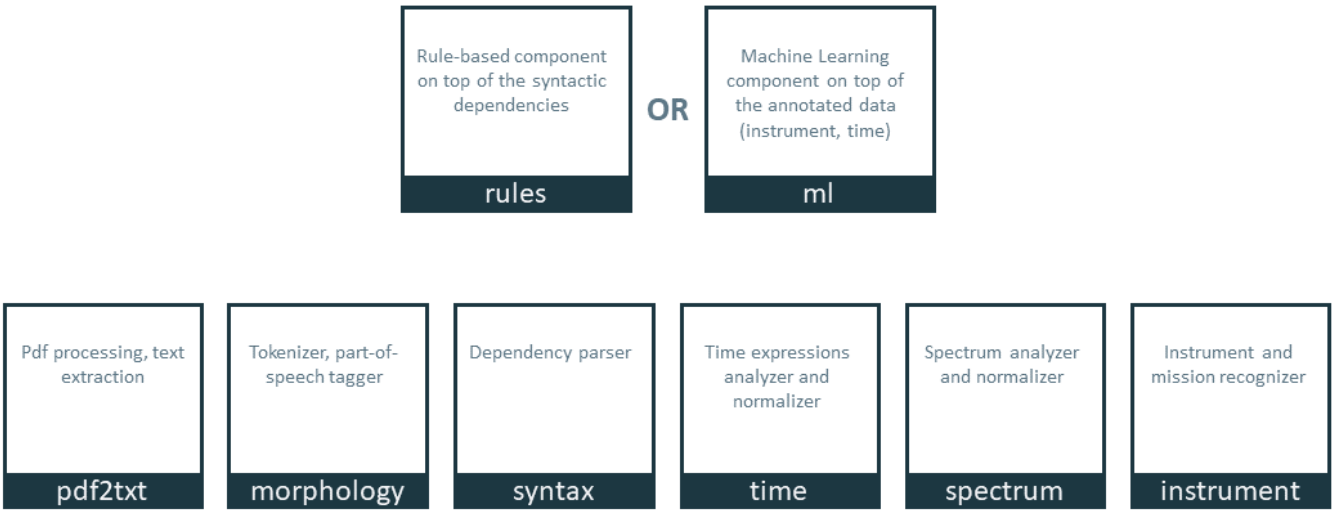


Figure 1: The components of the proposed solution

measurements, missions and experiments are intertwined in the papers. Specific references are a minority. The papers are dense in named entities, time expressions and units of measurement, and they contain complex formalisms. Sometimes there exist dedicated chapters in the papers where data and observations are described. However, in the majority of the papers both data and observations can appear anywhere in the paper. Regarding ESA there exist on-topic and off-topic papers: sometimes ESA is at the epicenter, sometimes there is just one mention possibly regarding a single comparison. Instrument acronyms, instrument names and mission names seem systematic throughout papers. Finally, matrices and images contain information regarding the observations, the measurements and the instruments of the missions.

The type of text and the variety of the linguistic phenomena that we have found in the papers lead us to adopt techniques from the domains of NLP and machine learning in order to address the project's objective. We plan to apply morphological, syntactic and semantic processing to recognize the semantic relations between the elements of the papers, such as missions, instruments, temporal expressions and observation IDs.

3. METHODOLOGY

We first search for the names of the mission instruments in the text that we extract from the papers' pdf files. We then locate the time expressions and we recognize the dependency structures which constitute the directed grammatical relations that hold between the words of the sentence. Once we know the grammatical functions and the time expressions, then we can create a rule-based or statistical algorithm that will link the constituents of the sentence to the ESA database. Figure 1 depicts the components of the methodology that we propose. The description of each component follows.

3.1. Pdf2txt

There exist many tools that convert pdf to text. Since the older papers are in image format while the recent papers are in textual format, we must be able to deal both with images and text. Also, the selected tool should be able to correctly convert tables and hyphens. Our experiments showed that no tool outperforms the others in all cases, so the final solution will combine and select the best output of various tools in an ensemble approach. In the cases where text is not already embedded in the pdf, we will need to use OCR to extract the text.

3.2. Morphology

Tokenization is the process of breaking up the given text into units called tokens. The tokens may be words or numbers or punctuation marks.

The process of classifying words into their parts of speech and labeling them accordingly is known as part-of-speech tagging or POS tagging, or simply tagging. Parts of speech are also known as word classes or lexical categories. The collection of tags used for a particular task is known as a tag set. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. POS tagging serves as an input to more complex linguistic analysis such as chunking and parsing.

doi:10.2760/848593

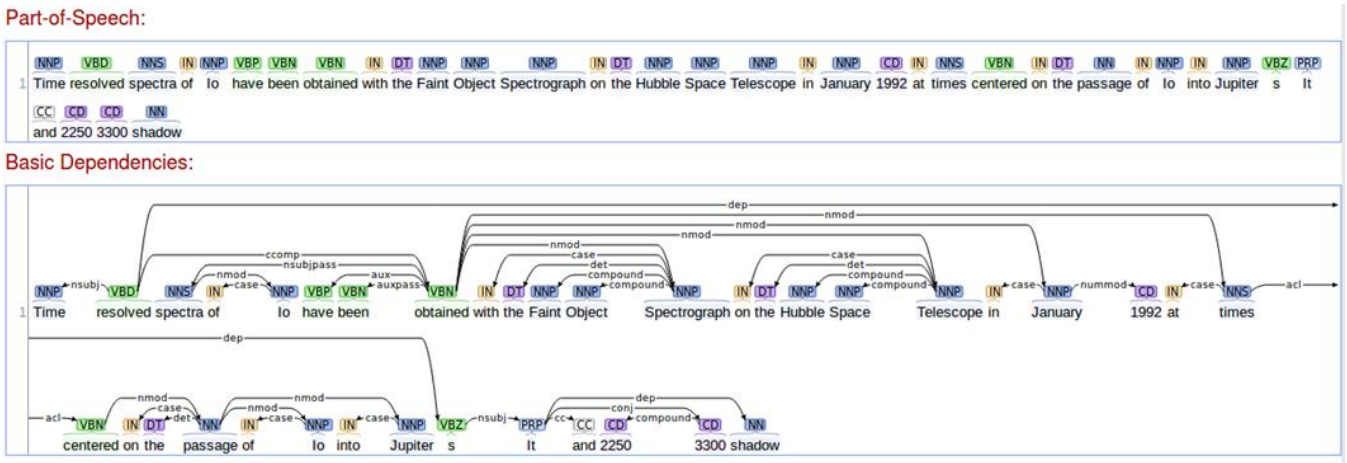


Figure 2: An example of dependencies among words of a sentence that contains an instrument, a mission and a temporal expression

3.3. Syntax

In a dependency grammar the syntactic structure of a sentence is described solely in terms of the words (or lemmas) in a sentence and the associated set of directed binary grammatical relations that hold among the words. The most widely used syntactic structure is the parse tree which can be generated using various parsing algorithms. Parse trees play a critical role in the semantic analysis stage.

The traditional linguistic notion of grammatical relation provides the basis for the binary relations that comprise the dependency structures. The arguments to these relations consist of a head and a dependent. In dependency-based approaches, the head-dependent relationship is made explicit by directly linking heads to the words that are immediately dependent on them. In addition to specifying the head-dependent pairs, dependency grammars allow us to further classify the kinds of grammatical relations in terms of the role that the dependent plays with respect to its head. Familiar notions such as subject, direct object and indirect object are among the kind of relations that we analyze. In our pilot we have used the Stanford Parser which is considered as one of the best parsers that exist in the language processing market [1].

3.4. Time

The task is to automatically detect, bracket and normalize relevant time expressions mentioned in the papers. Detection refers to the systems' capability to recognize time expressions within an input text. Bracketing concerns systems' capability to correctly determine the extension of a detected time expression.

Temporal normalization (or resolution, grounding) is the task of mapping from a textual phrase describing a potentially complex time, date, or duration to a context-independent, easy-to-use temporal representation. For example, possibly complex phrases such as *the week before last* are often more useful in their normalized form – e.g., *August 1 - August 7*.

This is in its own right a difficult problem. To illustrate, *the past* and *3 months* each have independent and very different interpretations, yet *the past 3 months* is completely different from either. In addition, temporal expressions are often ambiguous, either in the syntactic structure (e.g. *[last Friday] the 13th* vs. *last [Friday the 13th]*) or its pragmatic content (e.g., *Friday* could be either the previous or next Friday). In our pilot we have experimented with Duckling which parses temporal expressions described in many ways [2].

3.5. Spectrum

Spectral entity recognition is a subtask of the broader named-entity recognition task that seeks to locate and classify named entities into predefined categories such as persons, locations, organizations, companies, quantities, time expressions, monetary values, etc.

Temperature and energy, angular distances, spectral lines and ions, are types of named entities that need to be annotated in order to train a named entity recognition engine that will be able to accurately detect these types in the papers. Once we have adequate training data, we can use a named entity recognition system in order to build our recognizer [3].

3.6. Instrument

Names and acronyms of missions and instruments appear in the papers in their standard form, which means that a name search will be adequate to detect them. In most papers the first mention uses the full name while subsequent mentions use the acronym.

3.7. Linking

Building semantic representations from text corpora is the first step to perform more complex tasks such as text entailment, enrichment of knowledge bases, or question

doi:10.2760/848593

answering. The dependency parsers are commonly used as a semantic representation in natural language understanding and inference systems [4].

Our task in the project is to link the instruments to the measurements/observations and to the time expressions in the sentences where they co-occur. Once we find the relation between instrument, observation and time, then we will be able to link the specific time expression to the ObsIDs in the ESA database. Two general methodologies that we can apply in order to link the papers' observations to the ESA database are semantic parsing and classification.

3.7.1. Semantic parsing

Semantic refers to meaning, and parsing means resolving a sentence into its component parts. As such, semantic parsing refers to the task of mapping natural language text to formal representations or abstractions of its meaning. . There are several models built using the results of the syntactic analysis, which are usually referred to as shallow semantic processing: semantic role labeling [5], conceptual dependencies, logical forms, etc. The main reason computational systems use semantic roles is to act as a shallow meaning representation that can let us make simple inferences that aren't possible from simple representations such as the bag-of-words, or even from the parse tree [6].

Broadly speaking, we can classify the attempts to add semantic knowledge to a parser in two sets: using large semantic repositories, such as WordNet or similar ontologies, and approaches that use information automatically acquired from corpora. One method that has showed good results is to semantically enrich the input by substituting content words with their semantic classes.

In Figure 2, which depicts the result of the dependency parser, we use the semantic class *observation* in order to represent all the content words such as *obtained* that are identified by the parser and that correspond to the meaning of a measurement/observation within the papers. We can see the arrows (dependencies) between the named entity of the instrument Faint Object Spectrograph of the Hubble Space Telescope and the verb *obtained*, as well as between the date January 1992 and the verb *obtained*. Thus, we can link the instrument to the date and the observation and then we can link the canonical form of the date to the database which also contains the canonical form of the dates of the observations.

3.7.2. Classification

Another way to address the problem is as a classification task where, given a sentence that contains mentions of instruments (recognized by the given seed list), observations (semantic class) and time expressions (temporal recognizer), the system has to decide whether they are linked or not. In the last years, the most successful methods model the structural information of the sentence into a character, word or sentence embedding, by using convolutional neural networks with various configurations [7]. Embedding layers take a sequence of words as an input and produce a sequence of corresponding vectors as an output. In order to learn the embeddings we can use a pre-trained set of embeddings and jointly fine-tune it for our particular dataset.

Neural network text classifiers typically follow the same architecture: embedding, deep representation, fully-connected part [8]. The most straightforward and reliable architecture is a multilayer fully connected text classifier applied to the hidden state of a recurrent network. An alternative way to train a deep text classifier is to use convolutional networks.

4. CONCLUSION

We conducted a small-scale experiment on 40 papers equally distributed among four ESA missions. The purpose of the pilot was to inspect the data and to understand the linguistic phenomena that occur in the texts, in order to identify a methodological framework that can be applied and meet the project's requirements, as described in the previous sections.

As a next step, we will proceed with the full-scale project of one mission, i.e. Rosetta. Since the techniques described above are data-driven, we will need the full dataset of the Rosetta papers in order to design the modules of the system and to fine-tune and run the algorithms. Machine learning systems are data intensive and the corresponding algorithms will learn on the full dataset in order to achieve optimal accuracy. Once we successfully tackle the problem for Rosetta, then we can move to the other ESA missions by replicating and fine-tuning the methodology, and learning from new data, as well as our experience from working with the Rosetta material.

5. REFERENCES

[1] Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser Using Neural Networks. In Proceedings of EMNLP 2014.
[2] https://duckling.wit.ai/
[3] http://services.gate.ac.uk/annie/
[4] Statement presented by Chris Manning at the SEM 2013 Panel on Language Understanding, http://nlpers.blogspot.com/2013/07/the-sem-2013-panelon-language.html.
[5] Foland Jr, W. R. and Martin, J. H. (2015). Dependency-based semantic role labeling using convolutional neural networks. In SEM 2015), pp. 279–289.
[6] Fillmore, C. J. and Baker, C. F. (2009). A frames approach to semantic analysis. In Heine, B. and Narrog, H. (Eds.), The Oxford Handbook of Linguistic Analysis, pp. 313–340. Oxford University Press.
[7] Dos Santos, Cícero Nogueira, and Maira Gatti. "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts." COLING. 2014.
[8] https://blog.statsbot.co/text-classifier-algorithms-in-machine-learning-acc115293278

doi:10.2760/848593

SATELLITE REMOTE SENSING OF OZONE USING A FULL-PHYSICS INVERSE LEARNING MACHINE

Jian Xu, Klaus-Peter Heue, Diego G. Loyola, Dmitry S. Efremenko

German Aerospace Center (DLR), Remote Sensing Technology Institute (IMF)

ABSTRACT

The new generation of environmental satellites with increased spatial and spectral resolutions imposes critical challenges for the processing of the Big Data. This work employs the newly-developed full-physics inverse learning machine (FP-ILM) to estimate vertical distributions of ozone from Global Ozone Monitoring Experiment – 2 (GOME-2) measurements and analyzed its performance. The obtained ozone profile shapes are further used to derive the vertical column density of ozone. The main advantage of FP-ILM is that, unlike classical retrieval algorithms, the ozone profile retrieval is formulated as a classification problem, producing a significant speed-up and reliable accuracy. The time-consuming radiative transfer computations and neural network training are performed off-line and do not introduce additional performance bottlenecks in the whole processing chain. Therefore FP-ILMs are suitable for processing remote sensing Big Data.

Index Terms— Atmospheric remote sensing, ozone, FP-ILM, machine learning

1. INTRODUCTION

Investigation of vertical distributions of atmospheric ozone offers useful information on photochemical and dynamical processes as well as the transport of pollutants. For several decades, remote sensing instruments mounted on various platforms (e.g., satellite, balloon, aircraft) have been utilized to measure vertically integrated column density and the concentration profile of ozone. In particular, satellite remote sensing of ozone measuring the UV radiation has been rapidly reaching maturity. Past European sensors dedicated to ozone monitoring include GOME (Global Ozone Monitoring Experiment) on the ERS-2 satellite, SCIAMACHY (SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY) on the Envisat satellite, OMI (Ozone Monitoring Instrument) on the NASA's Aura satellite, and the GOME-2 spectrometers aboard MetOp series of satellites. In October, 2017, TROPOMI (TROPOspheric Monitoring Instrument) mounted on the Copernicus Sentinel-5 Precursor (S5P) satellite was launched, which focuses on spatial and temporal variations of tropospheric ozone and other trace gases related to air quality and climate change, as well as clouds and aerosols. Last but not least, the European satellites for atmospheric monitoring will be followed by the future Sentinel-4 and Sentinel-5 missions.

These instruments have been expected to provide accurate and timely observations of key atmospheric species, for services on air quality, climate forcing, UV and the ozone layer. The daily global observations will be used for improving air quality forecasts as well as for monitoring the concentrations of atmospheric constituents. Trend monitoring is very important to verify that policies implemented to control emissions to the atmosphere are effective. In particular, with its global coverage and improved spatial resolution, TROPOMI/S5P will open a new era of challenges regarding big data and the processing capability.

Although ozone profiles can also be retrieved from these aforementioned remote sensing instruments, current algorithms are time-consuming and therefore may not be suitable for near-real-time applications. As these algorithms often use means of regularization methods to tackle the ill-posedness and some a priori knowledge (such as climatology) about the atmospheric state variables to impose the constraint. However, convergence issues may arise when the constraint strength is not chosen wisely [9], or when the constraint shape cannot represent the solution. Precise knowledge of an ozone profile shape is essential for accurate determination of ozone vertical column density through the atmosphere. Most ultraviolet (UV) based total ozone retrieval algorithms use an external ozone climatology that may differ from the actual vertical distribution of ozone.

In atmospheric remote sensing, we often require a radiative transfer modeling that is sometimes a bit too complicated and has to be precisely carried out. Furthermore, multiple calls to a forward model can be computationally prohibitive. Machine learning turns out to be a promising solution to these issues. A new retrieval framework based on machine learning techniques for estimating ozone profile shape was proposed in [10], and the comparison of retrieved ozone profiles from GOME-2 data between our algorithm and the optimal estimation method [7] reaches an encouraging agreement.

This paper extends the ability of the FP-ILM retrieval to use ozone profile shapes on total ozone retrieval for GOME-2 measurements.

doi:10.2760/848593

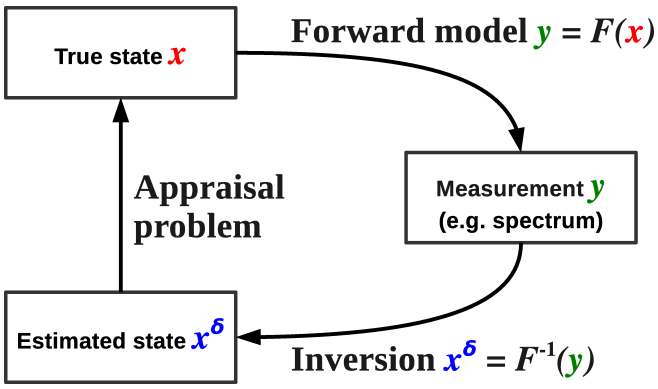


Fig. 1. Forward and inverse problems.

2. METHODOLOGY

In atmospheric remote sensing, the inverse problem is the process of deriving geophysical quantities from a given set of measurements. It is often referred to as a retrieval problem or simply *inversion*.

In the classical approach, the inverse problem is solved by reducing it to an exercise in optimization. The main idea behind this method is to minimize the cost function based on the residual between simulated and observed data by finding an appropriate state vector. A non-linear inverse problem is solved iteratively. Assuming an *a priori* state vector, a non-linear forward model is linearized around it. Then, the linearized model can be easily inverted and a new estimation for the state vector can be found. However, this inversion method is very time-consuming, due to repeated calls to complex radiative-transfer (RT) forward models that simulate radiances and Jacobians, and subsequent inversion of relatively large matrices.

Most machine learning algorithms do not consider the optimization problem explicitly. Rather, they *learn* from a given dataset and make predictions regarding parameters of interest. In this context, we have developed a new type of algorithm designed for solving inverse problems, called full-physics inverse learning machines (FP-ILMs). Conceptually, the FP-ILM consists of a training phase, wherein the inversion operator is obtained using synthetic data generated using a radiative transfer model (which expresses the "full-physics" component), and an operational phase, in which the inversion operator is applied to real measurements. The main advantage of the FP-ILM over the classical optimization approach is that the time-consuming training phase involving complex RT modeling is performed off-line; the inverse operator itself is robust and computationally simple.

Our objective can be seen as a handling of the fundamental functional relationship between the forward problem and the inverse problem, which is illustrated in Fig. 1. The forward problem deals with the simulation of spectral radiances in the Hartley-Huggins absopption band.

The design of the FP-ILM algorithm was initially inspired in [5], and was specially adapted for deriving volcanic SO_2 plume height from GOME-2 UV data in Efremenko et al. [3]. In this work, the FP-ILM algorithm was designed to estimate an ozone profile shape from a given spectral spectrum. The principal advantage of this data-driven approach is its "simplicity" — the retrieval of the atmospheric parameter of interest can be easier than the classical inversion methods after the coefficients are determined from the training phase. Also note, that in the FP-ILM approach the inverse operator is trained directly, so no additional inversion is required.

2.1. FP-ILM training phase

The FP-ILM algorithm for ozone profile shape retrieval is described in detail by Xu et al. [10], and is summarized here for completeness. The training phase consists of the following main steps:

- clustering of various ozone profile shapes extracted from two ozone climatologies;
- computing GOME-type UV spectra with representative O_3 profiles from each cluster;
- deriving differential spectra using a low-order polynomial;
- assigning an O_3 profile class corresponding to a given spectrum; and
- estimating the O_3 profile shape by scaling to a given total column density.

Thus, the initial problem of ozone profile retrieval is splitted into to separate problems: the ozone total column retrieval and the ozone profile shape retrieval. A schematic diagram of the FP-ILM algorithm during its training and operational phases can be found in Fig. 1 of [10].

The clustering of ozone profile shapes was implemented by the k-means clustering procedure. The reference ozone profiles were extracted from the Bodeker Tier 1.4 database [1] which describes ozone profile variations as functions of time and region, including the seasonal evolution of the Antarctic ozone hole. The Bodeker database was merged with the McPeters/Labow ozone climatology [4] combining data from the MLS (Microwave Limb Sounder) aboard the Aura satellite with data from balloon sondes (1988–2010) in order to better represent tropospheric ozone concentrations.

The UV spectra that resemble GOME-type measurements from representative ozone profiles from each cluster were simulated by the radiative transfer model VLIDORT (Vector LInearized Discrete Ordinate Radiative Transfer) [8]. The computations were performed in the Hartley-Huggins absorption band (280-335 nm). Note that the spectrum in the 280-325 nm wavelength range is sensible to the ozone profile, while the rest part of the spectrum is usually used

doi:10.2760/848593

for ozone total column retrieval. the influence of the slit function is modelled by convolving the ozone absoption cross-section with the slit function. Particularly, the "smart sampling" technique [6] was employed to optimally cover the multi-dimensional input space and to concurrently minimize the number of samples to be generated (i.e. the calls to the radiative transfer model) to describe the output space. Essentially, the smart sampling technique is based on the Halton series which are superior over the equidistant grid series. Then, a third-order polynomial fit in the wavelength domain was performed in order to obtain differential spectra $I_c^\delta(\lambda) = I^\delta(\lambda) - P_N(\lambda, \mathbf{p}_c)$, where $I^\delta(\lambda)$ and $P_N(\lambda, \mathbf{p}_c)$ are synthetic noisy spectra and a polynomial with an order of N, respectively, and the vector of polynomial coefficients $\mathbf{p}_c$ tackles the minimization problem

$$\mathbf{p}_c = \arg\min_{\mathbf{p}} \|I^\delta(\cdot) - P_N(\cdot, \mathbf{p})\|^2 \,. \tag{1}$$

In practice this step can largely remove instrumental artefacts including degradation in the observed spectrum and other factors which are not directly accounted for in the forward model.

For the last steps, we utilized two sets of neural network (NN) techniques performing different functionalities. The first NN classifies the ozone profile shape corresponding to a given data vector and its corresponding model parameter vector, while the second NN ensemble derives a nonlinear scaling function for each individual cluster, yielding scaled ozone profile shapes according to the total ozone.

2.2. Total ozone retrieval

Many operational algorithms for total ozone retrieval are based on the Differential Optical Absorption Spectroscopy (DOAS) techniques, i.e., the vertical column density V is determined from the fitted slant column density S using the iterative scheme

$$V^{i+1} = \frac{S/M^i_{\text{ring}}}{(1-\phi)\,A^i_{\text{clear}} + \phi A^i_{\text{cloud}}} \,, \tag{2}$$

where the superscript indicates the number of iteration step. A_{clear} and A_{cloud} are the air mass factors for clear-sky and cloudy scenarios, respective, which require our retrieved ozone profile shapes. M_{ring} is the molecular Ring term, and ϕ is the intensity-weighted cloud fraction.

During the operational phase, we implemented the inverse functions (i.e., both trained NNs) derived from the training phase in the framework of total ozone retrieval. Since the conversion from A to V is an iterative process, the profile shape estimated from at the current iteration was used to obtain the next iterate of V. With the newly retrieved V, the ozone profile shape was further adjusted.

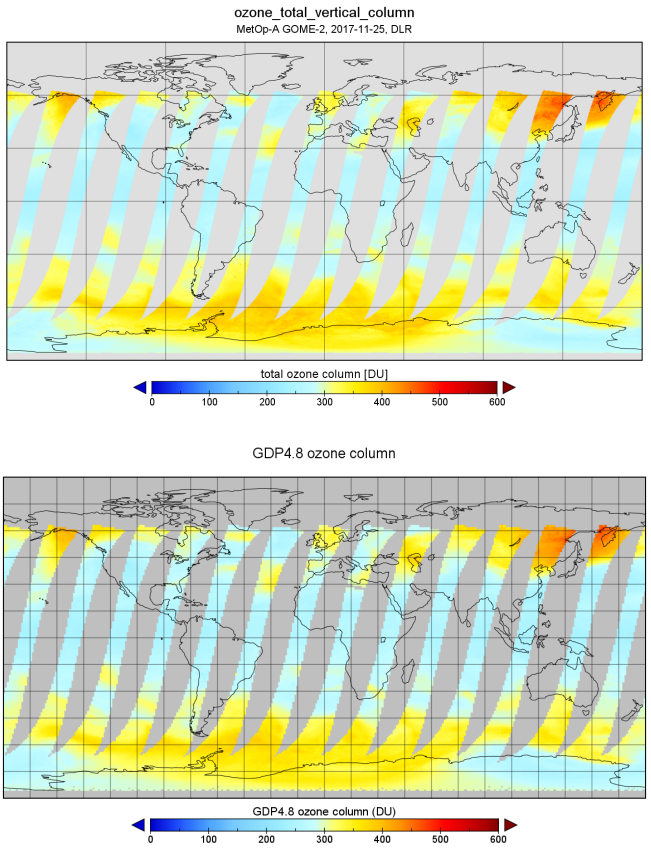


Fig. 2. Comparison of retrieved total ozone vertical column densities between the DOAS-based retrieval using the FP-ILM profile (top) and the GDP 4.8 product (bottom) from GOME-2 data.

3. FIRST RESULTS

Figure 2 depicts the retrieved total ozone from GOME-2 data on November 25, 2017. It can be identified that both retrievals using the two ozone profile schemes agree well, indicating that the FP-ILM profile shape used in the total ozone retrieval seems reasonable and may reflect the actual measurement conditions.

More importantly, the computational effort was drastically reduced using the FP-ILM algorithm. According to [2], an OEM-based single profile retrieval converges in less than 10 iterations and typical computation time is 20–30 s; whereas the FP-ILM prototype algorithm normally takes less than 0.5 s. The operational processor using FP-ILM ran about almost three orders of magnitude faster. It should be noticed that the most time-consuming steps were the radiative transfer computations and the following NN trainings,which are conducted off-line.

4. CONCLUSIONS

The FP-ILM algorithm uses typical machine learning techniques to characterize ozone profile shapes from GOME-type sensors in a very efficient manner and can be easily adapted to other hyperspectral UV instruments. The FP-ILM comprises the radiative transfer model. However the time consuming computations are performed off-line. The resulting operator is computationally simple and fast.

Optimization is currently ongoing to better integrate the FP-ILM algorithm to the operational processing system Universal Processor for UV/VIS Atmospheric Spectrometers (UPAS) at DLR. Further work will place an emphasis on the NN jacobians (derivatives of the outputs with respect to the inputs) for a retrieval error characterization. The retrieval sensitivity of the profile shape retrieval will also be investigated.

The FP-ILM framework will be used for the near-real-time processing of the new European Sentinel sensors with their unprecedented spectral and spatial resolution and corresponding large increases in the amount of data.

REFERENCES

[1] G. E. Bodeker, B. Hassler, P. J. Young, and R. W. Portmann. A vertically resolved, global, gap-free ozone database for assessing or constraining global climate model simulations. *Earth Syst. Sci. Data*, 5(1):31–43, 2013. doi: 10.5194/essd-5-31-2013.

[2] Johan de Haan. Sentinel-5p TROPOMI ozone profile and troposopheric profile. Algorithm theoretical basis document, Royal Netherlands Meteorological Institute (KNMI), 2015. URL https://earth.esa.int/documents/247904/2476257/Sentinel-5P-TROPOMI-ATBD-O3-Profile.

[3] Dmitry S. Efremenko, Diego G. Loyola R., Pascal Hedelt, and Robert J. D. Spurr. Volcanic SO_2 plume height retrieval from UV sensors using a full-physics inverse learning machine algorithm. *Int. J. Remote Sensing*, 38(sup1):1–27, 2017. doi: 10.1080/01431161.2017.1348644.

[4] Gordon J. Labow, Jerald R. Ziemke, Richard D. McPeters, David P. Haffner, and Pawan K. Bhartia. A total ozone-dependent ozone profile climatology based on ozonesondes and Aura MLS data. *J. Geophys. Res. Atmos.*, 120(6):2537–2545, 2015. doi: 10.1002/2014JD022634.

[5] Diego G. Loyola R. Applications of neural network methods to the processing of earth observation satellite data. *Neural Netw.*, 19(2):168–177, 2006. doi: 10.1016/j.neunet.2006.01.010. Earth Sciences and Environmental Applications of Computational Intelligence.

[6] Diego G. Loyola R, Mattia Pedergnana, and Sebastián Gimeno García. Smart sampling and incremental function learning for very large high dimensional data. *Neural Netw.*, 78:75–87, 2016. doi: 10.1016/j.neunet.2015.09.001. Special Issue on "Neural Network Learning in Big Data".

[7] G. M. Miles, R. Siddans, B. J. Kerridge, B. G. Latter, and N. A. D. Richards. Tropospheric ozone and ozone profiles retrieved from GOME-2 and their validation. *Atmos. Meas. Tech.*, 8(1):385–398, 2015. doi: 10.5194/amt-8-385-2015.

[8] R.J.D. Spurr. VLIDORT: A linearized pseudo-spherical vector discrete ordinate radiative transfer code for forward model and retrieval studies in multilayer multiple scattering media. *J. Quant. Spectrosc. Radiat. Transf.*, 102(2):316–342, 2006.

[9] Jian Xu, Franz Schreier, Adrian Doicu, and Thomas Trautmann. Assessment of Tikhonov-type regularization methods for solving atmospheric inverse problems. *J. Quant. Spectrosc. Radiat. Transf.*, 184:274–286, 2016. doi: 10.1016/j.jqsrt.2016.08.003.

[10] Jian Xu, Olena Schüssler, Diego Loyola R, Fabian Romahn, and Adrian Doicu. A novel ozone profile shape retrieval using Full-Physics Inverse Learning Machine (FP-ILM). *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 10(12):5442–5457, 2017. doi: 10.1109/JSTARS.2017.2740168.

MULTI-TEMPORAL LAND COVER CLASSIFICATION USING SENTINEL DATA AND THE EO-LEARN OPEN-SOURCE PYTHON PROJECT

M. Lubej, M. Aleksandrov, M. Batič, M. Kadunc, G. Milčinski, D. Peressutti, A. Zupanc

Sinergise d.o.o., Ljubljana, Slovenia

ABSTRACT

Land cover mapping and monitoring have a paramount role in understanding and managing changes in territory and their impact on the ecosystem. The advent of open-access satellite data and advances in data processing and analysis has opened up new possibilities to compute large-area land cover maps at high spatial resolution.

In this paper, we present a start-to-end workflow to generate a land cover map for Slovenia. We do this by using eo-learn, a framework to handle multi-temporal and multi-source satellite data, both in raster and vector format. The framework allows splitting the area-of-interest (AOI) into smaller patches, that can be processed with limited computational resources, and allows automation of the processing pipeline. The framework consists of open-source Python packages and is designed to facilitate prototyping and building of end-to-end Earth Observation applications.

In the presented workflow, we use the annual Sentinel-2 images to construct a machine learning model. We report on the steps required to build such a pipeline and on how to best optimize them. To foster the exploitation of open-access data and the uptake of technologies, the complete pipeline – from data processing to predicting the labels – is open-source. This allows for the reproducibility of the process or even its further optimization, which fits one's purpose.

***Index Terms*—** machine learning, land cover classification, open-source

1. INTRODUCTION

Accurate land cover mapping has a paramount role in describing and analyzing the environment and its changes, in particular concerning the management and monitoring of natural resources, human-made activities, and their impact on the ecosystem dynamics [2, 3]. In recent years, local, national and international authorities have been increasingly relying on land cover information for territory management and policy making, although decision-makers are not yet exploiting the full potential of such information [1].

The advent of open-access satellite imagery (e.g. available through the Copernicus programme), and new image processing and machine learning techniques have allowed the production of land cover maps for large areas using multi-temporal and multi-source data. Wulder *et al.* [3] describe this evolution in generating land cover maps as the *Land Cover 2.0* paradigm. Instances of this paradigm include the work of Midekisa *et al.* [5], where fifteen years of Landsat images were processed to produce a land cover and land use change maps over continental Africa. Authors visually inspected images retrieved from Google Earth to generate training labels and employed random forest for the supervised classification. Inglada *et al.* [4] proposed a method for the automatic land cover mapping of the entire region of France using annual Landsat imagery. Training labels were created by combining existing and outdated land cover maps. A stratified classification based on climate areas was used, using random forests as base classifiers. Prediction labels and prediction confidence were generated for the 2014 year. For a more comprehensive review of land cover mapping methods using time-series, refer to [2].

Despite these substantial advances in the field, land cover monitoring mostly remains a research topic, with land cover maps at national and global scale being mainly produced by large consortia of research and commercial centers. Main reasons for a delayed commercial exploitation lie in the high computational resources required for downloading, processing and storing data. In this paper, we propose a machine learning framework to generate land cover maps using open-access, multi-temporal Sentinel-2 images and open-source tools. The framework focuses on optimizing data processing using limited resources and data parallelization. The framework could enable entry into the field for small-size research and commercial entities, as well as citizens. For this reason, to foster the uptake of technologies, improve on the current state-of-the-art, and stimulate the development of Earth Observation applications, we open-source the code used to generate the land cover map.

2. MATERIALS

In this work, the framework was used to generate a land cover map of the Republic of Slovenia for the year 2017. The inputs

`eo-learn` was developed under the Perceptive Sentinel European H2020 project.

to the framework are a shape-file defining the geometry of the AOI, the Sentinel-2 L1C images for the entire year, and a set of training labels. A key advantage of the framework over existing ones is the possibility to automatically split the AOI into smaller areas, which can be more easily handled with limited computational resources. Avoiding to download and process entire tile products (e.g. Sentinel-2 granules) provides flexibility and facilitates automation of processing pipelines.

2.1. Image data

Splitting the AOI into smaller patches and downloading Sentinel-2 L1C bands for each patch was executed using `sentinelhub-py`, a Python package[1] that acts as a wrapper for the Sentinel-Hub OGC web services. Sentinel-Hub services are subscription-based, although free accounts for research institutes and start-ups are available. An alternative method to retrieve satellite imagery for areas of any given size is by querying products encoded as Cloud-Optimised GeoTiffs (COG). Sentinel-2 L2A products could be used instead of L1C products, or additional imaging sources (e.g. Landsat-8, Sentinel-1) could be similarly added to the processing pipeline.

Cloud masks were generated using the open-source `s2cloudless`[2] Python package, which applies a pre-trained machine learning classifier to Sentinel-2 L1C products to produce cloud probability and cloud masks.

2.2. Label data

The pipeline presented here uses supervised machine learning models for land cover classification, however, the framework can be used for any processing or machine learning task. Training and validation labels were obtained from the Slovenian Ministry of Agriculture, Forestry and Food. Datasets from the year 2002 are freely available to download in vector format and can be used for land cover change studies. The original labels were mapped to 10 land cover classes (*cultivated land*, *forest*, *grassland*, *shrubland*, *water*, *wetlands*, *tundra*, *artificial surface*, *bareland*, *snow and ice*) using the definition from the GlobeLand30 map [6]. These labeled datasets are also available on the open-access cloud-based GIS Geopedia, so that they can be queried and retrieved through OGC web requests.

If labeled datasets are not available, as can be the case in real-world applications, a consensus dataset can be derived from a combination of open-source land cover maps at different spatial resolutions (e.g. Corine Land Cover, OpenStreetMap, Climate Change Initiative Land Cover), as described in [4].

[1] `sentinelhub-py` is available on PyPI and GitHub
[2] `s2cloudless` is available on PyPI and GitHub

2.3. Processing framework

The `eo-learn`[3] Python package was used to build the processing pipeline to train and validate a machine learning model. `eo-learn` allows processing multi-temporal and multi-source remote sensing data, both in raster and vector format. A pipeline in `eo-learn` is defined as a connected acyclic graph of well-specified tasks to be performed on the data. Example tasks include data retrieval and rasterization, cloud and snow masking, co-registration, interpolation, feature manipulation, and geometric sampling. Tasks are modular and allow users to easily implement their own. The Python environment allows to use the many available packages for data analysis and machine learning, and quickly prototype and test EO applications. `eo-learn` supports parallelization of operations, such that the same workflow (e.g. data preparation for land cover classification) can be run in parallel for the smaller patches constituting the AOI. Logging and reporting allow to monitor and debug the execution of the processing pipeline. Being an open-source project, contributions from users help to improve the scale and scope of the supported features.

3. METHOD

The automated pipeline for predicting the land-cover labels is similar to [4] and follows the following setup:

1. Split the AOI into manageable chunks (based on available computing resources);
2. Create AOI patches and fill them with information (Sentinel-2 band data, cloud maps, reference maps, etc.);
3. Spatially sample pixels inside each patch;
4. Interpolate the time-series and re-sample to unified dates;
5. Train and validate the machine-learning model.

The code implementing this pipeline is available in the `examples/land-cover-map` directory of the `eo-learn` GitHub project.

3.1. Splitting the AOI

The boundary of the Republic of Slovenia was taken from Natural Earth and a buffer was added. The bounding box of the AOI has a size of about $250\ \mathrm{km} \times 170\ \mathrm{km}$. The AOI was split into 25×17 equal parts, which resulted in about 300 patches of about $10^3\ \mathrm{px} \times 10^3\ \mathrm{px}$ at a $10\ \mathrm{m}$ resolution. The splitting choice depends on the amount of available resources, so the pipeline can be executed on a high-end scientific machine with a large number of CPU's and a large memory, as well as on a laptop. The output of this step was a list of bounding boxes, covering the AOI as shown in Figure 1.

[3] `eo-learn` is available on PyPI and GitHub

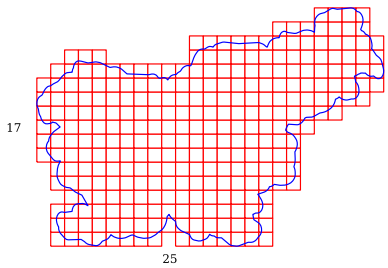


Fig. 1. Splitting of the country into smaller chunks.

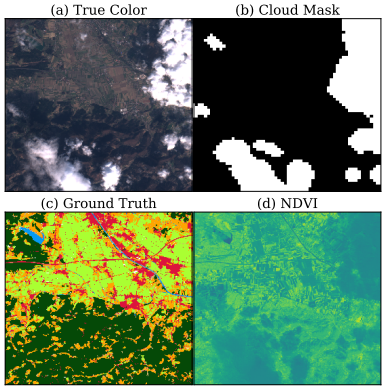


Fig. 2. Example of a patch with the contained information: true color (a), cloud mask (b), ground truth (c) and NDVI (d).

3.2. Creating and adding information to AOI patches

The bounding boxes covering the AOI were used to create patches where the information for the corresponding area was stored. Given the time interval of interest (e.g. from 2017-01-01 to 2017-12-01), `eo-learn` used the `sentienlhub-py` package to download the six Sentinel-2 bands (B2, B3, B4, B8, B11, B12) for each patch. The `s2-cloudless` cloud detector was subsequently used to obtain the cloud probabilities and cloud masks. From the cloud masks and from the output of Sentinel-2 image data one can construct a mask of valid pixels over the entire AOI and time interval, providing information about non-cloudy scenes in the satellite swath. Additional band combinations were calculated in order to obtain more complex features such as normalized difference vegetation and water indices (NDVI, NDWI) and the euclidean norm of the used bands (NORM).

Ground truth information is obtained in a vector format, which is then rasterized over the given bounding box and added to the patch. Figure 2 shows information contained within an example patch.

3.3. Spatial sampling

A random spatial sampling of patches is performed to select time-series of pixels which are used for ML model training.

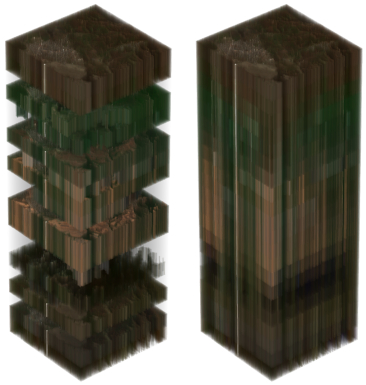


Fig. 3. Visualization of a time-series interpolation. The image on the left shows the time-series with missing data (invalid pixels, clouds, etc.), while the image on the right shows the time-series after the interpolation procedure.

The random sampling is uniform throughout the patch, independently of the labels. Prior to sampling, erosion with a disk size of 1 pixel is performed, in order to remove single pixels, pixel-wide structures, and pixels on the borders of land cover classes, where substantial class mixing is possible. An alternative sampling strategy would be to over-sample and under-sample label classes to provide the same number of samples per class.

3.4. Time-series interpolation and resampling

Due to non-constant acquisition dates of the satellites and irregular weather conditions, missing data is very common in the field of EO. One way to tackle this problem is to apply the mask of valid pixels in the time series and interpolate the values in order to "fill the gaps" due to missing data, as visualized in Figure 3. After the interpolation, values at uniform or non-uniform dates can be evaluated to unify the dates among all the patches for an arbitrary size of the multi-temporal stack.

3.5. Training the ML model

As the pixels in the patches are spatially sampled and their values interpolated, the pixels and labels are accumulated over all the patches. Time frames of all the features act as independent features of the ML model, so training data has shape $n \times m$, where n represents the number of all training points, where each point has $m = f \times t$ features, corresponding to f features in the information dimension, and t corresponding to the number of resampled dates. The data sample is split at the patch-level into training and test samples, according to the $80 : 20$ rule, and the obtained training data is further split into the training and cross-validation sample, according to the same rule as above. With the model setup of $f = 9$, $t = 45$ and with the number of training pixels of about $n = 7.5 \times 10^6$, the ML model takes about and

Table 1. Per-class validation metrics of the optimal ML model. Values are in %.

Class	F_1 score	Precision	Recall
Cultivated land	89.9	86.6	93.5
Forest	98.5	99.3	97.7
Grassland	88.5	91.3	85.8
Shrubland	51.6	40.3	71.9
Water	93.4	94.7	92.0
Wetlands	16.0	11.3	27.3
Artificial surface	88.9	87.0	91.0
Bareland	86.4	88.6	84.4

Table 2. Results on the cloud coverage effect.

Set-up	Weighted F_1 score	Overall accuracy
Case A_1	93.1	92.8
Case A_2	94.2	94.0
Case A_3	94.4	94.1

hour to train with the default hyper-parameter settings. The cross-validation sample is used for optimization, such as ML model hyper-parameter optimization, while the model validation is performed on the test sample. Light Gradient Boosting Machine (`LightGBM`) was used as machine learning model.

4. RESULTS

The trained model was used to predict the labels on the test sample and the obtained results were then validated. In Table 1 we present the per-class F_1 score, precision and recall for the optimal model. The overall accuracy and F_1 score are reported for the B_3 case in Table 3. Additionally, we performed several experiments where, for example, we evaluated how different cloud coverage or temporal resampling affect the ML model performance. Few of these experiment results are shown in Table 2 with cases A_1 (no cloud filtering and no gap-filling), A_2 (with cloud filtering and no gap-filling) and A_3 (with both cloud filtering and gap-filling), or in Table 3 with cases B_1 (uniform temporal resampling at 16-day rate), B_2 (uniform resampling at 8-day rate) and B_3 (non-uniform resampling with frequency equal case B_2). In general, poor prediction was achieved for under-represented classes such as *wetlands* and *shrubland*. The full extent of experiments will be shown at the conference.

Table 3. Results on different resampling techniques.

Set-up	Weighted F_1 score	Overall accuracy
Case B_1	94.4	94.1
Case B_2	94.5	94.3
Case B_3	94.6	94.4

5. DISCUSSION AND CONCLUSIONS

The aim of this work was to showcase the use of open-source tools to build Earth Observation applications. In particular, land cover mapping over large areas using machine learning and annual Sentinel-2 images was described. The pipeline was inspired by the work of Inglada *et al.* [4], although the modularity of the framework supports implementation and extension to custom algorithms and processing tools. The proposed pipeline is designed to be generic and applicable to AOI of different size and location, although optimization of the parameters was investigated for a simple use-case (i.e. Slovenia land cover for the year 2017). Therefore, accuracy results may not generalize to other AOI. Moreover, in this work, high-quality label data was employed. The use of different open-source labels for the training of the machine learning model, as in [4], might lead to worse accuracy results due to discrepancies in spatial resolution and out-of-date information. Future work will investigate and report on the influence of such an approach on the results. The influence of different classification algorithms, such as convolutional networks, will also be further investigated and reported.

The proposed pipeline was built using open-source Python tools that allow to automatically process remote sensing data with limited computational resources. The proposed framework can be used to build Earth Observation applications quickly and efficiently. We open-source the code used to generate the land cover map to facilitate entry to the field of Earth Observation for small and medium-sized enterprises and research centers.

REFERENCES

[1] Wulder M., and Coops N., "Satellites: Make Earth observations open access", *Nature*, 513, DOI: 10.1038/513030a, 2014.

[2] Gomez C., White J., and Wulder M., "Optical remotely sensed time series data for land cover classification: A review", *ISPRS Journal of Photogrammetry and Remote Sensing*, 116, DOI: 10.1016/j.isprsjprs.2016.03.008, 2016.

[3] Wulder M., Coops N., Roy D., White J., and Hermosilla T., "Land cover 2.0", *International Journal of Remote Sensing*, 39(12), DOI: 10.1080/01431161.2018.1452075, 2018.

[4] Inglada J., Vincent A., Arias M., Tardy B., Morin D., and Rodes I. "Operational High-Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series", *Remote Sensing* 9(12), DOI: 10.3390/rs9010095, 2017.

[5] Midekisa A., Holl F., Savory D., Andrade-Pacheco R., Gething P., Bennett A., and Sturrock H., "Mapping land cover change over continental Africa using Landsat and Google Earth Engine cloud computing", *PLoS ONE*, 12(9), DOI: 10.1371/journal.pone.0184926M, 2017.

[6] Jun C., Ban Y., and Li S., "Open access to Earth land-cover map", *Nature*, DOI: 10.1038/514434c, 2014.

HOW MANY ROADS? OBJECT SEGMENTATION ON SATELLITE IMAGERY IN A PRODUCTION ENVIRONMENT

Iris Wieser, Peter Schauer, Martin Angelhuber, Martin Riedl, Paul Fischer, Elisa Canzani

Industrieanlagenbetriebsgesellschaft (IABG) mbH
Ottobrunn, Germany

ABSTRACT

Developing automated information extraction methods is indispensable for handling the large amount of satellite imagery operated by governments and businesses around the world. Recent advances in the area of deep learning have successfully contributed to automating traditional object detection tasks. This work applies a well-known convolutional neural network (CNN) to extract roads from satellite images which are diverse in terms of spatial resolution, landscape, viewing angle, and other properties. We first conduct several experiments by training a U-Net based network on high-resolution satellite imagery of 11 regions in 3 continents. In a second step, we apply another network trained on thinned targets, followed by a centerline algorithm and a custom simplification algorithm to transform the results to a connected vector representation. In this paper, we propose a complete workflow for automated road extraction with a special focus on the applicability in a vector data production environment.

***Index Terms*—** Deep Learning, Semantic Segmentation, Remote Sensing, Convolutional Neural Network, Road Extraction

1. INTRODUCTION

Retrieving information from high-resolution satellite imagery is essential for a wide range of application fields, such as mobility, agriculture, and defense. A common approach to extract features is to manually segment or vectorize the images. This is a time-consuming task and requires a high level of expertise. Therefore this approach does not scale well to the drastically increasing amounts of satellite data available.

Most established approaches for automating object recognition tasks are limited to specific scenarios (e.g. land cover) or satellite specifications (e.g. resolutions), such as [9]. Due to these generalization limitations, they often fail to meet customer quality requirements.

Novel network architectures proposed in the field of deep learning, such as U-Nets [10], received particular attention for their performance in pixel-wise semantic segmentation and their generalization capabilities. They have been applied successfully to one-class segmentation tasks, such as road detection [1, 7], and multi-class segmentation tasks [3, 5], such as detection of vegetation, buildings, water bodies and so on.

The segmentation output is usually represented as a raster graphic. However, for many practical applications the output needs to be in a vector format. There are several methods to convert raster representations of line features, like roads or rivers, to vector representations and create simplified and connected graphs. For example, Haunert and Sester [2] present an automatic approach to create road centerlines using special characteristics of straight skeletonization algorithms. Máttyus et al. [6] suggest a shortest path algorithm to close missing connections in the extracted road graph.

While most research investigates single aspects of road detection with a major focus on algorithmic challenges [8], we explore the application of deep learning for road detection in geo-data production environments towards leveraging rapid-mapping applications like disaster mapping. For this, the models need to a) generalize well to different satellite scenes with different spatial resolutions, regions in the world, seasons, and so on and b) be efficient in terms of training and prediction time, and c) provide a vector data output for subsequent processing.

In this paper, we propose a workflow that is applicable in a geo-data production environment. We refer to road detection for demonstration purpose. In general, our approach can be applied to other domains to support different object segmentation tasks in satellite image analysis, like coniferous forest segmentation. Based on manually - and thus highly accurate - labeled satellite scenes taken from different regions in the world, our approach is highly generic in terms of different scenarios and satellite specifications. By using model ensemble methods and postprocessing steps, we can generate an accurate vector representation of the road network from satellite imagery, while taking into account the requirements of production environments.

2. DATA

For training, high-resolution satellite images with a spatial resolution of $1\,\mathrm{m}$ are used. So far, we selected 19 areas of interest (AOI) and labeled them manually with ArcGIS. The

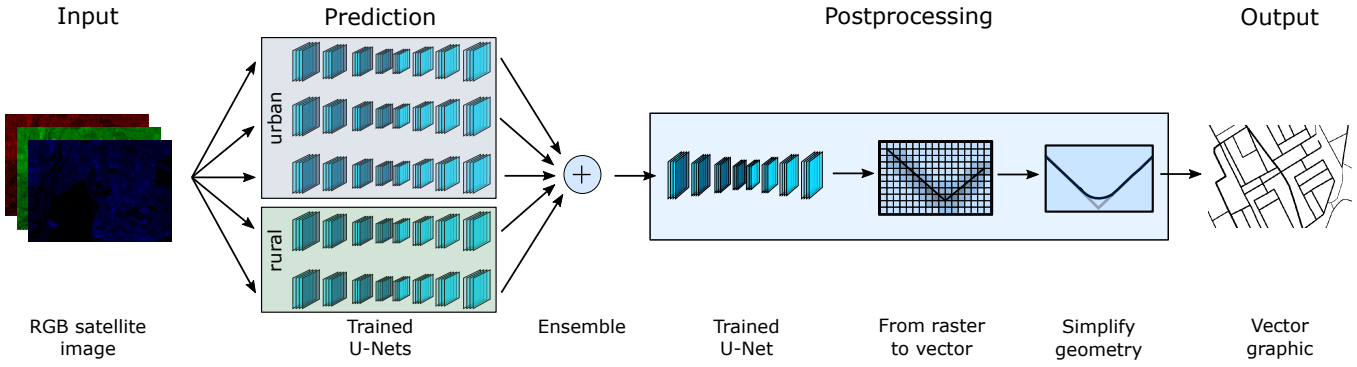


Fig. 1. Our proposed workflow to extract road vectors from satellite imagery in a production environment.

selected AOIs contain both rural and urban areas of 11 regions in 3 different continents (i.e. Europe, Africa, and Central America). In total, the imagery currently covers a surface of about $420\,\mathrm{km}^2$.

To test the transferability of our workflow to other scenarios (e.g. seasons, climate zones) and sensor specifications (e.g. spatial and radiometric resolution, viewing angles), we use satellite imagery from various satellites, such as SkySat, Planet Dove, Sentinel-2 and others.

3. METHODOLOGY

Fig. 1 illustrates our proposed workflow, which we implemented in Keras. It consists of two phases: prediction and postprocessing. We use an input size of $(512, 512)$ with 3 channels (RGB). These input images are classified at pixel level as roads in the prediction phase. In the postprocessing phase, the conversion from raster to vector representation takes place in three steps. First, we use another U-Net to thin the detected roads and fill gaps in the road network. This is an important preparation for the second step, the conversion from raster data to centerlines, and eventually vector data. The third step is the simplification of the vector data.

For the prediction, we use the network architecture shown in Fig. 2, which is based on the U-Net architecture originally proposed by Ronneberger et al. [10]. The U-Net is an adaption of a fully convolutional network specifically designed to solve segmentation problems.

Our architecture consists of 3 downsampling blocks for encoding and 3 upsampling blocks for decoding. A downsampling block has 3 convolutional layers and a max pooling layer of size $(2, 2)$. Each convolution layer consists of a convolution, a batch normalization and a ReLu activation function. Conversely, an upsampling block has 3 convolutional layers and a transposed convolutional operation. The number of filters is increased after every block by a factor of two in the decoding part and decreased by a factor of two in the encoding part. For training we use an Adam optimizer with a loss function that is a convex combination of the cross-entropy and

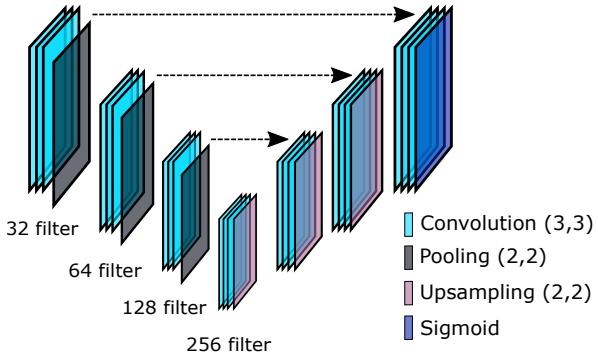


Fig. 2. The applied network architecture based on a fully connected U-Net. The amount of filters duplicates by each downsampling block.

the negative logarithm of the Jaccard index as in [1], i.e.

$$L(X, \hat{X}) = \alpha H(X, \hat{X}) - (1 - \alpha) \log(J(X, \hat{X})),$$

where H is the binary cross-entropy and J is the Jaccard index.

For the prediction phase of our workflow, we have trained 3 models on images that contain mainly urban areas, while 2 models are trained on mainly rural-like regions. The images are clustered as "rural" and "urban" by the amount of road pixel in the target. Then, we average the predictions of these 5 models. The ensemble of these models allows an accurate segmentation in both rural and urban areas.

For the postprocessing, we apply another network of the same architecture which is trained using centerlines derived from vector data as labels and the predictions of the first network as features. We smooth these predictions by a filter and use it as input for the raster to vector conversion. Then, we convert the road polygons to their centerline. To fit quality requirements for road vector data, junctions are generalized and reconstructed with respect to topological and geometrical features. We finally simplify the road lines by removing unnecessary vertices within a specified distance.

(a) Original satellite scene (b) Prediction (raster) (c) Postprocessing (vector) (d) Ground truth

Fig. 3. Intermediate outputs of the steps in our workflow.

4. RESULTS AND DISCUSSION

An example of intermediate and final outputs of our workflow is shown in Fig. 3. From left to right, it shows a satellite image with a spatial resolution of 1 m (a) on which a pixelwise object segmentation is applied by an ensemble of 5 models (b) and is then converted to the final vector output (c). For comparison, the manually created ground-truth vector is also displayed (d). Note that all satellite scenes shown and discussed in this section have not been included in the training.

We have used 4 GPUs for training. Based on our train set of roughly $420\,\mathrm{km}^2$ the training of 5 models took about 3 h.

Even though the prediction in Fig. 3 (b) appears rather accurate, we face challenges in converting this raster graphic to vectors, as it contains small segments of false positives or false negatives. Therefore, postprocessing is necessary and essential to obtain a meaningful vector representations.

During development we noticed that commonly used metrics, such as the Jaccard index or F1 score, are not sufficient to describe the correctness of the output. As they measure the accuracy based on the raster graphic, they do not necessarily yield a good estimate of the quality of the road network as vectors. For example, those scores do not indicate if and how many small missing segments exist. Thus, they can not infer the connectedness of the vector graph which is an important graph property for the usability of the output.

In a production environment it is important that the objects can be segmented a) to high accuracy, b) as vector representations, and c) in a minimal amount of time. The big advantage of the type of networks we implement is that they can be adjusted to arbitrary size of input images. Hence, when using a sliding window approach for predicting large satellite scenes, one can tune the window size to best use the available hardware. This allows deployment of the models on systems with limited computational power and flexible distribution to scale with increasing demand. A prediction of a $20\,\mathrm{km}^2$ scene with a spatial resolution of 1 m takes 4 min using a NVIDIA GTX 1060 graphic card (136 min using CPU only). Thus, the user can obtain a very good first impression of unknown regions already before further postprocessing steps. Fig. 4 shows a satellite image of Nigeria. The OpenStreetMap data is very sparse in this area, as shown in blue. However, our prediction, illustrated in pink, allows the user to quickly extract how many roads roughly exists in this scene.

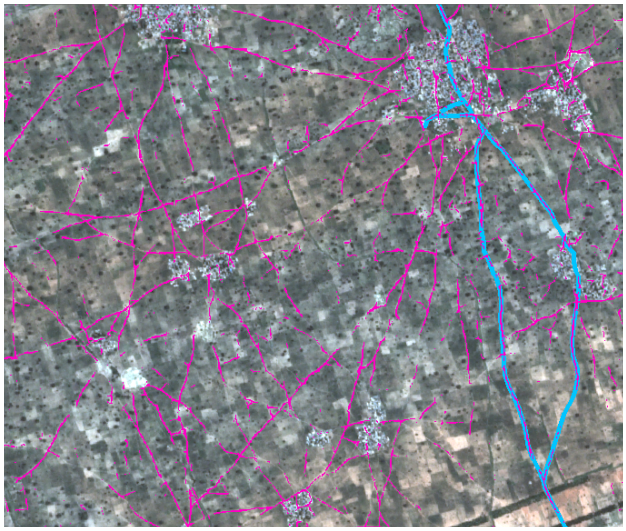


Fig. 4. Exemplary Planet Dove satellite image showing a region in Nigeria with a spatial resolution of 3 m. The corresponding OpenStreetMap data is illustrated in blue. Our prediction is shown in pink, given an impression of how many roads truly exist in this region.

Moreover, our workflow can be applied to satellite imagery with various satellite specifications. Fig. 5 shows 4 satellite images of different regions. They vary in their spatial resolutions (from 1 m to 10 m) and include both urban and rural areas. By the combination of prediction and postprocessing our workflow outputs connected vector representations for various satellite images. The final vector representations are illustrated in the second row on top of the corresponding satellite image. Thus, our workflow can be easily scaled for deployment in production as a part of automatic feature labeling systems for satellite imagery analysis.

5. CONCLUSION

Satellite imagery is highly important for a wide range of applications, such as topographic, land cover and disaster

doi:10.2760/848593

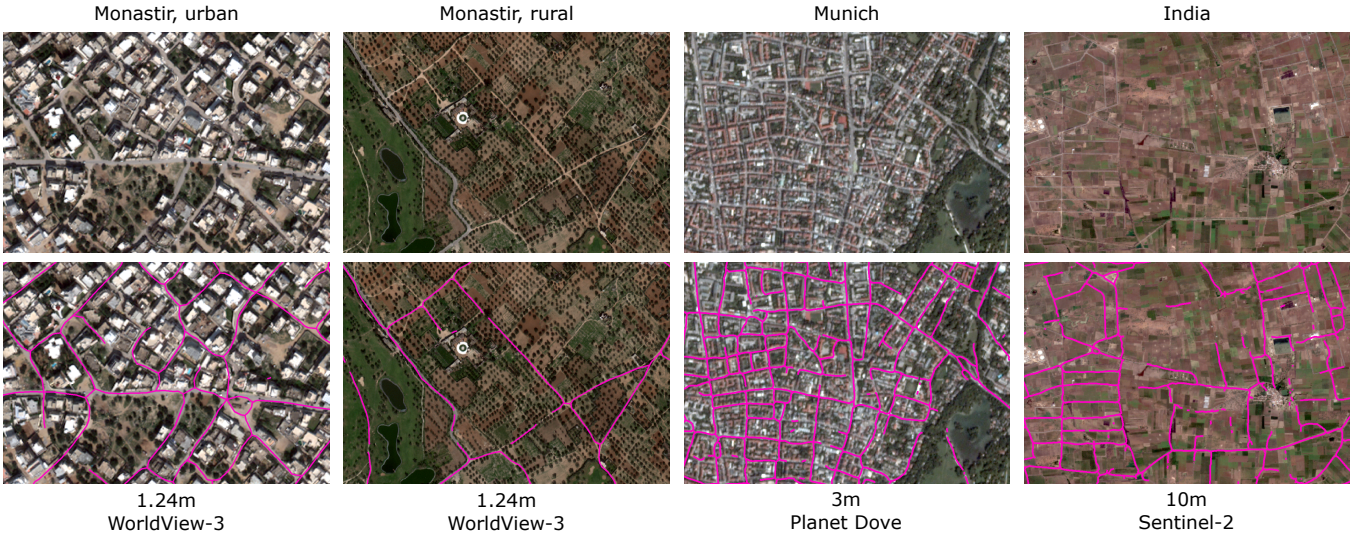


Fig. 5. Exemplary results of our proposed workflow. 4 different satellite images are shown in the top row. The bottom images show the satellite scene with the vector output of our workflow (shown in pink).

mapping, as well as change detection. Due to the increasing amount of satellite data and rapid change of infrastructures, it becomes extremely relevant to automate the process of object segmentation. We propose a workflow that enables object segmentation in satellite imagery in a production environment. In particular, we demonstrate our workflow by applying it to road segmentation tasks. Our generic method comprises a prediction phase with several architectures followed by various postprocessing operations. This workflow allows an automatic extraction of a vector representation from a satellite scene.

We obtain high computational efficiency by using fully convolutional neural networks. The algorithm is able to predict a $20\,\mathrm{km}^2$ scene with a spatial resolution of $1\,\mathrm{m}$ within a few minutes on a middle class graphic card. This allows the user to obtain almost immediately a first impression of unknown regions. Another advantage is the diversity of our dataset. We train our models on both rural and urban areas to achieve a high transferability to regions not included in the training set, while obtaining comparable results.

REFERENCES

[1] Alexander Buslaev, Selim Seferbekov, Vladimir Iglovikov, and Alexey Shvets. Fully convolutional network for automatic road extraction from satellite imagery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

[2] Jan-Henrik Haunert and Monika Sester. Area collapse and road centerlines based on straight skeletons. *GeoInformatica*, 12(2): 169–191, 2008.

[3] Vladimir Iglovikov, Sergey Mushinskiy, and Vladimir Osin. Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. *arXiv preprint arXiv:1706.06169*, 2017. URL https://arxiv.org/pdf/1706.06169.pdf.

[4] P. Kaiser, J. D. Wegner, A. Lucchi, M. Jaggi, T. Hofmann, and K. Schindler. Learning aerial image segmentation from online maps. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11):6054–6068, Nov 2017. ISSN 0196-2892. doi: 10.1109/TGRS.2017.2719738.

[5] Martin Längkvist, Andrey Kiselev, Marjan Alirezaie, and Amy Loutfi. Classification and segmentation of satellite orthoimagery using convolutional neural networks. *Remote Sensing*, 8 (4):329, 2016.

[6] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3458–3466, 2017.

[7] Volodymyr Mnih and Geoffrey E Hinton. Learning to detect roads in high-resolution aerial images. In *European Conference on Computer Vision*, pages 210–223. Springer, 2010.

[8] Javier A. Montoya-Zegarra, Jan Dirk Wegner, L'ubor Ladicky, and Konrad Schindler. Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques. 2015.

[9] Alameen Najjar, Shun'ichi Kaneko, and Yoshikazu Miyanaga. Combining satellite imagery and open data to map road safety. In *AAAI*, 2017.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

REMOTE SENSING DATA ANALYTICS WITH THE UDOCKER CONTAINER TOOL USING MULTI-GPU DEEP LEARNING SYSTEMS

Gabriele Cavallaro[1], *Valentin Kozlov*[2], *Markus Götz*[2], *Morris Riedel*[1]

[1]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany
[2]Steinbuch Centre for Computing, Karlsruhe Institute of Technology, Karlsruhe, Germany

ABSTRACT

Multi-GPU systems are in continuous development to deal with the challenges of intensive computational big data problems. On the one hand, parallel architectures provide a tremendous computation capacity and outstanding scalability. On the other hand, the production path in multi-user environments faces several roadblocks since they do not grant root privileges to the users. Containers provide flexible strategies for packing, deploying and running isolated application processes within multi-user systems and enable scientific reproducibility. This paper describes the usage and advantages that the *uDocker* container tool offers for the development of deep learning models in the described context. The experimental results show that *uDocker* is more transparent to deploy for less tech-savvy researchers and allows the application to achieve processing time with negligible overhead compared to an uncontainerized environment.

Index Terms— Containers, uDocker, multi-GPU, deep learning, classification, remote sensing.

1. INTRODUCTION

In this era of a growing number of earth observation satellite and aerial platforms the volume, variety and acquisition rate of remote sensed images have been dramatically increased. This introduced remarkable challenges that lie within the entire acquisition and processing data pipeline—i.e., the Vs of big data [1, 2]. The interpretation of remote sensing images is not straightforward and requires complex algorithms since their content depends upon various factors, e.g., the sensor resolution, the equipment unreliability, the type and amount of noise, etc. Furthermore, the increased data volume and demands of real-time applications require the use of high scalable and parallel processing approaches. While modern desktop computers and laptops having unprecedented performance, e.g., multi-core architectures and built-in accelerators, they are still limited in terms of computable problems due to their memory constraints and raw floating-point operations per second.

Having massive numbers of processors and memory available, multi-GPU systems can overcome these limitations and provide processing capacity that well exceed traditional laptops and work stations. Moreover, the utilized dedicated high-speed networks, such as InfiniBand, enable strong vertical and horizontal scaling of applications. Despite the impact of these new architecture on traditional simulation sciences, parallel computing is currently experiencing focus and advancements due to the current deep learning trend. Both of the latter domains influence each other in numerous ways [3] such as, among others, refreshed attention to hardware and performance engineering around tensor operations, the explorations of scalability boundaries as well as the envisioning simplified, parallel programming models. At the same time, deep learning has made revolutionary achievements for the analysis of remote sensing images [4] possible.

Nevertheless, there are major factors that prevent multi-GPU and -user systems from being the platform of choice for researchers developing new deep learning models. It starts with getting access and computing time on these machines, but goes well beyond that. Users who develop deep learning workflows want to focus first and foremost on the purpose and the realization of their analysis pipeline. This in turn requires them to be in full control of their programming library stack and underlying system. However, in multi-user systems administrators are usually in charge of the maintenance and supervision of the systems; users do not have privileges to install or modify software and can therefore not easily catch up with up-to-date libraries. Instead, a user is usually faced with either a long-pending installation request or a user-land compilation of their custom software, which needs to be repeated for every actively working scientist of the research collaboration. At the same time, users seek reproducible science through computational mobility [5], i.e., the possibility to restore a software environment as closely as possible to verify and continue past research. Containers have drawn a lot of attention in recent years in the parallel computing domain since they allow the simplification and acceleration of the application build and deployment process. Furthermore, for users working in the parallel computing and deep learning domains, containerization offers the benefits of scalability without performance penalties compared to traditional virtual machines. Gomes *et al.* [9] have proposed *uDocker* , a novel container tool that allows the execution of Docker containers without

doi:10.2760/848593

Table 1. Comparison of state-of-the art container technologies suitable for execution on multi-user systems. The displayed table is a heavily modified variant of the previous work from Priedhorsky et al. [6] and Kurtzer et al. [5].

	Docker [7]	Singularity [5]	Shifter [8]	Charlie Cloud [6]	*uDocker* [9]
Privilege model	Root daemon	SUID/UserNS	SUID	UserNS	chroot-like
Current production Linux distros support	✗	✓	✓	✗	✓
No privileged or trusted daemon	✗	✓	✓	✓	✓
Access to the host filesystem	✓	✓	✓	✓	✓
Support for GPU	✗ [a)]	✓ [b)]	✗	✗	✓ [b)]
Support for MPI	✓	✓	✓	✓	✓ [c)]
Pulling from Docker Hub	✓	✓	✓	✓	✓
No system admin intervention required	✗	✗	✗	✗	✓
No escalation of permissions	✗	✓ [d)]	✓	✓	✓
Works with all HPC schedulers	✗	✓	✗	✓	✓

[a)] Can be realized by installing nvidia-docker runtime
[b)] Experimental feature
[c)] Container MPI version has to match the HPC one
[d)] There was a number of high severity security issues in Singularity

the necessity for administrative privileges, i.e., no need to install additional system software. This paper describes the usage of *uDocker* [9] container tool for the development of an exemplary deep learning model for remote sensing images pixel-wise classification. The experimental results show that *uDocker* is comparable to a bare-metal installation, only entailing around a 1% computation time overhead, while simplifying the setup drastically.

2. BACKGROUND

The development of applications on shared multi-GPU systems is a difficult operation which requires that the system administrators build ad-hoc environments, i.e., software modules. For instance, a simple application upgrade can demand updating several environment modules. Furthermore, multi-GPU applications usually require running across multiple platforms and environments and utilize site-specific resources while resolving complicated software-stack dependencies. These are time-consuming tasks which add more work to the administrators, who have to maintain the multi-user systems and assure that the users have the tools and support to make the most efficient use of the computing resources.

Inspired by the shipping containers in inter-modal global transport, i.e., standardized containers that can be directly transferred with different shipping methods without any additional preparation, software containers utilize the same strategy. They are in many ways the next logical progression from virtual machines [10]. However, containers are a type of lightweight virtualization technology, which encapsulates system environments into standard units of software that are: portable, easy to build and deploy, have a small footprint, and low runtime overhead. As researchers started embracing containers for science, their usage within parallel computing environments grew as well. Despite all the issues of using containers in multi-user systems, they have been developed to meet their needs including security, MPI compatibility and GPU access. Since the introduction of Docker [7], the development of technologies associated with containers raised. Table 1 shows the most leading container technologies with their main features.

3. UDOCKER CONTAINER

uDocker is a software technology that allows the reuse and execution of Docker containers in user mode [9]. A container in turn is an isolated environment mimicking an operating system and its installed software. It is created by making use of layering file system, where every change made to the image, e.g., the installation of a software, adds a new layer to the image. These layers can then be shared and reused or further extended to a customized versions. In contrast to traditional virtualization technology, like virtual machines, containers are often referred to as light-weight, as they do not "pull-in" the entire operating system, but reuse the host operating system kernel when executed. This does not only reduce the memory footprint of such a container, but also reduces the computational performance impact.

While there is a plethora of containerization technologies currently being developed and researched on, first and foremost Docker, they often have a particular usage scenario in mind, requiring intervention of a priviliged user, e.g., administrator, for at least one step of the creation or execution of an isolated environment. In multi-user systems, especially with multiple GPUs used for deep-learning, the assumption about privileges does not hold in hindsight of security issues and direct use of containerization technology is not viable. At the same time, the operations requiring containers to have administrator privileges, are in most cases not needed for (scientific) deep learning application like in remote sensing. There-

doi:10.2760/848593

fore, *uDocker* attempts to offer a compromise between both worlds.

Through a second layer of virtualization technologies, like PTRACE, UserNamespaces or `libfakechroot` [9], it emulates as many container technology functions as possible in an unpriviliged userland environment. While actual privileged operation, like access to high-ports or password management, will obviously fail, enabling security by design, access to deep-learning essentials like GPUs is possible. *uDocker* offers multiple execution modes, where each is referring to a particular realization of the secondary virtualization technology—*P* uses PTRACE, *F* `libfakechroot`, *R* UserNamespaces and *S* Singularity as engine. This alongside numeric levels for the execution modes, e.g., *P1* or *P3*, allow the fine tuning of the *uDocker* for the particular execution scenario.

uDocker syntax is designed to be very similar to Docker's interface in order to allow users to reuse documentation material and container technology manager to transfer their knowledge. At the same time, *uDocker* is able to reuse openly published Docker containers, e.g., on DockerHub, enabling a rapid development cycle, custom extension and exchange with large community. A remote sensing scientist, who developed a new classification algorithm for example, may want to establish it either as a generally usable service or open-source it alongside a publication. In this scenario, the respective container can be created directly using *uDocker* on his experimentation device and then later shared with other scientists to verify or build on the results.

4. EVALUATION

4.1. Experimental Setup

The experiments have been performed on the LSDF setup, which is a single computer with all hardware available locally. Its configuration parameters are listed in Table 2. The operating system is a RedHat Enterprise Linux 7.5, CUDA Toolkit 9.0.176 and cudnn 7.0.5 library are installed system-wide. We first created virtual environment and run baremetal tests by means of Keras 2.2.2, TensorFlow 1.8.0 (GPU), and the neural network code[1] described in the next section. Versions of all relevant Python packages were fixed with `pip freeze`, so that exactly same versions are used in all the tests, including created docker image[2]. Note, that the utilized Python versions are slightly different in case of baremetal and the docker image: 2.7.5 and 2.7.12 respectively. *uDocker* is executed in 'F3' mode (Fakechroot) with '`--nvidia`' flag specified, devel branch of *uDocker* from GitHub is used.

[1] Source code: https://github.com/vykozlov/semseg-bids19
[2] https://hub.docker.com/r/vykozlov/semseg/, tag 'bids19-gpu'

Table 2. LSDF setup used in the experiments.

CPU	RAM	Nvidia GPU (driver version)
2 × Intel Xeon E5-2630 v3	128 GB	4 × K80, 12 GB (396.26)

4.2. Dataset and Deep Learning Model

The Vaihingen dataset [11] includes 33 orthorectified image tiles acquired by an aerial camera (i.e., infrared, green and red bands) over the town of Vaihingen (Germany) [3]. Since this dataset was released as a benchmark for a 2D semantic labeling contest, only 16 out of the 33 tiles are annotated (i.e., at pixel level with a spatial resolution of 9 cm). For the experiments, the annotated tiles that are used for the training and validation have ID= 1,3,5,7,11,13,17,21,26,28,34,37 and ID= 30,32, respectively. The semantic segmentation task involves the discrimination of 6 land-cover classes: *impervious surfaces* (i.e., roads, concrete surfaces), *buildings*, *low vegetation*, *trees*, *cars* and a class of *clutter* representing uncategorizable land covers (i.e., excluded in the prediction). The training data was randomly augmented using 90 degree rotations and horizontal and vertical flips.

The deep learning model is a 50-layer Residual Network (ResNet) [12] that was adapted into a Fully Convolutional Network (FCN) with connections from the last 3232, 1616, and 88 layers of the ResNet: ResNet50 FCN [4].

The model was trained using a random initialization for 20 epochs with 4166 samples per epoch (2083 original images and 2083 augmented) with a batch size of 16 using the Adam optimizer with Keras default settings (e.g., a learning rate of 0.001). The network takes 256×256 windows of data as input. To generate predictions for larger images, we made predictions over a sliding window (with 50% overlapping of windows) and stitched the resulting predictions together.

4.3. Results

The code allows to use more than one GPU for training by means of Keras' `multi_gpu_model` function, we therefore perform training on one, two, and four GPUs for every case. Each experiment is run three times under the same conditions in *baremetal* installation and via *uDocker* . In order to compare our results we used mean value and estimated standard error for the sample calculated based on the three runs. Every run consists of 20 epochs of training. Results for the total training time are shown in Table 3. As one can see, in

[3] http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html
[4] https://www.azavea.com/blog/2017/05/30/deep-learning-on-aerial-imagery/

doi:10.2760/848593

either case we see no statistically significant difference between *baremetal* and *uDocker* modes of running. There is also a clear performance improve in processing time when using four over one GPU (the higher variance is due to the data parallelization). The scaling with number of GPUs is however imperfect. This can be attributed to the communication overhead of the way Keras synchronizes weight gradients between multiple GPUs in the training's backpropagation step.

Table 3. Total training time of the neural network. Each result is an average of three runs with its standard error. Every run takes 20 epochs of training on either one, two, or four GPUs.

Number of GPUs	Training time, s	
	baremetal	*uDocker*
1	3710 ± 10	3730 ± 10
2	2390 ± 30	2360 ± 16
4	1860 ± 40	1880 ± 10

We note here, that *uDocker* also allows to pass environment settings at container instantiation phase, therefore one can e.g., specify which GPU card to use by setting CUDA_VISIBLE_DEVICES environment.

5. CONCLUSIONS

This paper describes the usage of the *uDocker* container tool within a multi-GPU system for the development of a deep learning classification task. *uDocker* allow to run the classifier in a Docker container without using Docker, root privileges and additional system software. It is run as a normal user without the intervention of the system administrators. The paper shows that researchers can adopt *uDocker* to facilitate the deployment of new analytical models and workflows on multi-user systems and enable scientific reproducibility. Furthermore, the experimental results demonstrated that the overhead introduced by the container is negligible when compared to an uncontainerized environment.

6. ACKNOWLEDGMENTS

uDocker is being developed within DEEP HybridDataCloud project, which receives funding from the European Union's Horizon 2020 research and innovation programme under agreement RIA 777435.
This project has received funding from the European Union's Horizon 2020 research and innovation program under the Grant Agreement No. 754304 DEEP-EST.
The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [11] http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html.

REFERENCES

[1] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, Y. Zhu, SunZhongyi, J. Shen, and Y. Zhu, "Big Data for Remote Sensing: Challenges and Opportunities," *Proc. IEEE*, 2015.

[2] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, and W. Jie, "Remote Sensing Big Data Computing: Challenges and Opportunities," *Future Generation Computer Systems*, vol. 51, pp. 47–60, 2015.

[3] T. Ben-Nun and T. Hoefler, "Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis," 2018.

[4] X. X. Zhu, D. Tuia, L. Mou, G. S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources," 2017.

[5] G. M. Kurtzer, V. Sochat, and M. W. Bauer, "Singularity: Scientific Containers for Mobility of Compute," *PLoS ONE*, 2017.

[6] R. Priedhorsky, T. C. Randles, and T. Randles, "Charliecloud: Unprivileged Containers for User-Defined Software Stacks in HPC," *SC17: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2017.

[7] D. Merkel, "Docker: Lightweight Linux Containers for Consistent Development and Deployment," 2014.

[8] D. M. Jacobsen and R. S. Canon, "Contain This, Unleashing Docker for HPC," *Cray User Group 2015*, 2015.

[9] J. Gomes, E. Bagnaschi, I. Campos, M. David, L. Alves, J. Martins, J. Pina, A. López-García, and P. Orviz, "Enabling Rootless Linux Containers in Multi-User Environments: The Udocker Tool," 2018.

[10] J. Smith and R. Nair, *Virtual Machines: Versatile Platforms for Systems and Processes*. 2005.

[11] M. Cramer, "The DGPF-Test on Digital Airborne Camera Evaluation Overview and Test Design," *PFG Photogrammetrie, Fernerkundung, Geoinformation*, vol. 2010, no. 2, pp. 73–82, 2010.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

AN EXPLORATION OF CONVOLUTIONAL RECURRENT NETWORKS FOR LARGE-AREA LAND COVER PREDICTION USING MODIS ARCHIVES

Alejandro Coca-Castro[1], *Marc Rußwurm*[2], *Mark Mulligan*[1]

[1]King's College London
Department of Geography
30 Aldwych, WC2B 4BG, London
{alejandro.coca_castro, mark.mulligan}@kcl.ac.uk

[2]Technical University of Munich
Chair of Remote Sensing Technology
Arcisstraße 21, 80333, Munich
marc.russwurm@tum.de

ABSTRACT

Current pipelines for *Land Use and Land Cover (LULC)* classification partially exploit all available spatiotemporal information offered by the Earth Observation archives. Thanks to the advances in computing resources, we explored the feasibility of an architecture based on convolutional recurrent networks for large-area LULC classification. We tested it across the Brazilian Amazon biome using MODIS archives providing surface reflectance time series data and multitemporal land cover maps at $500m$ for 2009. According to the experimental sets, the architecture outperformed baseline methods such as Random Forest and Support Vector Machines. Moreover, the trained architecture using $500m$ MODIS data produced spatially consistent predictions over $250m$ MODIS data under the same distribution of bands and observations periodicity. We conclude with a set of key elements and challenges to consider for future implementation of the method assessed, particularly for LULC classification tasks.

Index Terms— land use and land cover, recurrent neural networks, deep learning, time series classification, the Brazilian amazon

1. INTRODUCTION

A variety of space-born sensors monitor the Earth's surface since many decades. Initially driven by US-American advances in *Earth Observation (EO)* research, the data of some satellites have been made available to the public since the early 70s. In recent years, Europe's Copernicus Program has taken over incentive and an increasing number of Sentinel satellites collect data at large scale.

EO satellite data provide the basis of data-driven research at global scale. With the advance of deep learning, a variety of research fields and applications, such as Land Cover and Land use *(LULC)* Classification, Vegetation Monitoring, Change Detection, or Atmosphere Science, profit from this abundance of data.

In this work, we explored *Convolutional Recurrent Neural Networks (ConvRNNs)* for the task of multitemporal LULC using Remote Sensing imagery. We used MODIS data that is available since the early 2000s and allow a continuous mapping of our large area of interest in the Brazilian Amazon Biome.

2. RELATED WORK

LULC classification is one of the most common task of the EO community and crucial for many scientific and operational applications. To date, the most widely accepted LULC classification methods have limited capabilities to fully use all available spatiotemporal information offered by the EO archives at reasonable complexity [4].

Thanks to the advances in computing and storage resources, the EO community have recently started using *Recurrent Neural Networks (RNNs)* architectures for LULC classification and change tasks [4]. RNNs are specifically designed to model time series data. Therefore, they are good candidates to cover the limitations by traditional supervised *machine learning (ML)* based methods.

In order to complement the aforementioned efforts regarding the feasibility of RNNs for analysing sequential data as the EO observations, our aim was first to evaluate the capability of one of them, applied originally for crop classification using Sentinel 2 data [4], but for a large-area prediction based on MODIS satellite time series and multi-temporal LULC maps available at $500m$. Then, under the assumption of the spectral similarities between two MODIS $500m$ and MODIS $250m$ products, we visually inspected the capability of the trained RNNs using MODIS $500m$ data for predicting over MODIS $250m$ data.

3. METHODOLOGY

RNNs iteratively encode a sequence of T observations $\boldsymbol{x} = \{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_T\}$. Since the data is processed sequentially,

The author acknowledges the International Center of Tropical Agriculture (CIAT) for granting the access to the resources used for processing data and models linked with this work.

deeper representations are created through sequential updates $\boldsymbol{h}_t \leftarrow \boldsymbol{x}_t, \boldsymbol{h}_{t-1}$ with context information from the previous representation $\boldsymbol{h}_{t-1}$. Hence, few stacked recurrent layers produce deep high-level representations for the classification task [2]. This recurrent networks design, however, struggles with learning long-term relationships due to vanishing and exploding gradients when back-propagating corrections though time. This has been addressed by additional cell gates, initially in *Long Short-Term Memory (LSTM)* networks and later in *Gated Recurrent Units (GRU)*. These gates control the gradient flow through time and enable learning of long temporal relationships. The initial formulation of Recurrent neural networks was taylored towards temporal sequence processing, as is commonly used in natural language processing. Here, trained weights are applied by matrix multiplication to the input $\boldsymbol{x} \in \mathbb{R}^d$ in a fully-connected fashion. To process spatiotemporal data these matrix multiplications can been replaced by convolutions. With this, inputs $\boldsymbol{x} \in \mathbb{R}^{w \times h \times d}$ of given height h, width w and depth d can be processed and local pixel neighborhoods are considered. These spatiotemporal implementations are referred to as *convolutional LSTM (ConvLSTM)* [5] or *convolutional GRU (ConvGRU)* [1] throughout this work.

4. DATA

The study area corresponds to the Brazilian Amazon biome, a region covering roughly 4.1 million km^2 in South America.

4.1. Satellite data

12-month time series of MODIS Terra 16-day composite $250m$ (*MOD13Q1*) and MODIS surface reflectance 8-day composite $500m$ (*MOD09A1*) images were used for the study area for the January through December period of 2009. MOD09A1 provides surface reflectance in 7 bands (blue, green, red, Near-Infrared (NIR), Short-Wave Infrared 1 (SWIR1), SWIR2, SWIR3) with resolution of $500m$. MOD13Q1 included data for vegetation indices, and surface reflectances from bands blue, red, NIR, and SWIR3 with $250m$ resolution. Due to a part of this work aimed to inspect the feasibility of the trained models using MODIS $500m$ data to predict over MODIS $250m$ data, certain assumptions were made. These assumptions included to retain 8-day intervals of MODIS $500m$ to 16-day intervals of the MOD13Q1 $250m$ product. Additionally, the shared bands between the products (blue, red, NIR, and SWIR3) were used as input features. Both datasets were extracted from the *Google Earth Engine (GEE)* platform which also includes the $500m$ MODIS land cover product used to extract the labelled dataset.

4.2. Ground truth

The LULC ground truth labels were obtained from the MODIS land cover product (MCD12Q1). For this study, the MCD12Q1 International Geosphere-Biosphere Program (IGBP) classification scheme, which classifies pixel into one of 17 classes was used. We worked with the Collection 5.1 MCD12Q1 product, which covers the years 2001-2012 at a spatial resolution of 500 m.

We aimed to reduce label noise by identifying *reliable* pixels. These pixels consisted of those with unchanged land cover during a given time period. For instance, *reliable* pixels of 2009 consisted of those unchanged from MODIS LULC maps between 2008, 2009 and 2010.

4.3. Data partition

We divided the area of study in blocks of 207 pixels per 207 pixels (or $\sim 100km$ per $\sim 100km$) based on the $500m$ MODIS data. This size matches multiples of 23 pixels which correspond to the patch size used to train the deep neural networks described in the next section of experimental sets. The 4:1:1 ratio proposed by [4] was maintained to randomly assign the training, validation and evaluation blocks across the study area (Fig.1). This split ratio represented approximately 12.1, 3.4 and 3.4 millions of *reliable* pixels for training, validating and evaluating the deep learning methods.

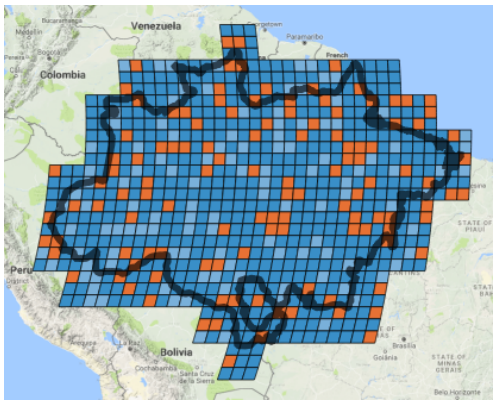


Fig. 1. Illustration of the study area (delimited by the black line) with non-overlapping 207 pixels x 207 pixels blocks partitioned for training (blue), evaluation (light blue), and validation (orange).

5. EXPERIMENTS

We compared the architecture initially proposed by [4] using ConvRNNs cells with two common machine learning methods, random forest and support vector machines (here both of them referred as baseline methods).

RF an SVM: Here the input is a vector with values of band blue, red, NIR, and SWIR3 extracted from MODIS $500m$ with a sample size of 100 pixels per class per 16-days interval MODIS image. For this work, each sample to train the models represented a flattened vector of the reflectance bands (4) and number of time intervals (23). We implemented the

RF and SVM models within GEE. The main hyperparameters by model were determined by manual grid search based on the performance on the validation dataset. The SVM model was trained with a *radial basis function (RBF)*, *gamma* value of 0.5 and *cost* parameter set to 10. For RF, it was trained with 20 *number of trees (ntress)*. The remaining parameters (*variablesPerSplit* and *minLeafPopulation*) were set by their default value in GEE (0 and 1, respectively). Due to the perfomance of RF and SVM vary by changes in the input data, ten runs were generated by model using different samples (or seeds) distributed across the training/testing blocks.

ConvRNNs: We oriented our implementation on related work and slightly modified the network configuration. The input vector consisted of the bands band blue, red, NIR, and SWIR3. Additional to the spectral information, the day-of-year of the individual observations was added as matrix to the input tensor. ConvRNNs were trained on 23 pixels 23 pixels tiles with separate experiments according to two types of RNN cells, ConvLSTM or ConvGRU with 128 *recurrent cells*, *batch size* equals to 15, a *single* layer, and *krnn/kclass* of 3 pixels. We applied batch normalization (BN) to the input. The type of BN used normalizes a tensor by mean and variance [3]. The ConvRNNs-based experiments using either ConvLSTM or ConvGRU cells were trained on a Tesla M60 GPU for 60 epochs.

5.1. Prediction and evaluation

The prediction strategy varied by method. For both RF and SVM classifiers, the predictions were computed by single image. The set of 23 classified maps were then aggregated into a single map using the maximum voting. Regarding the ConvRNNs architectures, they were set under a many-to-one approach which allows producing a single map derived by the sequence of images seen.

The final accuracy metrics were calculated using the pixels located over the evaluation blocks, which remained untouched during all experiments. To facilitate the interpretability of the metrics per class, these metrics were computed over the predicted and target maps aggregated to a customised land cover nomenclature of 7 major classes (Woodland, Grassland, Cropland, Crop/Nature mosaic, Water/Wetland areas, Artificial surfaces, Bare/Spare vegetation). Classes which were aggregated to the major classes included *Evergreen needleleaf, Evergreen broadleaf forest, Deciduous needleleaf forest, Deciduous broadleaf forest, Mixed forest, Closed shrublands, Closed shrublands and Open shrublands* as Woodland; *Woody savannas, Savannas and Grasslands* as Grassland; *Snow and ice, Water, Wetlands* as Water/Wetland. The remained classes of the 17 IGBP classification system remained with their original name (*Cropland* and *Cropland/natural vegetation mosaic*), except for *Urban and built-up* and *Barren or sparsely vegetated* renamed to to Artificial and Bare/Spare, respectively.

6. RESULTS AND DISCUSSION

In this section, we first assess the performance of the architecture using ConvRNNs in regards to the baseline classification methods. Then, we visually inspected the spatial patterns of predictions over $500m$ and $250m$ satellite data by each model trained with the $500m$ MODIS data (Fig.2).

Table 1 shows the values of five evaluation metrics computed between the true 'reliable' LULC pixels and the predictions of the ConvRNNs and baseline methods. The ConvRNN structures presented higher values in all metrics in comparison with the baseline methods.

Table 1. Performance evaluation of two baseline ML-based methods, RF and SVM, in comparison to the ConvLSTM and ConvGRU models trained and evaluated using MODIS $500m$. The metrics were weighted by the frequency of samples in each class to avoid biases of the non-uniform class distributions. The standard deviation of the multiple runs by baseline method is denoted by $\pm$.

Measure	RF	SVM	Conv LSTM	Conv GRU
accuracy	96.6±0.2	98.2±0.2	99.5	99.3
kappa	78.9±1.0	71.3±2.5	95.1	94.6
precision	94.8±0.2	92.9±0.4	98.5	98.5
recall	93.5±0.3	92.2±0.6	98.5	98.4
f1-score	94.1±0.2	91.8±0.7	98.5	98.4

Due to the kappa metric showed the highest difference between the baseline and ConvRNN structures, this metric was computed by major land cover class (Table 2). Overall, both ConvRNN architectures outperformed the baseline methods for most of the classes, except by the bare/sparse class. This class was underrepresented (less than 600 samples out of 15.5 million of samples within the training/testing dataset). In particular, ConvRNN methods produced classification improvements for the cropland/nature mosaic and artificial classes.

Table 2. Kappa values (%) per major land cover class by method for models trained and evaluated using MODIS $500m$. The standard deviation of the multiple runs by baseline method is denoted by $\pm$.

Major class	RF	SVM	Conv LSTM	Conv GRU
Woodland	85.6 ± 0.6	75.6 ± 2.5	97.2	97.1
Grassland	78.9 ± 1.4	69.2 ± 3.4	96.2	95.7
Cropland	68.2 ± 3.8	69.1 ± 4.0	92.6	91.5
Crop/Nat.	15.5 ± 1.1	10.0 ± 2.0	60.0	58.5
Water/Wet.	69.3 ± 2.6	74.9 ± 2.3	88.5	88.5
Artificial	4.2 ± 0.3	3.3 ± 0.4	32.8	46.7
Bare/Spare	47.0 ± 18.9	45.5 ± 20.4	1.7	1.8

doi:10.2760/848593

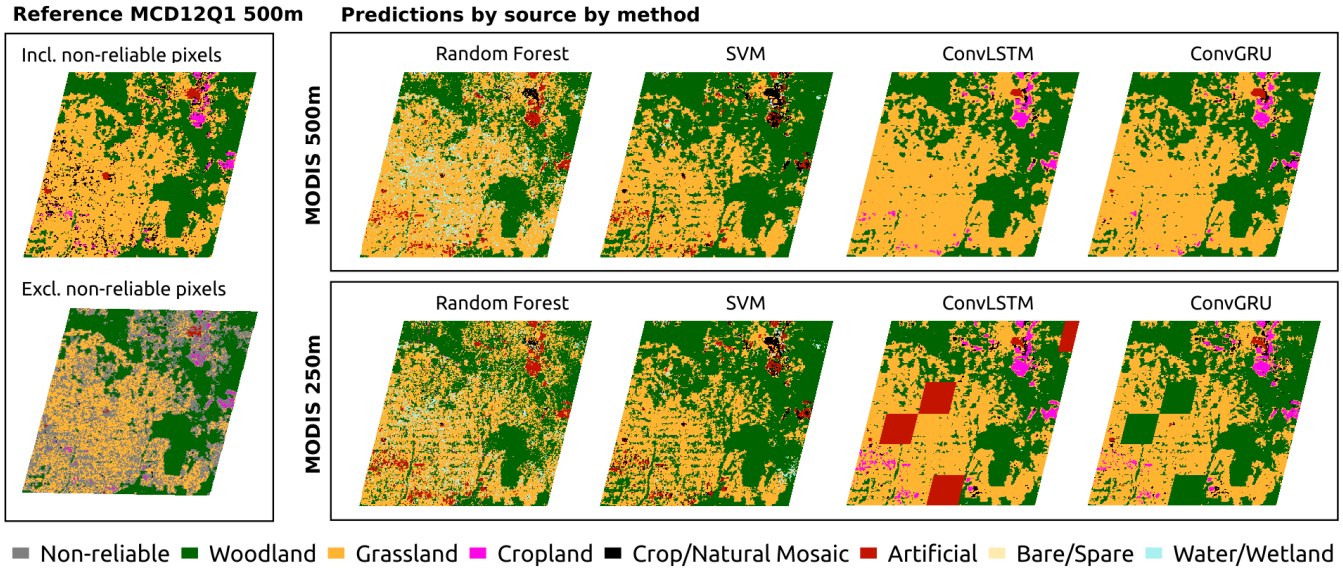


Fig. 2. Visualization of the spatial structure of the predictions by method and source of major land cover classes within a single evaluation block of 207 pixels x 207 pixels. The block represents an area dominated by grassland with certain patches of artificial and cropland pixels located at the upper right of the block. Red and green mini-blocks (patches of 23 pixels x 23 pixels) of the ConvRNN models' predictions over MODIS 250m are related with high percentages of missing observations.

To examine the spatial patterns, we exported the predictions over a representative evaluation block of 207 pixels x 207 pixels as depicted in Fig.2. In general, for both resolutions of MODIS archives, the baseline implementations tend to confuse classes, mainly cropland with artificial and artificial with grassland. SVM produced slightly less noise pixels of water/wetland class than RF within the grassland area. In comparison to the baseline methods, the ConvRNN architectures precisely retained most of the spatial patterns of the artificial and cropland land cover classes.

The presence of artifacts, which represent areas of 23 pixels per 23 pixels, for the predictions over MODIS 250m data (Fig.2) were attributed to a high percentage of missing data from this particular product. The current source code of the method only works with images without missing or masked values and further research might be done in this regard.

7. CONCLUSION

This work evaluated the performance of recurrent networks with SVM and RF for a large-area classification using MODIS archives. In comparison to the baseline methods, the end-to-end setting of the architecture allows learning both feature extraction and classification solely from the provided data. The quantitative and qualitative outcomes showed that the assessed RNNs can be used for large-area spatiotemporal information extraction.

Since MODIS provides more than a decade of EO observations, we will expand this classification method to a multi-year evaluation. Further work is also needed specially identify how the model can handle missing input data. Overall, we would like to utilize this method on data acquired by the recently launched Sentinel 3 satellite.

REFERENCES

[1] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *arXiv preprint arXiv:1511.06432*, 2015.

[2] Alex Graves. Supervised sequence labelling. In *Supervised sequence labelling with recurrent neural networks*, pages 5–13. Springer, 2012.

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. URL http://arxiv.org/abs/1502.03167.

[4] Marc Rußwurm and Marco Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4), 2018. ISSN 2220-9964. doi: 10.3390/ijgi7040129. URL http://www.mdpi.com/2220-9964/7/4/129.

[5] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Advances in Neural Information Processing Systems 28*, 2015. URL https://arxiv.org/pdf/1506.04214v2.pdf.

doi:10.2760/848593

CEOS ANALYSIS READY DATA FOR LAND – SUPPORTING THE EARTH OBSERVATION COMMUNITY TO GET THE BEST VALUE FROM THE BIG DATA WAVE FROM SPACE

Andreia Siqueira[a], Adam Lewis[a], Medhavy Thankappan[a], Zoltan Szantoi[b,c], Philippe Goryl[d], Takeo Tadono[e], Ake Rosenqvist[f], Jonathon Ross[a], Steven Hosford[d], Susanne Mecklenburg[d], Kurtis Thone[g] Steven Labahn[h], Brian Killough[i], Jennifer Lacey[h]

[a]Geoscience Australia, Cnr Jerrabomberra Ave and Hindmarsh Drive, Symonston, Australia, 2609
[b]European Commission, Joint Research Centre, Directorate D - Sustainable Resources, Ispra 21027, Italy
[c]Department of Geography & Environmental Studies, Stellenbosch University, Matieland 7602, South Africa
[d]European Space Agency, Largo Galileo 1, Italy 00044
[e]Japan Aerospace Exploration Agency, Tsukuba, Japan 305-8505
[f]solo Earth Observation, Tokyo, Japan 104-0054
[g]NASA Goddard Space Flight Center
[h]USGS EROS, Sioux Falls, SD, USA 57198
[i]NASA Langley Research Center, Hampton, VA, USA 23681

ABSTRACT

Public and private agencies have been committed to address the "big data" challenge by producing Analysis Ready Data products (ARD) for their users. The ARD products are enabling users to get first hand satellite data that are 'ready to use' for a wide range of applications, including time-series analysis and the way forward to multi-sensor interoperability. The Committee on Earth Observation Satellites (CEOS) is leading the CEOS Analysis Ready Data for Land (CARD4L) initiative, which is focused on a framework implementation and the development of Product Family Specifications (PFSs) across the optical, thermal and radar domains. CARD4L aims to enable non-expert users access to products that have been processed 'far enough' to be suitable for immediate analysis for a range of applications, while ensuring they are not too specific to only be used for a particular area. The aim of this paper is to give a brief overview on the CARD4L Framework and to introduce the Product Alignment Assessment (PAA) process.

Index Terms— Analysis Ready Data, CARD4L, Product Family Specifications, Product Alignment Assessment, Earth Observation

1. INTRODUCTION

Many satellite data users lack the expertise, infrastructure, and internet bandwidth to efficiently and effectively access, pre-process, and utilize the growing volume of space-based data for local, regional, and national decision-making. Even sophisticated users of Earth Observation (EO) data typically invest a large proportion of their effort into data preparation. This is a major barrier in realizing the full potential and the successful utilization of space-based imagery data. This barrier presents a major obstacle to mainstreaming the use of EO data, and a threat to the success of major global and regional initiatives supported by CEOS. As data volumes grow, this barrier is becoming more significant for the majority of users.

Countries and international organizations have expressed a desire for support from CEOS to facilitate access to and processing of satellite data into CEOS Analysis Ready Data for Land (CARD4L) products [1]. Systematic and regular provision of CARD4L is expected to greatly reduce the burden on global satellite data users and, as a direct consequence, boost data use. The provision of this data is possible through many options including systematic processing and distribution, processing on hosted platforms, and processing via toolkits provided to users.

CARD4L products are intended to be flexible, accessible and suitable for a wide range of users and a wide variety of applications, including time-series analysis and multi-sensor interoperability. They are also intended to support rapid ingestion and exploitation via high-performance computing, cloud computing and other data architectures.

CARD4L will be an important enabler of the Open Data Cube (ODC) initiative [2]. Through CARD4L, users will be able to locate products that are suitable for ingestion into Data Cubes [3], and will have confidence that these different CARD4L products will limit as far as possible barriers to interoperability.

2. CARD4L FRAMEWORK

Fig. 1 presents the overall CARD4L Framework and its three elements; a) Definition, b) Product Family Specifications and c) Product Alignment Assessment.

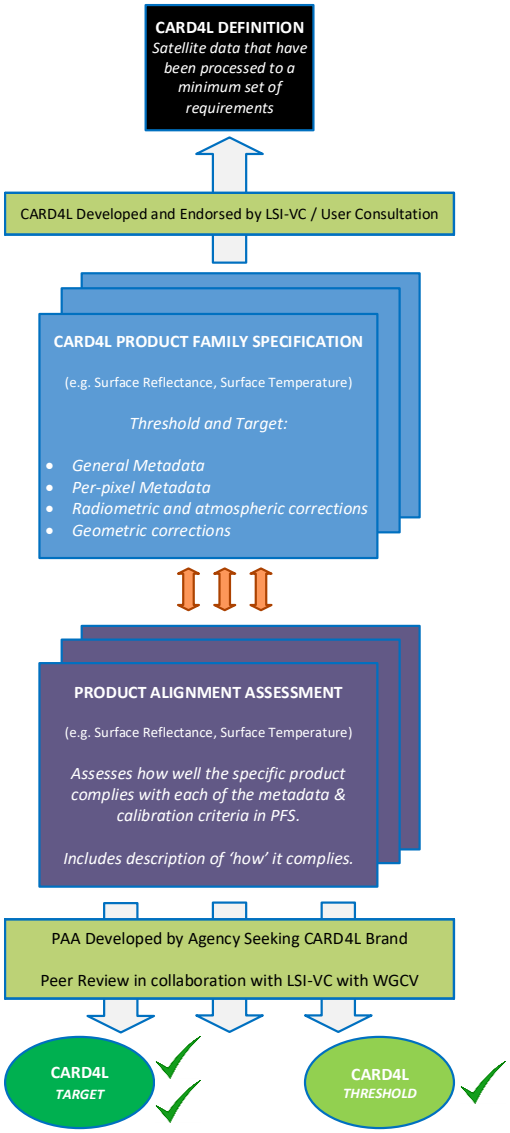


Fig. 1 CARD4L Framework Components: Definition, Product Family Specifications and Product Alignment Assessment.

2.1. CARD4L Definition

The definition of CARD4L is not exclusive or prescriptive (Fig. 2). It is expected that a range of data products will be produced by CEOS agencies [4] to meet the needs of the diverse user community, and that these products will be fully "analysis ready" for their users.

CARD4L DEFINITION

CARD4L are satellite data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets.

Fig. 2 CARD4L Definition.

2.2. CARD4L Product Family Specifications

The Product Family Specifications provide details on what type of corrections the data providers need to carry out in order to deliver EO data in an analysis ready form to their user community.

A number of parameters described in the PFSs are used to assess the minimum requirements (threshold) of CARD4L, which differ for each sensor type. Any higher level derived data products (for instance composites or indices), or additional corrections that meet or exceed these minimum requirements, may also be considered CARD4L. Beyond the minimum requirements, there is a "target" requirement, where the products labeled as such will be evaluated based on more stringent evaluation criteria. All PFSs, independent of the sensor type, include requirements for:

- General metadata: These requirements are metadata records describing a distributed collection of pixels. The collection of pixels referred to must be contiguous in space and time. General metadata should allow the user to assess the overall suitability of the dataset.
- Pixel-level metadata: Per-pixel metadata should allow users to choose between observations on the basis of their individual suitability for an application, and include 'quality flags'. Whether the metadata are provided in a single record relevant to all pixels, or separately for each pixel, is at the discretion of the data provider. Similarly, the mechanism or form of the per-pixel metadata (additional data bands, mask layers, etc.) is open to the provider.
- Radiometric corrections: Radiometric corrections are designed to perform adjustments for sensor/instrument gains, biases, offsets and adjustments for sensor viewing angle with respect to the pixel position on the surface. These type of corrections should allow the majority of users to apply the data directly rather than, in general, undertaking these steps themselves. Ideally, CARD4L should provide geophysical quantities such as surface reflectance, temperature, or backscatter amplitude facilitating the use of observations from multiple platforms and sensors.
- Geometric corrections: Geometric corrections are designed to establish ground position, to take into account terrain (ortho-georeference) and ground control points and to assess absolute position accuracy. Geometric calibration allows products to be used with other spatial data, and in particular to be 'stacked' as time-series. Adjustments for ground variability typically use a Digital Elevation Model (DEM).

PFSs covering optical sensors only include additional requirements for:

- Solar and view angle correction: These include adjustments for local solar and view angles with respect to the pixel position.

doi:10.2760/848593

- Atmospheric correction: These include adjustments for atmospheric effects (absorption and scattering) due to water vapor, ozone, molecular scattering, and aerosols.

Moreover, PFSs covering radar sensors only contain additional requirements on:

- Geometric Corrections: Similar to the optical imagery, the data is orthorectified using a DEM to compensate for the terrain distortion introduced by the slant looking geometry of the SAR. The geometric calibration of the imaging system and the precision of orbit ephemeris allow the accurate stacking for time-series analysis.
- Per-pixel metadata: Per-pixel metadata for providing information like shadow, layover area or no-value pixel. By the nature of the SAR processing that compresses thousands of raw lines into one, no per-pixel quality data is envisaged, so far.
- Radiometric Corrections: Classical adjustments for antenna/instrument gains, biases, offsets are performed. In addition, radiometric terrain correction is performed to remove the radiometric bias (slope dependent) introduced by the topography allowing to efficiently combine data from different geometries.

Currently there are three established PFSs; 1) Surface Reflectance, 2) Surface Temperature and 3) Radar Backscatter – *please see brief descriptions below*. Additional PFSs in the Synthetic Aperture Radar (SAR) domain are currently being assessed and developed by CEOS experts.

1. Surface Reflectance (SR): Produced using data collected with multispectral sensors operating in the VIS/NIR/SWIR wavelengths. These sensors typically operate with ground sample distance and spatial resolution in the order of 10 to 100m, however, the specification is not inherently limited to this resolution and can be applied to higher and lower resolution data.
2. Surface Temperature (ST): Produced using data collected with multispectral sensors operating in the thermal infrared (TIR) wavelengths. These sensors typically operate with ground sample distance and spatial resolution in the order of 10 to 100m.
3. Radar Backscatter (RB): Produced using data collected with Synthetic Aperture Radar (SAR) sensors, operating in the microwave (MW) domain. The CARD4L backscatter product is developed to facilitate a growing range of applications of radar data that draw on time-series observations from different instruments, radar bands and/or observation modes. A key objective is to expand the use of radar remote sensing beyond the current expert user community to enable a set of new, generalist users to access and apply SAR data in geographical analyses to produce improved products.

2.3. CARD4L Product Alignment Assessment Process

Product Alignment Assessment is the process that is intended to help data providers to 'self-assess' their product alignment with the CARD4L specifications. PAA will also allow an independent assessment and a peer-review to be done. Thus, PAA process is the quality assurance element of the CARD4L framework. In the absence of the PAA, any product could potentially be badged as CARD4L and the value of the concept would be lost.

Fig 3. presents the detailed CARD4L PAA process, which is divided into 6 steps. For instance, if a data provider would like to have a dataset to be considered as a CARD4L dataset, the following steps will take place:

1. The data provider will contact the Land Surfacing Imaging Virtual Constellation Secretariat (LSI-VC SEC) to indicate their interest to submit their dataset through the PAA process. *One of the responsibilities of LSI-VC SEC is to monitor the incoming requests and keep a registry of the proposals*,
2. The LSI-VC SEC will register the proposal and will connect the data provider to a product-specific LSI-VC Point of Contact (POC) – *A LSI-VC POC will be identified for each PFS. The POC will be responsible for verifying the data provider's self-assessment and for obtaining feedback from the Working Group on Calibration & Validation (WGCV) on the product's Calibration/Validation process (the peer review process)*,
3. The LSI-VC POC will interact with the data provider on their documentation using the PFS as the basis for the assessment – *this step is called self-assessment*,

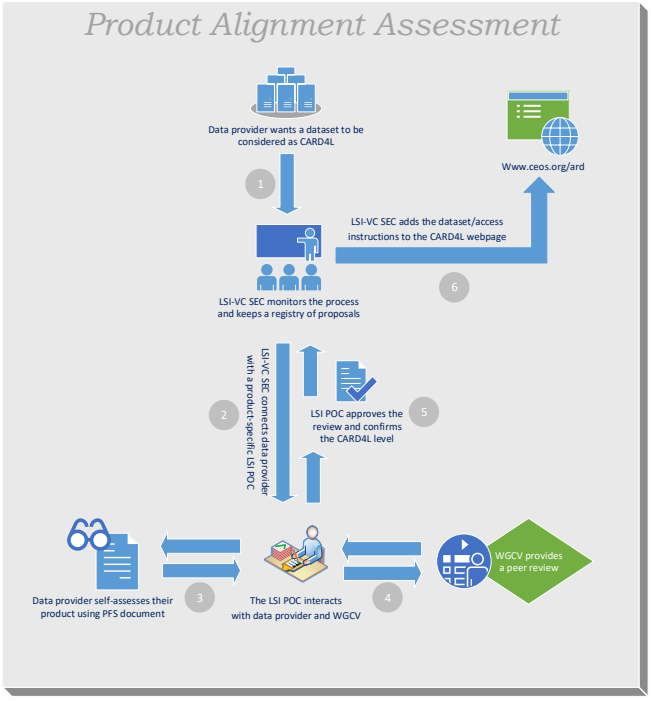


Fig. 3 CARD4L Product Alignment Assessment Process.

4. The LSI-VC POC will also interact with the WGCV POC, to facilitate the peer review assessment. The interactions in steps 3 and 4 are expected to be contemporaneous and may take several interactions to be completed – *The WGCV has proposed a framework for the "WGCV peer review process", in support of the PAA.*
5. Once the self-assessment and the WGCV peer review processes are completed, the LSI-VC POC will approve the review and confirm the CARD4L level (threshold or target level). Then the outcome of the review will be communicated to the LSI-VC SEC.
6. The LSI-VC SEC will then publish the outcome of the review on the CEOS Analysis Ready Data webpage (http://ceos.org/ard/).

Therefore, a particular product will be approved as CARD4L dataset when:

- The product has been assessed as meeting CARD4L requirements by the agency responsible for production and distribution of the product; and
- The assessment has been peer-reviewed by the CEOS Land Surface Imaging Virtual Constellation (LSI-VC) in consultation with the CEOS Working Group on Calibration and Validation (WGCV).

Whilst these mechanisms are within the CEOS community, the Framework is a public document allowing the wider community to produce analysis ready products, assess those against the PFS, and to independently assess product alignment through the PAA process.

3. THE WAY FORWARD

The next steps foreseen for the CARD4L implementation phase are:

- Pilot the CARD4L data provider's self-assessment and production process for the Surface Reflectance, Surface Temperature and Radar Backscatter PFSs.
- Develop the WGCV "peer review" process to evaluate datasets that are candidates to become a CARD4L dataset.
- Promote and communicate information on CARD4L products including implementing discoverability of CARD4L products.
- Identify data products that are on-track to become CARD4L.
- Engage the CEOS WGCV to define Quality Assurance (QA) protocols and cross-validation projects across all product families.
- Examine extending the CEOS ARD framework to include the ocean and atmosphere domains.
- Continue reviewing the PFS through a controlled process, e.g. go through a cycle to gather feedback from technical experts, collate the feedback, submit the proposed changes through a drafting process and accept them into the PFS document.

REFERENCES

[1] CEOS Analysis Ready Data for Land – Description Document (http://ceos.org/document_management/Meetings/Plenary/30/Documents/5.5_CEOS-CARD4L-Description_v.22.docx)
[2] www.opendatacube.org
[3] Strobl, Peter, Peter Baumann, Adam Lewis, Zoltan Szantoi, Brian Killough, Matthew B. J. Purss, Max Craglia, Stefano Nativi, Alex Held, and Trevor Dhu. "The Six Faces of the Data Cube." In Proc. of the 2017 Conference on Big Data from Space (BiDS'17). Toulouse, France: Luxembourg: Publications Office of the European Union, 2017, 2017. https://doi.org/10.2760/383579
[4] http://ceos.org/about-ceos/agencies/

SENTINEL-2 AND LANDSAT-8 ANALYSIS READY DATA: TOWARDS A SERVICE PROTOTYPE FOR ON-DEMAND PROCESSING USING THE ESA RESEARCH AND SERVICE SUPPORT

Roberto Cuccu[1,2], *José Manuel Delgado*[1,2], *Giovanni Sabatino*[1,2], *Mauro Arcorace*[1,2], *Giancarlo Rivolta*[1,2], *Joost van Bemmelen*[3], *Steven Hosford*[4], *Ferran Gascon*[3]

1 ESA – Research and Service Support (ESA-RSS)

2 Progressive Systems s.r.l.

3 European Space Agency (ESA)

4 Centre national d'études spatiales (CNES)

ABSTRACT

The era of Copernicus Sentinel missions for Earth Observation started with the launch of Sentinel-1A in April 2014 and has been providing an unprecedented amount of data freely available to EO users. With an increasing number of EO missions and initiatives from National and International Organizations, more and more innovative applications are becoming possible. More often EO data users need to analyze data by running their algorithms on what it is usually referred to as "Analysis Ready Data" (ARD), which in the simplest interpretation might be thought as data with a standardized product regardless of the acquisition sensor. In this paper the features of an Analysis Ready Data On-Demand demonstrator service for Sentinel-2 and Landsat-8 implemented by ESA-Research and Service Support (RSS) will be presented and discussed. The tool provides a web interface where users can select L1C data, visualize their footprint on a map, select one of the available Atmospheric Correction Algorithms, the output projection and format and submit a processing task On-Demand.

***Index Terms*—** Analysis Ready Data, Sentinel-2, Landsat-8, Data preparation, Atmospheric Correction Algorithm

1. INTRODUCTION

The growing number of EO satellite missions and the ease of access data in the last years, have lowered the barrier in EO data exploitation, in turn stimulating new ideas for applications and new initiatives among the community. EO data usage has increased enormously and EO data, today, is no longer the prerogative of EO experts. EO data is being largely used in sectors such as maritime surveillance, agriculture industry and in general in all types of human activity monitoring. Governmental organizations and public institutions are gaining awareness of the advantages derived from EO data usage for their purposes. The drawback is that more and more non-expert users will need to preprocess and analyze EO data from different missions, which often come in different formats and projections, are referred to a different grid or which have been obtained with different preprocessing algorithms. In other words while more data are becoming available, such data are not ready for immediate use and even sophisticated users typically invest a large portion of their time in data preparation.

In response to such a need, several initiatives are arising with the purpose to provide data ready for immediate analysis or that require a minimum set of operations to be ready. The CARD4L [1] initiative, for example, aims at providing specifications and definitions of Analysis Ready Data for Land products. There is some effort in this direction also by USGS [2] which proposes U.S. Landsat Analysis Ready Data consistently processed according to the scientific standards and level of processing required for immediate exploitation in monitoring landscape change. In addition, several ongoing projects which work on the harmonization of the Landsat and Sentinel-2 products in terms of bandwidth and spectral bands [3], may help the utilization of both datasets in synergy.

2. ANALYSIS READY DATA

The definition of ARD, is still matter of discussion due to the fact that products ready for some applications might not be considered as ready for other applications: some applications may require different map projections, different metadata or format. From the user-perspective, ARD probably means data as close as possible to the products useful for analysis: containing already the bands in the same projection, same format and same resolution, regardless of the specific sensor. On the other hand projection, format and resolution are application dependent, hence the question: is there a unique definition of ARD product good for all? The answer is no.

Then, how to find a way to select common criteria for large-use ARD?

Currently, USGS makes available Landsat products in ARD format limited to the USA. Similarly, Sentinel-2 L2A products are systematically provided over a limited area covering Europe. Hence, users who need atmospherically corrected products outside those areas have to find a way to process the data themselves, using one of the most common open source algorithms or other commercial solutions.
In this paper we present an example of ARD On-Demand demonstrator prototype made available by ESA-RSS [4].

2.1. The case of Sentinel-2 and Landsat data for Land Applications

Land applications domain in Optical Remote Sensing plays a fundamental role: with a spatial resolution of the order of 10 m and a revisit time typically of few days, Landsat and Sentinel-2 in synergy offer an unprecedented opportunity for applications in land cover studies, agricultural fields monitoring and inland or near-shore water monitoring (in general landscape monitoring). For some applications, which are relatively new to Earth Observation, specific processing expertise and know-how owned by the users could be limited. Therefore the possibility to have ARD On-Demand could really make the difference: it would be possible to generate ARD over a specific Area of Interest simply configuring some basic parameters without the need to know in detail how an algorithm actually works.

2.2. ARD On-Demand demonstrator implemented by RSS

At the beginning of summer 2018, ESA-RSS has been requested to provide a feasibility study and to implement an Analysis Ready Data On-Demand demonstrator which takes in input Sentinel-2 L1C and Landsat-8 L1 products. The demonstrator should have a web based GUI allowing registered users to select the input dataset (Landsat or Sentinel), an Area of Interest, a time-frame, the Atmospheric Correction Algorithm to use, the type of projection and output format. In addition the user should be able to set some configuration parameters specific for each Atmospheric Correction Algorithm. The algorithms selected for the atmospheric correction are:

- Sen2Cor [5]
- LaSRC [6]
- iCOR [7]
- MAJA [8]

It is important to note that each of the aforementioned processors provide by default an output with different format, type and bit depth, hence being necessary to harmonize the results to provide as much as possible a uniform atmospheric corrected product.
In Tab.I some of the most relevant characteristics of each software, such as main language employed, output format, etc…, are summarized. The processing time has been obtained testing the processors on the same input product. The larger processing time for MAJA is due to the fact that MAJA uses a multi-temporal approach.

TABLE 1. CHARACTERISTICS OF THE SOFTWARE INTEGRATED IN THE ESA PROCESSING ENVIRONMENT.

Algorithm name	Sen2cor	LaSRC	iCOR	MAJA
Provider	ESA	USGS – NASA/GSFC	VITO	CNES
Developed by	CS/ Telespazio	USGS – NASA/GSFC	VITO	CNES, CESBIO and DLR
Algorithm version	2.5.5	1.4.1	1.0	1.0
Source Code	Yes*	Yes	No	No
Main language	Python	C and Fortran	Python	C++
Required dependencies	No (standalone)	Yes	No (standalone)	Yes
Output data format	SAFE / JPEG2000	GeoTiff / ESPA	GeoTIFF	GeoTIFF
Bit depth	Unsigned Integer 16	Integer 16	Float 32	Integer 16
Sensor supported	Sentinel-2	Landsat (Sentinel-2)*	Sentinel-2 Landsat-8	Sentinel-2 Landsat-8
Observed processing time	½ ~ 1 hour	< 5 minutes (on the tested scene)	½ ~ 1 hour	2 ~ 3 hours
Multithread	No	Yes*	No	Yes (configurable)
Projection	From original	From original	From original, but loosing CRS	From original
BRDF adjustment support	Yes	No	No	No

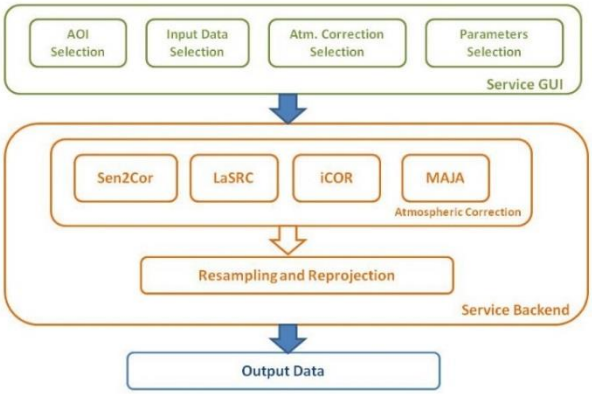


FIGURE 1. ARD ON-DEMAND DEMONSTRATOR BLOCK DIAGRAM.

Figure 1 depicts schematically the workflow of the ARD On-Demand demonstrator. The main processing block is divided in two steps: (1) processing of L1 product using one of the selected algorithms and (2) product resampling and re-projection according to the user's selection. Output data are available for download when the processing is finished.
During the feasibility study tests have been done for each of the selected algorithms to understand their readiness, performance and requirements in terms of hardware resources, auxiliary and metadata files needed, supported products, etc. This has allowed to assess and configure the service backend environment where the ARD On-Demand demonstrator runs: a dedicated cluster of four Virtual Machines with CentOS 7 operating system equipped with 32

GB of RAM and 8 CPU cores, has been deployed on a Cloud infrastructure where Sentinel-2 and Landsat input data are co-located. RSS Team has implemented the ARD On-Demand demonstrator integrating in one single processing service the four selected Atmospheric Correction Algorithms. The service demonstrator was ready at the end of July 2018 and it has been opened to a selected number of beta test users, with the purpose of testing the service and reporting to ESA-RSS their impressions and feedback.

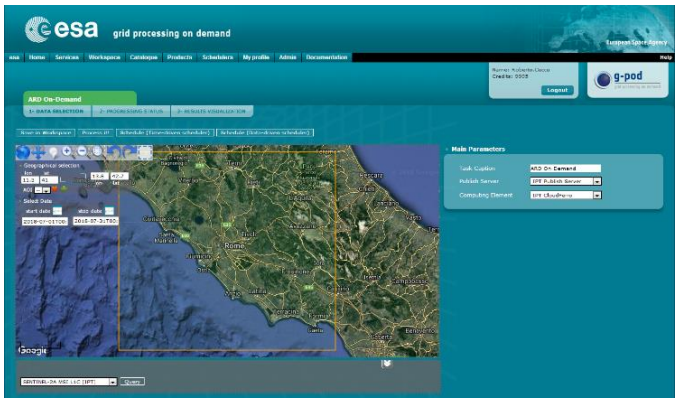


FIGURE 2. ARD ON-DEMAND DEMONSTRATOR SERVICE IMPLEMENTED IN RSS PROCESSING PLATFORM (G-POD).

Figure 2 shows the service GUI: Area of Interest, time-frame and the input dataset to be queried are all selectable from the GUI. For the purpose of the demonstrator, Landsat-8 and Sentinel-2 data products are provided for the last few months (May 2018 to July 2018).

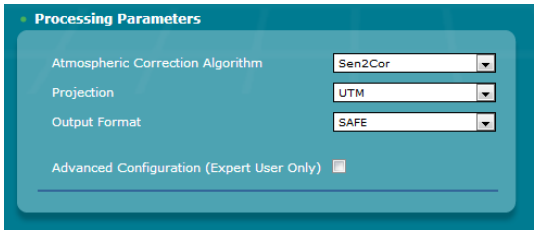


FIGURE 3. PROCESSING PARAMETERS SELECTION

From the same GUI, users can select the Atmospheric Correction Algorithm to use via a drop-down menu (Sen2Cor, LaSRC, iCOR, MAJA), the output projection (currently supported projections are UTM and Geographic Lat/Lon (WGS84)), and the output format (GeoTiff, HDF, NetCDF and SAFE - only for Sen2Cor -). The Resampling option is not yet available in the beta version of the ARD demonstrator. Advanced configurations for each Atmospheric Correction Algorithm are also possible for expert users: if a checkbox is ticked, a set of advanced parameters is displayed to the user (only Sen2Cor and iCOR advanced configuration is implemented in the current version).

The service has been tested by about 10 users who provided feedback and suggestions for future improvement of the demonstrator. Figure 4 illustrates the different atmospherically corrected Sentinel-2 products obtained for each of the integrated processors.

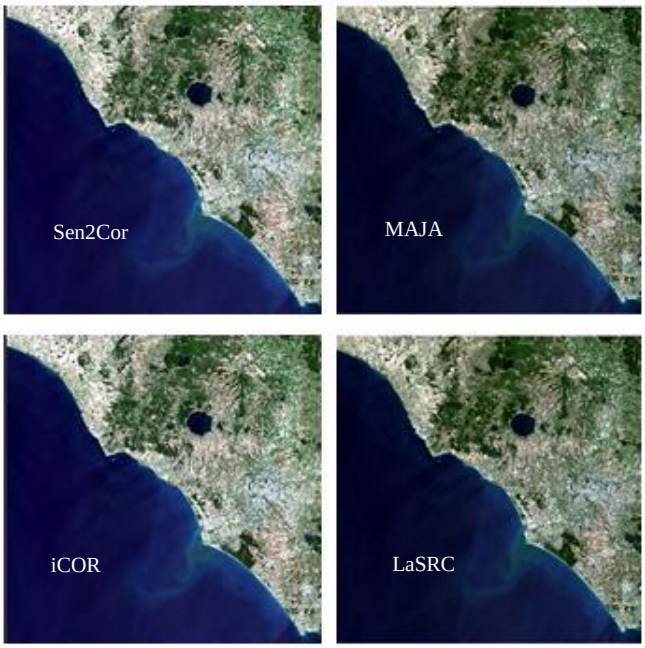
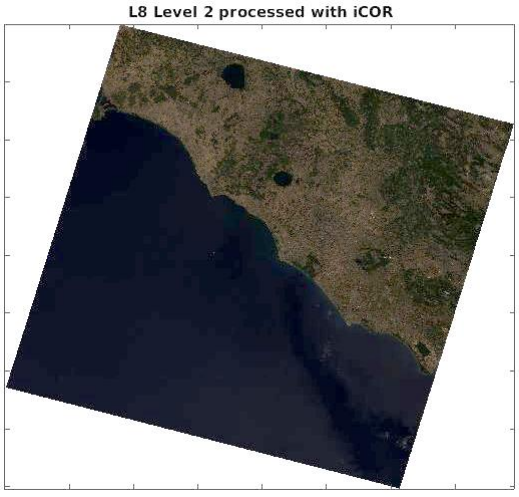


FIGURE 4. RGB TRUE COLOR COMPOSITE OF SENTINEL-2 OUTPUT PRODUCTS PROCESSED WITH THE ARD ON-DEMAND DEMONSTRATOR.

Figure 5 shows two images: the first one, acquired by Landsat 8 over the area of Rome - Italy, is the RGB composite of the atmospherically corrected product obtained with the ARD On-Demand demonstrator using iCOR. It is provided in UTM projection. The second image over the same area of interest is acquired by Sentinel-2 and is processed On-Demand using exactly the same options of the previous image (iCOR algorithm, UTM projection and the same file format). This is a basic example of analysis ready data starting from two different sensors.



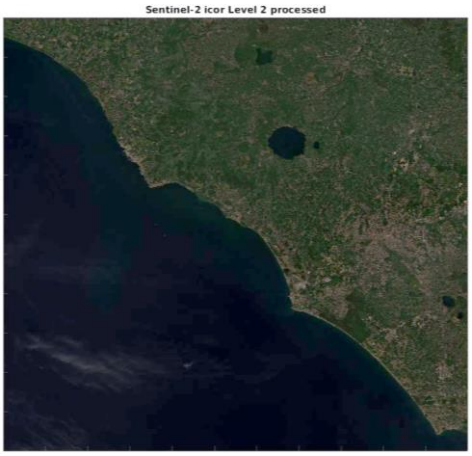


FIGURE 5. TOP: LANDASAT 8 L2 IMAGE PROCESSED WITH ICOR. BOTTOM: SENTINEL 2 L2 IMAGE PROCESSED WITH ICOR. BOTH IMAGES ARE IN UTM PROJECTION AND ARE PROVIDED IN THE SAME FILE FORMAT BY THE ARD ON-DEMAND DEMONSTRATOR.

3. LESSONS LEARNT AND RECOMMENDATION FOR FUTURE WORK

The demonstrator service has been opened to a selected number of beta-tester users (around 10 users at the moment of writing this abstract) with the objective to test the service and report internally feedback, suggestions and recommendations. Many comments arrived proposing often new features and desired functions that a future ARD On-Demand operational service should have. One of the most remarkable comments was the request to provide beside the results already provided by each Atmospheric Correction Algorithm (ACA), a quicklook of the results to easily visualize what was produced.

To respond to such request, RSS has implemented a web oriented visualization tool where a RGB composite of the results is produced and visualized together with basic information like: mission, product name, ACA processor used, projection and data format. This is one of the important improvements already implemented, but we are sure that more comments and recommendations for the future will come when the service will be opened to a wider community.

4. CONCLUSIONS

The need to have Analysis Ready Data for EO applications is clearly highlighted in this paper. Many initiatives are arising from EO providers to satisfy this requirement and other will come.

ESA-RSS has implemented and tested an ARD service demonstrator able to generate On-Demand Sentinel-2 or Landsat-8 atmospherically corrected products using different algorithms (Sen2Cor, LaSRC, iCOR, MAJA), projections (UTM and Geographic Lat/Lon) and data formats (GeoTiff, HDF, NetCDF). Improvements are still needed to satisfy new requirements and the service is being constantly updated taking into account recommendations and feedback provided by the beta tester user community.

5. REFERENCES

[1] http://ceos.org/ard/

[2] https://landsat.usgs.gov/ard

[3] Masek, J. G., Claverie, M., Ju, J., Vermote, E., & Justice, C. O. (2015, December). A Harmonized Landsat-Sentinel-2 Surface Reflectance product: a resource for Agricultural Monitoring. In AGU Fall Meeting Abstracts. https://agu.confex.com/agu/fm15/webprogram/Paper85646.html

[4] Marchetti P.G., Rivolta G., D'Elia S., Farres J., Mason G. and Gobron N., "*A Model for the Scientific Exploitation of Earth Observation Missions: The ESA Research and Service Support*." *IEEE Geoscience and Remote Sensing Society Newsletter* , Vol. 162, pp. 10-18, March 2012. http://www.grss-ieee.org/wp-content/uploads/2010/03/ngrs_NL_0312-Webv2.pdf

[5] Muller-Wilm, U., Louis, J., Richter, R., Gascon, F., & Niezette, M. (2013, September). *Sentinel-2 level 2A prototype processor: Architecture, algorithms and first results. In Proceedings of the 2013 ESA Living Planet Symposium*, Edinburgh, UK (pp. 9-13). http://seom.esa.int/lps13/abstracts/849980.html

[6] Vermote, E., Justice, C., Claverie, M., & Franch, B. (2016). *Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. Remote Sensing of Environment*. http://dx.doi.org/10.1016/j.rse.2016.04.008.

[7] De Keukelaere, L., Sterckx, S., Adriaensen, S., Knaeps, E., Reusen, I., Giardino, C., ... & Vaiciute, D. (2018). *Atmospheric correction of Landsat-8/OLI and Sentinel-2/MSI data using iCOR algorithm: validation for coastal and inland waters. European Journal of Remote Sensing*, 51(1), 525-542. https://doi.org/10.1080/22797254.2018.1457937

[8] Lonjou, V., Desjardins, C., Hagolle, O., Petrucci, B., Tremas, T., Dejus, M., ... & Auer, S. (2016, October). *Maccs-atcor joint algorithm (MAJA). In Remote Sensing of Clouds and the Atmosphere XXI (Vol. 10001, p. 1000107). International Society for Optics and Photonics*. https://elib.dlr.de/107293/

SELECTIVE DATA PROCESSING IN DIAS FOR LOCALIZED TIME SERIES ANALYSIS - A SPECIFIC USE CASE FOR A GENERIC DIAS PROCESSING SUITE

Nils Junike[1], Bernard Pruin[1], Alexander Strecker[1]
[1]Werum Software & Systems AG, Wulf-Werum-Str. 3, 21337 Lüneburg

ABSTRACT

We present a systematic production solution for DIAS with an application for local subsetting and data stacking for time series analysis. The scenario makes optimal use of Copernicus data vicinity within DIAS services to extract a relevant analysis ready dataset for local time series analysis and data mining. A pre-existing processing management solution is configured and instrumented to implement the workflow steps for the generation of a subsetted stack of co-registered high resolution Sentinel-1 GRD products. The ESA Snap toolbox batch application is exploited to provide the required algorithm implementations. An example scenario is presented for a representative administrative region that shows that, with a given spatial focus, time series analysis functionality can be made available with minimal resource and bandwidth requirements.

Index Terms— *Time Series, DIAS, Analysis ready data*

1. INTRODUCTION

Time series analysis without geographical constraints requires a large amount of additional memory to store data in an analysis ready format. For some missions and regions there are already significant analysis ready datasets available, e.g. for Landsat data over the USA [2]. For regions and missions, where data is not yet available in an analysis ready form or the form provided is not appropriate, e.g. due to specific needs on the the data tile structure and grid, alternative solutions are needed. As many applications for time series analysis have a focus on a specific region a lightweight local stacking approach is presented here based on a cloud-ready generic processing suite. Preparing only local data for time series analysis to create “data islands of interest” allows lowering the infrastructure requirements for time series generation.

An integrated software solution for time series data preparation is presented that addresses the various functionalities required for data preparation for time series analysis. The system is comprised of interface adapters to Copernicus data sources (DIAS), full term subsetted raw data archive to avoid lengthy data re-fetching exercises for potential re-gridding, integrated stack building software (SNAP Toolbox), stack product storage and dissemination functions.

The main building block of the overall stack building solution is a generic DIAS processing Suite (DIAS Suite) that comprises functionality for

- DIAS integration for data reception, storage and dissemination
- Systematic processing management
- Processing workflow management
- Processor modules integration
- Processing resource management
- System status monitoring

This generic processing suite can be instrumented to perform arbitrary production workflows. In the context of this paper, the configuration for a subsetting and stack building workflow is demonstrated on a typical region of interest. We have chosen the administrative district “Landkreis Harburg” for this exercise. The choice of the region is essentially arbitrary, but for demonstration purposes, it is useful that within the “Landkreis Harburg” significant construction has occurred since the launch of Sentinel-1 A. A large logistics facility with 64 000 m^2 area that has been constructed in 2016/2017 provides a good landmark for change observations.

2. DIAS INTEGRATION

The Copernicus Data and Information Access Service (DIAS) setups provide convenient and efficient access to Copernicus data. There are 5 DIAS systems in place, while the DIAS services provide the same data, the details of the data access interfaces are not harmonized and the access to computing and storage resources is specific to the individual DIAS instance. The solution thus follows a concept of interface adapters for the data discovery, data retrieval and data storage.

We have deployed the current demonstration solution in in CREODIAS (https://creodias.eu/). Deployment into other DIAS instances is also foreseen. It should be noted that the workflow itself is independent of the selected DIAS and remains unchanged, independent of the specifics of the DIAS instance chosen for integration.

3. SUBSETTED INPUT DATA ARCHIVE

The DIAS Suite provides systematic production capability that is triggered by emergence of new products

within a DIAS Service. The triggering mechanism allows filtering by product type as well as additional metadata constraints. For subsetting, of course, a matching of footprint versus region of interest is included.

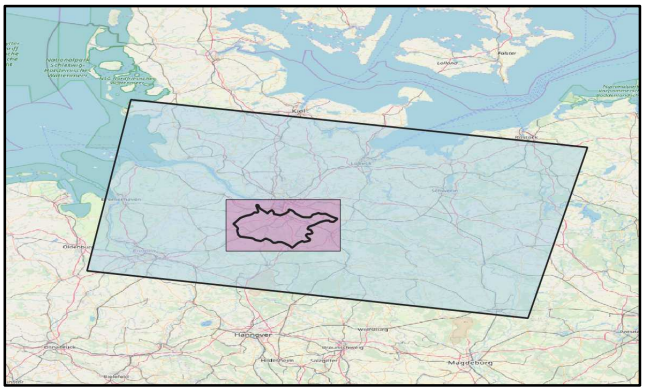


Figure 1 Area of Interest

Figure 1 shows the geospatial relations of the areas relevant in this exercise. A localized interest into a specific administrative region, here the district of Harburg is assumed. The district of Harburg, south of Hamburg, has an area of 1250 km^2. An average Sentinel 1 GRDH product (large quadrangle) has a coverage of 44 000 km^2. A reasonable bounding box covering the complete district has an area of 3 000 km^2. Subsetting the individual Sentinel-1 products that, to a large extend, only cover parts of the region of interest lead to an overall compressed products of 5 to 10% of the full product set size.

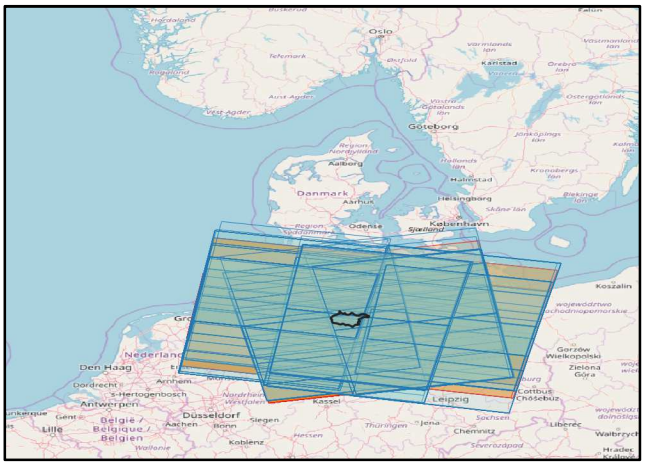


Figure 2 Covering Products

Since launch up to 22.10.2018, there have been 818 products from Sentinel-1 A and 339 from Sentinel-1 B intersecting the region of interest. Just these products requires nearly 1.2 TByte of storage space. A storage of the subsets is less than 100 GByte. Additionally, a proportional amount of storage is needed in order to store restructured data for optimized time series access. The above argument is true for the original band data (amplitudes). If other data bands are generated, e.g. as required to provide analysis ready data as per CEOS recommendations [7], again more storage space is needed. The current dataset that is generated as input into to 3D-pattern detection and deep learning algorithms, the initial bands however are sufficient.

Due to the complex interaction between the radar waves and individual surface objects, the representation of objects is highly dependent on the viewing direction [3]. The results of time series analysis may thus be more meaningful when performing analysis on stacked data separated individually by relative orbit. The S1 satellites acquire data over the region of interest on relative orbits 44, 66, 117, 139 and 168. This leads to 10 different stack products.

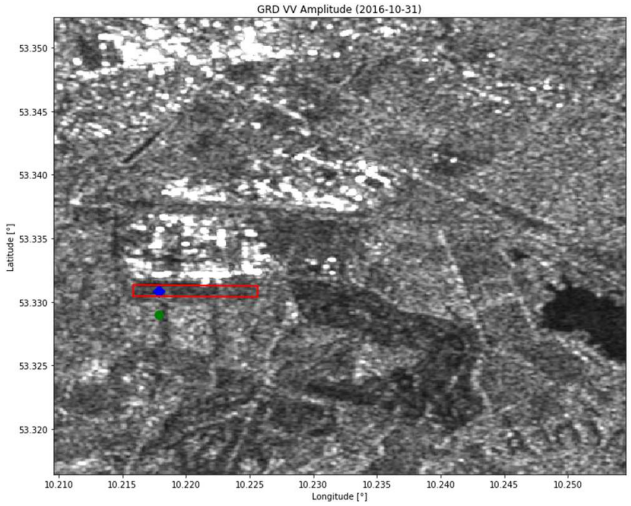


Figure 3 Amplitude from 2016-10-31 [4]

Figures 3 and 4 shows 2 VV polarized amplitude elements from a time series stack for orbit 117, zoomed in on probably the largest building of the district, a distribution center near Winsen (Luhe) covering approx. 64 000 m^2 (red square).

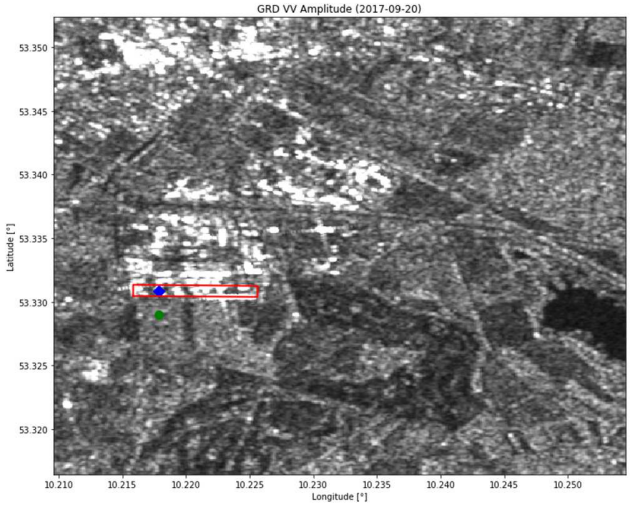


Figure 4 Amplitude from 2017-09-20 [4]

In the figures a 2 locations are highlighted, one within the building site and one over a nearby agricultural area. The time series as shown in Figure 5 reveals the clear signal from the new building.

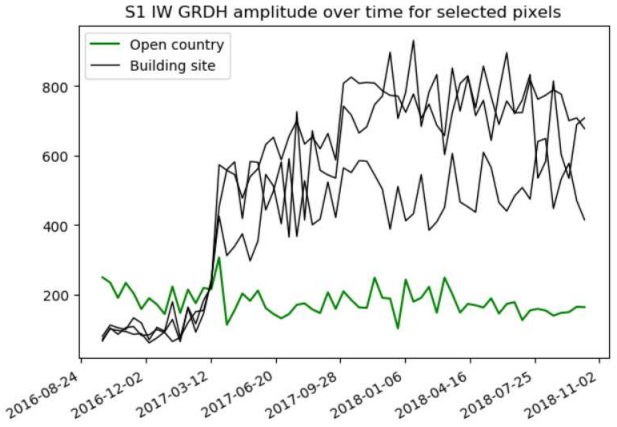


Figure 5 Timeline of amplitude at selected sites

To provide additional context information on the RADAR images of the figures 3 and 4, Figure 6 shows a natural color presentation of the same covered area as acquired by Sentinel-2 on 07.10.2018.

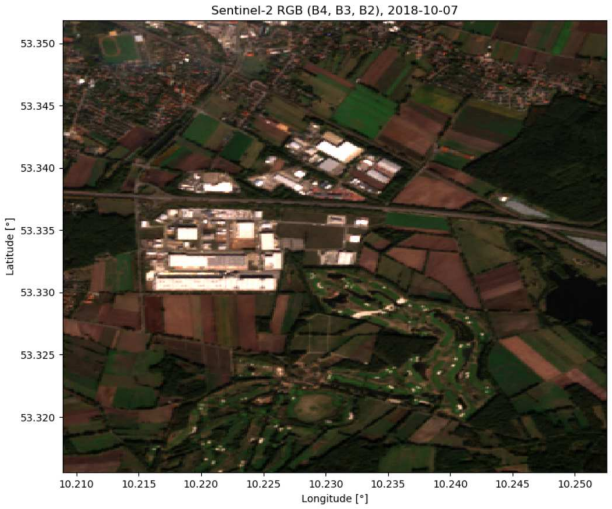


Figure 6 Natural Color Scene Sentinel-2 [4]

4. STACK GENERATION AND RESOURCE MANAGEMENT

The ESA SNAP toolbox [1] provides subsetting and stack building functionality that is employed for these functions for the supported product types. The SNAP toolbox also provides a batch mode that allows incorporating its workflows without recurrence to the graphical user interface. This batch interface is exploited to integrate the tool into the DIAS Suite. The suite supports a generic command line interface to incorporate third party processing tools and instrument processing facilities.

The SNAP batch interface is wrapped by a simple script to support the required interface and is called with differing parameters to call the required functions.

The main steps for stack generation are

- subsetting and terrain correction to build-up an input data archive (SUBTC)
- coregistration to create individually stacked images and combining to create an incremental stack product (STCK)

The executables are wrapped into a Docker image and are executed within the resource pool managed by the DIAS Suite. A Digital Elevation Model (DEM) is made available statically for use by all process instances, as per SNAP Toolbox requirements. Figure 7 provides a schematic overview of the steps that are performed and their respective inputs. The generated intermediate products are identified by scenario specific type names. The type names are defined in ad-hoc fashion and server to filter inputs for the different workflow steps.

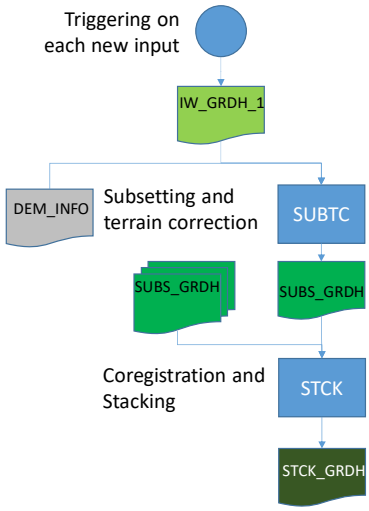


Figure 7 Stack Generation Workflow

The workflow depicted is encoded in a set of workflow rules and task table specifications that define a processor. The configuration language largely builds on a novel, domain specific workflow configuration language defined by ESA in the context of the EarthCARE mission [5].

Task Tables (Figure 8) are used to define the individual processing steps (e.g. SUBTC, STCK) and define their input output relation in machine-readable form for interpretation by a management layer.

Additional workflow specifications allow triggering of productions based on the emergence of a new product within the realm of the management layer. The generic rule solution also allows the scheduled execution of rules, but this feature is not needed for stack building as a stack only changes upon new product arrival.

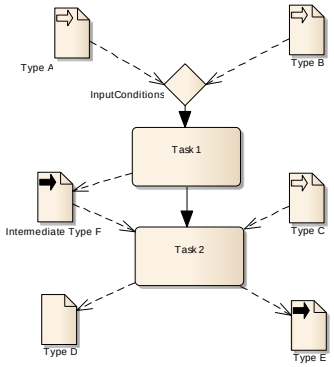


Figure 8 Task Table Elements (BPMN notation)

Intermediate and final products are stored in the managed archive of the suite. As the time series stack is incrementally increased, data is not purged automatically.

For the scenario of this paper, data from the past is used together with data that is not yet acquired. The overall stack results is created by manually initiating the production for already existing products with subsequent stacks being produced systematically and automatically whenever a new matching product emerges within the hosting DIAS instance.

5. STACK PROVISIONING

Each new stack product is provided to eligible users by providing it from the internal archive accessible via https. Alternatively, a "push" to any other infrastructure can be configured.

6. CONCLUSION AND OUTLOOK

We have presented a lightweight solution for the generation of localized analysis ready time series data. The solution is operational for demonstration purposes over an administrative region in northern Germany.

A number of improvements and extension of characteristics of the current stack output are considered:

- The current time series stacks simply use the geolocation grid of the first (oldest) product as coregistration reference. The use of a standard grid, e.g. the Equi7 grid [6] could be foreseen.
- No care is currently taken that the resulting stack is compatible with any standard. Compliance with the CEOS CARD4L product specification for normalized Radar backscatter [7]. should be considered. However, the current stack is based on original input IW_GRDH bands that contain backscatter amplitude values and are not yet normalized. To move towards CEOS ARD compliance, additional processing steps need to be included and the ESA Snap Toolbox functionality needs to be configured for compliance with the standard.
- Stacks are currently only generated for Sentinel-1 amplitude data starting from GRDH products. Starting from SLC products would allow the computation of additional quantities of interested for time series analysis depending on SNAP Toolbox capabilities and available funds for infrastructure resources. Coherence data for example may be useful for land cover classification [8]
- An extension to other missions, in particular Sentinel-2 is foreseen. For subsetting of the all bands however, a re-gridding of some bands need to be introduced due to limitations in the current SNAP Toolbox functionality set.
- The current workflow setup ends with the provision of the generated stack. As the DIAS Suite is generic, it can also be employed to add automated analysis functions to the workflow.

As an additional consideration, it should be mentioned that while the overall concept of the local stack is intended to provide a data set for flexible use in interactive processing outside the DIAS context, the data analysis can be performed within the cloud environment.

7. ACKNOWLEDGEMENTS

Background maps shown in this paper contain map data copyrighted OpenStreetMap contributors and is available from https://www.openstreetmap.org.

8. REFERENCES

[1] SNAP Toolbox Software. Version 6.0.0. (http://step.esa.int/main/toolboxes/snap/)

[2] U.S. Landsat Analysis Ready Data (ARD) (https://landsat.usgs.gov/ard).

[3] C.O. Dumitru, M. Datcu. Information Content of Very High Resolution SAR Images: Study of Feature Extraction and Imaging Parameters. IEEE TRANS. ON GEO SCIENCE AND REMOTE SENSING, VOL. 51, NO. 8, August 2013.

[4] Contains modified Copernicus Sentinel data 2016-2018.

[5] C. Caspar, N. Junike, B. Pruin, C. Stella, A. Strecker. EarthCARE processing facility and EarthCARE L2 testbed - A synergetic setup to support scientific algorithm development. IAC 2018. IAC-18,B1,IP,3,x43883.

[6] B. Bauer-Marschallinger, D. Sabel, W. Wagner, “Equi7 Grid. Optimisation of global grids for high-resolution remote sensing data,” Comput. Geosci., vol. 72, no. C, pp. 84–93, Nov. 2014.

[7] CEOS Analysis Ready Data Specification for Normalised Radar Backscatter. V3.2.1. (http://ceos.org/ard)

[8] F. Vicente-Guijalba et.al. Assessing hypertemporal Sentinel-1 Coherence maps for Land Cover monitoring. 2017 9th Int. Workshop on the Analysis of Multitemporal Remote Sensing Images. DOI: 10.1109/Multi-Temp.2017.8035240

PYROSAR: A FRAMEWORK FOR LARGE-SCALE SAR SATELLITE DATA PROCESSING

John Truckenbrodt[1], Felix Cremer[1], Ismail Baris[2], Jonas Eberle[1]

[1]Friedrich-Schiller-University, Institute of Geography, Department for Earth Observation, Jena, Germany
[2]German Aerospace Centre DLR, Microwaves and Radar Institute, Oberpfaffenhofen, Germany

ABSTRACT

With the increase in Synthetic Aperture Radar (SAR) data availability, the need for a management solution arises, which is capable of handling the various available data sources and formats, as well as interfacing with different processing software solutions. Optimally, a user is presented with a standardized output ready for analysis without further knowledge of SAR processing details.

pyroSAR addresses this need by providing a complete workflow from retrieving the raw data from its provider to filled data cubes ready for analysis. It keeps record of registered scenes and their metadata, enables scalable processing in different software via homogenized interfaces and tracks products in their source locations.

This way, pyroSAR can act as a central SAR data broker to keep track of what data is available in raw or processed format while reducing the need to learn different processing software solutions thus leaving more time for actual data analysis.

Index Terms— Processing, SAR, Data Cube, Data Management, SNAP, Gamma

1. INTRODUCTION

Since the launch of Sentinel-1A in 2014 an unprecedented amount of Synthetic Aperture Radar (SAR) data has become available. The Python package pyroSAR is being developed for easy handling of these large amounts of data by offering a complete data management and processing solution from the raw source product to analysis readiness in a data cube. Thus, pyroSAR is intended to reduce the time needed for handling data and software, which could otherwise be spent directly on the analysis of data and development of algorithms to derive information from it.

With pyroSAR, all available SAR-data can be downloaded, preprocessed to e.g. radiometrically terrain-corrected backscatter and exported to a selected format like GeoTiff by simply providing a test site geometry, the type of data required and some additional parameters, like spatial resolution and time frame. While similar workflows are possible in other SAR software, the learning curve might be quite steep and taking away valuable time from the analyst. pyroSAR gathers experiences made with different software and generalizes them to consistent workflows.

SAR data that has been downloaded in the past can be read by pyroSAR and its metadata be stored in a database so that all locally available data is accessible via a search catalogue. This way, a researcher can at any point get an overview of which data from current and past missions is available for a particular area of interest and prepare it for analysis.

To achieve this, pyroSAR has three central components. First, the identification of SAR scenes, extraction of their metadata and functionalities for format-specific file handling (Section 2). Several driver classes exist, which scan the provided scene archive, e.g. .SAFE for Sentinel-1, and read the scene metadata from relevant files. The SAR format drivers share a set of methods for general tasks and additionally offer methods specific to the respective format, e.g. downloading orbit state vector files.

Second, the scene metadata can then be registered in a SpatiaLite [1] database (Section 3). pyroSAR offers functionalities for maintaining this database and easily querying all scenes with a given set of metadata, which can then be passed to the processing framework (third), allowing for the flexible use of different SAR processing solutions (Section 4).

Currently, the ESA Sentinel Application Platform (SNAP) [2] and Gamma [3] are integrated. The numerous options of SAR processing and different approaches of processing software, e.g. SNAP XML workflows vs. Gamma shell commands, are broken down to simple homogenized Python functions, which select the necessary processing steps depending on the defined user requirements, e.g. study area, spatial target resolution and geocoding approach.

This way, pyroSAR goes beyond processing SAR scenes by providing a comprehensive toolchain for general spatial data handling connecting several open source software tools along the way. It creates a framework around existing SAR processing tools to ensure consistent output from each of them reducing their individual handling to a minimum.

2. METADATA HANDLING

The SAR image driver architecture is designed in a modular class inheritance scheme so that data from all sensors can he handled in the same way with mutual attributes and methods.

2.1. Metadata drivers

A fixed list of metadata attributes is demanded of each scene and the respective driver must support at least these attributes. For example, the attribute *start* needs to exist to describe the image acquisition start time. Hence, this information must be translated from original names by the driver, e.g. the ESA format metadata attribute *MPH_SENSING_START*. Several methods are available to homogenize the values of these fields to a common standard, for example conversion between different time stamp formats or reading corner coordinates into vector objects, which can then easily be intersected with an area of interest. In several cases pyroSAR internally uses the Geo Data Abstraction Library (GDAL) [4] to read metadata from SAR scenes. In cases where the returned information is not sufficient (e.g. CEOS format) or the GDAL driver was found to be too slow (e.g. Sentinel SAFE format), an own implementation is used. The list of standardized attributes and driver class serves the purpose of easily querying the scene database described in section 3, e.g. selecting all scenes acquired after a certain date via attribute *start*.

2.2. Naming scheme

A selection of these metadata attributes will be contained in the names of processed files, each being reserved a field of fixed length for its entry. This way it is ensured to easily keep track of processed files since they will all be named in a consistent way independent of raw data format, sensor and acquisition characteristics. This is shown in the following with curly brackets marking a metadata field and simple brackets defining the length of this field in the filename. All slots except processing steps have a fixed length, ensuring consistent naming and easier identification. In case the value of attribute *sensor* or *acquisition_mode* is shorter than this fixed length, the field is filled with underscores.

{sensor}(4)_{acquisition_mode}(4)_{orbit}(1)_{start}(15)_
{polarization}(2)_{processing steps}(*).tif

e.g.
S1A__IW___A_20141115T181801_VH_grd_mli_norm_geo.tif
ASAR_APP__D_20050123T092033_VV_pri_....tif
PSR1_PH___A_20100408T213246_HH_1.1_....tif

3. DATA ARCHIVE

pyroSAR works best by registering information about the scenes accessible to the processor in a data base. For ease of use without further knowledge about database software and in support of a lightweight portable solution, SpatiaLite was chosen. The pyroSAR database stores essential metadata of the scene, like sensor, acquisition mode and time. By using a Python function, the registered scenes can easily be queried and subsequently be passed to the processor. A processing directory can also directly be passed to the query function to filter out scenes, which had already been processed to this directory before.

4. PROCESSING FRAMEWORK

The Framework of pyroSAR integrates available Processing Software to enable easier and comparable usage. It will, for example, insert an extra processing step or a series of additional processing commands in case a vector geometry (e.g. shapefile) representing a test site is passed to the function in Python, including in-memory reprojection if necessary. It also handles the required ancillary data, like DEM files or orbit state vector information.

4.1. General Features

The following describes general concepts of the processing framework, which are shared among the individual processing software APIs.

4.1.1. Border Noise Removal

This step describes the masking of an image artifact specific to Sentinel-1, which is not yet sufficiently achieved by neither SNAP nor Gamma. By following the guidelines published by ESA [5], it was found to still have artifacts remaining in the images, which would have prevented further automated time-series analysis and computation of multi-temporal statistics. While this has become superfluous for images processed by the ESA Instrument Processing Facility (IPF) version 2.9 published in January 2018 or later [5], additional steps were found necessary to achieve a final data quality suited for time series analysis. A custom approach was implemented based on the Visvalingam-Whyatt method of poly-line vertex reduction [6]. First the border noise is reduced as recommended by [5] after which the remaining noise patterns are removed by cutting off irregular patterns at the image borders comparable to grinding off rough edges.

4.1.2. Ancillary Data

Another key requirement was the efficient use of ancillary data needed for processing, DEMs and orbit state vector files (OSV) in particular. SNAP automatically downloads the needed files into a defined directory structure during processing. While this stringent directory naming scheme is seen as an advantage, the automatic downloading during processing can be of disadvantage in a multi-node server scheme, where not each node might have internet access. Hence, pyroSAR offers functionalities to download and manage necessary ancillary data into this directory structure such that subsequent processing is possible without internet access. Furthermore, it is planned to apply this directory structure to the Gamma processing API for mutual use so that file storage can be reduced.

4.1.3. Error Handling

Any error message that originated from the processing software should be handled by pyroSAR internally and be passed to the user only if necessary. For example, an automated processing job of several hundred scenes should not be aborted if only one scene is corrupt. This error should instead be protocolled in log files for later error tracing. Being able to differentiate between different error messages and translating them to Python error types is seen as necessary to achieve this since error messages might be difficult to interpret by a user and differentiation of error severity is not possible by executing piped processes.

4.1.4. Parallelization

pyroSAR uses the Python package pathos [7, 8] for parallelization of several geo-processing tasks across CPUs. This package was found to be superior to native Python solutions due to the easy passing of objects across parallel processes. In a multi-node setup, the scoop package [9] is used for process coordination via SSH. pyroSAR offers API functionality with the aim to make it as easy as possible to work with both solutions. Both SNAP and Gamma can themselves be parallelized across multiple CPUs so that own approaches are only necessary for multi-node orchestration.

4.1.5. Spatial Data Handling

Initially, several classes and functions for handling spatial data using GDAL were directly available in pyroSAR. It was recently decided to outsource this functionality to a separate package spatialist [10], which might be used outside of the scope of pyroSAR. This package is also available via GitHub and PyPI and is currently developed parallel to pyroSAR with new requirements defined by developments in the latter. Like pyroSAR, spatialist works both on Windows and Linux, which is ensured via continuous integration. It is intended as a more user-friendly API to GDAL and further offers several convenience functionalities for easy spatial data handling.

4.1.6. Product Export

Currently, pyroSAR supports two schemes for handling data after processing. Initial development focused on mosaicking and resampling the individual GeoTiffs to specific test site boundaries so that images from different scenes acquired on the same satellite overpass are combined into one file and the individual tiles all share the same size. This way they can be treated as a single multi-layer file or 3D memory array.

Recent work focused on the exploitation of the Open Data Cube software [11]. Export of scenes processed by pyroSAR directly into such an Open Data Cube is now easily possible. pyroSAR currently offers a set of classes and functions to read relevant metadata from a collection of raster files and generalize this metadata into an Open Data Cube product definition YAML file as well as indexing YAML files for each GeoTiff exported by pyroSAR. These exported files can then be passed to the command line functionalities of the Open Data Cube software for creating new products and adding the selected SAR scenes to them. Internally, several checks are performed by pyroSAR to ensure homogeneity among scenes and compatibility of them with already existing products.

4.2. SNAP API

Processing with the Sentinels Application Platform is realized by internally parsing XML workflows depending on the user input to pyroSAR functions and subsequently executing these workflows with the SNAP Graph Processing Tool (GPT) via Python's subprocess module. This way, it is not necessary to start the graphical user interface at any time. The processing workflows created by pyroSAR are regularly checked for applicability in updated versions of SNAP. Additional configuration of GPT ensures a consistent output regardless of user input and SNAP default configuration. This way, the processing output is always one or several single-layer cloud-optimized GeoTiff files, each containing one polarization. The workflow is stored together with the image files sharing the same naming scheme so that at any point the processing steps of an individual dataset can be retraced and the scene database query results be filtered to scenes which have not yet been processed to the target directory before.

4.3. Gamma API

The execution of Gamma command line tools is realized in Python via using the subprocess module. Logfiles are automatically written to a user-defined directory. Several functions and classes are available to simplify parametrizing and connecting the numerous command line tools needed for processing the scenes. For example, parsing attributes of a Gamma parameter file into an object and directly exporting selected attributes to an ENVI HDR file for display in other software. Or, creating a DEM file in Gamma format from SRTM data including gap filling, geoid correction and reprojection by just defining a vector geometry and a directory containing the SRTM tiles. The most recent addition is an automatic parser for converting the documentation of the individual Gamma commands to Python functions. This gives the user the opportunity to easily wrap the processing steps into a clearly arranged Python script with named parameters without having to worry about keeping the right order of the positional parameter arguments passed to the Gamma commands. A custom wrapper for the subprocess module then executes the commands, writes log files and interprets error messages raised by Gamma into Python error types so that the error

can be reacted to appropriately. For example, an error indicating insufficient cross-correlation matches might just require another iteration with different parametrization instead of aborting the whole process.

5. SOFTWARE MAINTENANCE

pyroSAR is hosted on GitHub [12] and PyPI. A testing framework was established using pytest, which currently tests basic general functionalities and contains an initial collection of test data sets. Continuous integration into both Linux and Windows is achieved through Travis CI and AppVeyor respectively. Documentation is provided through sphinx and readthedocs.

6. DISCUSSION AND FUTURE WORK

In this work we described the Python library pyroSAR which allows for the seamless processing of large amounts of SAR data. Currently the developed functions, both for SNAP and Gamma, are optimized for processing Sentinel-1 IW GRD products to geocoded Gamma0. This includes subsetting in case a test site geometry is defined, removal of border noise, updating of orbit state vector information from external sources, Range-Doppler geocoding, topographic normalization and optional scaling to dB. Having developed a framework, which has proven reliable for processing several thousands of SAR scenes in a multi-node server environment, current focus lies in making the software more user-friendly and transparent. While frameworks for CI-testing and automated documentation have been implemented, further work is necessary for making both tests and documentation complete and improving the usability by providing specific use-cases and workflows in the documentation for demonstrating how the numerous functionalities within pyroSAR work together. Furthermore, testing does currently not include processing of SAR scenes, which is still done manually. It is seen inevitable to automate the testing of pyroSAR's processing capabilities to ensure reliability for different SAR sensors and products and immediately detect changes in the processing software solutions, which pyroSAR offers APIs for.

Furthermore, great potential is seen in further homogenizing the processing APIs in order to consistently map the workflows in a mutual format, e.g. XML, which enables tracing of data from raw download to availability in a data cube.

In a general perspective of keeping pyroSAR up to the rapidly evolving field of big earth data, focus will be placed on integration into platform environments and further adaptation of big data software solutions. Utilization of the Data and Information Access Services (DIAS) or Amazon Web Services (AWS) and integration of big data software such as e.g. Docker and Kubernetes for process orchestration as well as the Open Data Cube and rasdaman for data management are of particular interest.

7. ACKNOWLEDGMENTS

This work was funded primarily by the European Commission via EU-H2020 project Satellite-Based Wetland Observation Service (SWOS) (Grant No 642088) as well as the German Aerospace Centre (DLR) in the Sentinel4REDD project (FKZ:50EE1540).

8. REFERENCES

[1] A. Furieri. (2017). *SpatiaLite*. https://www.gaia-gis.it/fossil/libspatialite/wiki?name=SpatiaLite

[2] ESA. (2018). *SNAP - ESA Sentinel Application Platform*. https://step.esa.int/

[3] Gamma Remote Sensing. (2018). *Gamma Software*. https://www.gamma-rs.ch/

[4] GDAL/OGR contributors. (2018). *GDAL/OGR Geospatial Data Abstraction software Library*. https://gdal.org

[5] N. Miranda and G. Hajduch, "Masking "No-value" Pixels on GRD Products generated by the Sentinel-1 ESA IPF," CLS2018-01-29 2018, issue 2.1. https://sentinel.esa.int/documents/247904/2142675/Sentinel-1-masking-no-value-pixels-grd-products-note.

[6] M. Visvalingam and J. D. Whyatt, "Line generalisation by repeated elimination of points," *The Cartographic Journal*, vol. 30, no. 1, pp. 46-51, 1993/06/01 1993.

[7] M. M. McKerns, L. Strand, T. Sullivan, A. Fang, and M. A. G. Aivazis, "Building a Framework for Predictive Science," in *Python in Science Conference*, 2011.

[8] M. M. McKerns and M. A. G. Aivazis. (2018). *pathos: a framework for parallel graph management and execution in heterogeneous computing*. http://trac.mystic.cacr.caltech.edu/project/pathos/wiki.html

[9] Y. Hold-Geoffroy, O. Gagnon, and M. Parizeau, "Once you SCOOP, no need to fork," presented at the Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment - XSEDE '14, 2014.

[10] J. Truckenbrodt, I. Baris, and F. Cremer. (2018). *spatialist: A Python Module for spatial data handling*. https://github.com/johntruckenbrodt/spatialist

[11] ODC initiative. (2018). *Open Data Cube*. https://www.opendatacube.org/

[12] J. Truckenbrodt, F. Cremer, and I. Baris. (2018). *pyroSAR online repository*. https://github.com/johntruckenbrodt/pyroSAR

SAR ALTIMETRY PROCESSING ON DEMAND FOR CRYOSAT-2 AND SENTINEL-3 USING THE ESA RESEARCH AND SERVICE SUPPORT

Jérôme Benveniste [(1)], Salvatore Dinardo[(2)], Giovanni Sabatino[(3)], Marco Restano[(4)], Américo Ambrózio[(5)],

[(1)] ESA-ESRIN, Via Galileo Galilei, Frascati, Italy, Email: jerome.benveniste@esa.int
[(2)]He Space/EUMETSAT, [(3)]Progressive Systems/ESRIN-RSS, [(4)] SERCO/ESRIN,
[(5)] DEIMOS/ESRIN

ABSTRACT

The scope of this paper is to feature the G-POD (Grid Processing On Demand) SARvatore service to users for the exploitation of CryoSat-2 and Sentinel-3 data, which was designed and developed by the Altimetry Team at ESA-ESRIN EOP-SD. The G-POD service coined SARvatore (SAR Versatile Altimetric Toolkit for Ocean Research & Exploitation) is a web platform that allows any scientist to process on-line, on-demand and with user-selectable configuration CryoSat-2 SAR/SARIN and Sentinel-3 SAR data, from L1a (FBR) data products up to SAR/SARin Level-2 geophysical data products. Several years of CryoSat-2 FBR data are at the disposal of the user, plus the full power of the G-POD's cluster: 600 CPUs and over 500 TB of storage.

Index Terms - SAR ALTIMETRY, CRYOSAT, SENTINEL-3, GPOD, SAMOSA, SENTINEL-3 STM

1. INTRODUCTION

The SAR Versatile Altimetric Toolkit for Ocean Research & Exploitation (SARvatore) takes advantage of the G-POD (Grid Processing On Demand) distributed computing platform (600 CPUs in ~90 Working Nodes) to timely deliver output data products and to interface with ESA-ESRIN FBR data archive (439,184 SAR passes and 367,592 SARin passes for Cryosat-2 and 39'000 SAR passes for Sentinel-3A). The output data products are generated in standard NetCDF format (using CF Convention), therefore being compatible with the Multi-Mission Radar Altimetry Toolbox (BRAT) and other NetCDF tools. By using the G-POD graphical interface, it is straightforward to select a geographical area of interest within the time-frame related to the Cryosat-2 SAR/SARin FBR and Sentinel-3 L1A data products availability in the service catalogue. The processor prototype is versatile, allowing users to customize and to adapt the processing according to their specific requirements by setting a list of configurable options. Pre-defined processing configurations (Official CryoSat-2, Official Sentinel-3, Open Ocean, Coastal Zone, Inland Water (20Hz & 80Hz), Ice and Sea-Ice) are available. After the task submission, users can follow, in real time, the status of the processing, which can be lengthy due to the required intense number-crunching inherent to SAR processing. From the web interface, users can choose to generate experimental SAR data products as stack data and RIP (Range Integrated Power) waveforms. The processing service, initially developed to support the awarded development contracts by confronting the deliverables to ESA's prototype, is now made available to the worldwide SAR Altimetry Community for research & development experiments, for on-site demonstrations/training in training courses and workshops, for cross-comparison to third party products (e.g. CLS/CNES CPP or ESA SAR COP data products), for producing data and graphics for publications, etc. Initially, the processing was designed and uniquely optimized for open ocean studies. It was based on the SAMOSA model developed for the Sentinel-3 Ground Segment using CryoSat data (Cotton et al., 2008; Ray et al., 2014). However, since June 2015, the SAMOSA+ retracker is available as a dedicated retracker for coastal zone, inland water and sea-ice/ice-sheet. A new retracker (SAMOSA++) has been recently developed and will be made available in the future. It will make usage of the RIP providing in output mean square slope. Following the launch of Sentinel-3, a new flavor of the service has been initiated, exclusively dedicated to the processing of Sentinel-3 mission data products. The scope of this new service is to maximize the exploitation of the Sentinel-3 Surface Topography Mission's data over all surfaces providing user with specific processing options not available in the default processing chain. The services are open, free of charge (supported by the ESA SEOM Programme Element) for worldwide scientific applications and available at URL reported in section 7. In this paper, we present first the ESA G-POD framework and system. Then we describe in detail the CryoSat-2/Sentinel-3 SAR Processing service integrated in G-POD and we conclude with the output package description and information on the contacts and references.

2. G-POD SYSTEM

The ESA Grid Processing on Demand (G-POD) system is a generic GRID-based operational computing environment where specific data-handling Earth-Observation services can be seamlessly plugged into system. One of the goals of G-POD is to provide users with a fast computational facility without the need to handle bulky data.

The G-POD system hosts high-speed connectivity, distributed processing resources and large volumes of data to provide scientific and industrial partners with a shared data processing platform fostering the development, validation and operations of new Earth Observation applications. In particular, the G-POD environment consists of:

- Over 600 CPUs in about 90 Working Nodes (50 reserved for SARvatore with 8 CPUs & 32GB RAM each. 5 tasks are typically parallelized in each node).
- Over 400 TB of local on-line Storage plus 180 TB of EO data accessed directly from the PACs.
- Access to Cloud processing and data resources on demand (from Interoute and other providers).

- Online software resources: IDL, MATLAB, BEAT, BEAM, BRAT.

Considering the specs of each working node, SARvatore is averagely 50 times faster than a PC with similar specs using no parallelization (250 times faster, if all available resources would be allocated to a single user). As an example, GPOD can process 7 years of CryoSat-2 SAR collected over the Arctic during the month of September (approx. 7'500 passes) in 15 days (83.3 passes/day). An user would require 750 days to process the same amount of data.

Actually, G-POD has more than 400TB of EO data locally stored. EO Data available to G-POD services come either from ESA and non-ESA missions. The G-POD web portal (http://gpod.eo.esa.int/) is a flexible, secure, generic and distributed web platform where the user can easily manage all own tasks. From the creation of a new task to the output publication, including data selection and job monitoring, the user goes through a friendly and intuitive user interface accessible from everywhere. More detailed information on the G-POD Web Portal and System are available here: http://wiki.services.eoportal.org/tiki-index.php?page=GPOD+User+Manual#Annex

3. CRYOSAT-2/ SENTINEL-3 SAR PROCESSING ON DEMAND SERVICE

The ESA G-POD Earth-Observation Service, SARvatore (SAR Versatile Altimetric Toolkit for Ocean Research & Exploitation) for CryoSat-2 and Sentinel-3 is an Earth-Observation application that provides the capability to process remotely and on demand CryoSat-2 SAR and Sentinel-3 data, from L1a (FBR, Full Bit Rate) data products until SAR Level-2 geophysical data products (Jensen and Raney, 1998; Wingham et al., 2006; Martin-Puig et al., 2008; Raney, 2008; Raney, 2012; Raney 2013).

The service works over any kind of surfaces and has been enhanced for inland water, land, sea-ice and ice sheets, implementing the SAMOSA+ model. The service is based on the SAR Processor Prototype that has been developed entirely by the ESA-ESRIN EOP-SD Altimetry Team (the authors) for CryoSat-2 & Sentinel-3 validation purposes, with the following system features:

- SAR/SARin FBR(L1a)/L1b DATA Archiving and Cataloguing
- SAR/SARin L1b Processor Prototype (Standard Delay-Doppler Processing)
- SAR/SARin L2 Retracker Prototype with SAMOSA Analytical Model and LEVMAR Least Square Estimator (Cotton et al., 2008; Ray et al., 2014)
- Input: CRYOSAT SAR/SARIN FBR DATA; Sentinel-3 SAR L1a Data
- Output L1b → Radar Echogram
- Output L2 → SSH, SLA (w/o SSB), SWH, sigma0, wind speed

The ESRIN EOP-SD ALT Team succeeded to compile the processor for a 64-bit Linux platform and delivered to the ESA G-POD team the executable codes, the input archive (CryoSat SAR FBR) and satellite footprints (ASCII tracks).

Now, the toolkit has been fully integrated in the GPOD System for gridded and on-demand computation.

The objectives of the service integration in GPOD are:

- to experiment in-house research themes that will be further matured in the ESA-funded R&D projects;
- to provide expert users with consolidated SAR geo-products to get acquainted with the novelties and specifities of SAR Altimetry;
- to validate CryoSat-2 & Sentinel-3 ocean products.

The service is open, free of charge and accessible online from everywhere. In order to be granted the access to the service, you need an EO-SSO (Earth Observation Single Sign-On) credentials (for EO-SSO registration, go to https://earth.esa.int/web/guest/general-registration) and afterwards, you need to submit an e-mail to G-POD team (write to eo-gpod@esa.int), requesting the activation of the service for your EO-SSO user account.
After the registration to EO-SSO, users can freely access the online service at: https://gpod.eo.esa.int/services/CRYOSAT_SAR/, https://gpod.eo.esa.int/services/SENTINEL3_SAR/. The services are listed under the Marine Theme. You can find them using the search bar as well.

4. WEB USER INTERFACE

Once you get to the service page (Fig. 1), the first action is to select the zone of interest and the time of interest for the required run. Regarding the selection of the area of interest, the user can simply draw a rectangle on the world map, after clicking on the rectangle icon on the tool bar. Instead, for more precise geo-selection, the user can either type directly the geo-coordinates of the area of interest using the geographical bar or switch to the newly available Google Earth map which ease the selection of inland water bodies.

Regarding the time of interest, the user may set the start and stop dates in the calendar bar. By default, the start date is the time of CryoSat-2/Sentinel-3 launch and the stop date is set at 2 months prior to the current date. The GUI embeds all the standard buttons for image browsing as panning, zoom-in zoom-out, centering, undo, redo, reset, etc. Once the time and geo selection is done, clicking on the "QUERY" button, the service lists all the CryoSat-2/Sentinel-3 passes matching the time and space requirements. The CryoSat-2/Sentinel-3 SAR tracks, crossing the area of interest, are then shown on the world map in overlay. The graphical interface lists a maximum of 250 passes per page and informs users of the total number of found passes. The user can decide which passes to select by clicking on the passes, select all, or delete some specific passes from the list.

On the top right, user finds a preference panel wherein user can set:

- Name of the current task
- Ftp Server where to publish the results (portal or personal)
- Data compression (tgz, none, single file)
- Grid Computing Resources
- Task Priority

doi:10.2760/848593

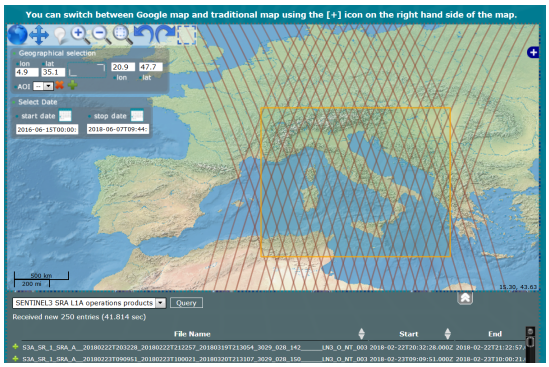


Figure 1: G-POD Sentinel-3 Service Main Interface.

The last step, before submitting the task, is to set the list of processing options. Indeed, the processor prototype is versatile in the sense that the users can customize and adapt the processing algorithms with flags and parameters, according their specific requirements, acting upon a list of configurable options. In the G-POD interface, users can easily enter this list of processing options via a series of drop-down menus. The configurable options are divided according to the processing level they refer to (L1b and L2). Starting from the first SARvatore release in 2014, the following upgrades have been introduced:

- Support for CryoSat-2 SARin Data.
- Advanced SAMOSA algorithm (SAMOSA+) for coastal & inland water domains.
- Added support for high posting rate (HPR) at 80 Hz in delivering the output geophysical parameters.
- Tide Model (TPX08),Geoid (EGM2008, EIGEN-4C6), Mean Sea Surface (CNES-CLS MSS2015) and Sea State Bias Solution (CLS Jason-2).
- Support for Sentinel-3A SAR data (NTC and STC data).
- ECMWF SWH and Wind Speed.
- NSIDC (sea ice concentration & age) and NCEP (Sea surface temperature and precipitable water) data.
- Joint & Share Forum (a meeting place to post questions and report issues).
- Data Repository (datasets processed for the users are available to the Altimetry Community).

Moreover, by selecting the processing options properly, users can mimic the CryoSat-2 or the Sentinel-3 processing baseline for an easy cross-comparison between missions. Pre-defined processing configurations (Official CryoSat-2, Official Sentinel-3, Open Ocean, Coastal Zone, Inland Water (20Hz & 80Hz), Ice and Sea-Ice) are available. Once the user has selected his processing options, in order to submit the task to G-POD Computing Elements, remains to click on the "PROCESS IT" button. After submission of a job, users will be directed to the workspace page where they can monitor in real time the status of the run and can be notified on the run status. The color code is:

- **Orange** → run under processing
- **Green** → run completed
- **Red** → run failed

Furthermore, by clicking on the task, the user can have more information, such as: Task Id, Task Creation Time, Processing Id, Grid Working Node Id, Task Progress (retrieving, processing, and publishing).After run completion, by clicking on the button "Jobs Information", the user can inspect:

- the GPOD log file (.stdout or .stderr) where eventual errors on data retrieving or data storing are reported;
- the prototype configuration file (L1b_CONFIG_FILE.log and L2_CONFIG_FILE.log) where are reported all the processing options;
- the prototype log files (L1b_start.log and L2_start.log) where are reported eventual prototype processing errors.

Users can also decide to change one or more processing options and then re-submit the task. In case of successful run completion (green status), the portal will provide an http link from where to download the output package on the user's own local drive. The users can order to post the package directly on a personal ftp server after having communicated to the web platform the ftp server credentials (through the "publish servers" sub-menu). This is the recommended option in case of processing of large amount of data.

Future releases will:

- Support the UPorto GPD wet correction.
- Support new Tide Models (FES 2014b & TPX09).
- Enhance retracking capabilities with the SAMOSA++ retracker.
- Support for Sentinel-3B SAR data.

5. OUTPUT PACKAGE & BRAT TOOLBOX COMPATIBILITY

The output package consists of:

- Satellite Pass Ground-Track in KML format
- Radar Echogram Picture in PNG format
- L2 Data Product in NetCDF format containing all the scientific results

The NetCDF format is self-explanatory with all the data field significance described in the attributes. The NetCDF Data Product follows the CF (Climate&Forecast) 1.6 Convention and can be opened with any standard NetCDF tools (ncdump, HDFview, etc.).

The following upgrades have been introduced for NetCDF Data Products:

- Inclusion of SAR echo and SAR RIP (Range Integrated Power) waveforms in the NetCDF files.
- Inclusion of STACK Data in the NetCDF files.

The recommended option is to ingest the NetCDF Data Products in BRAT Toolbox in order to exploit all the BRAT functionalities to browse and visualize the output content (Fig. 2). The Broadview Radar Altimetry Toolbox (BRAT) is a software suite designed to facilitate the use of radar altimetry data. It is able to read most

doi:10.2760/848593

distributed radar altimetry data, from ERS-1, ERS-2, TOPEX/Poseidon, Geosat Follow-On, Jason-1, Jason-2, Envisat, CryoSat-2, Jason-3 and Sentinel-3, to perform some processing, data editing and statistics, and to visualize the results. As part of the Toolbox, a Radar Altimetry Tutorial provides information about radar altimetry, the technique involved and its applications, as well as an overview of past, present and future missions, including information on how to access data and additional software and documentation. It also presents a series of data use cases, covering all uses of altimetry over ocean, cryosphere, inland water and land, showing the basic methods for some of the most frequent manners of using altimetry data. BRAT has been developed under contract with ESA and CNES (http://www.altimetry.info and http://earth.esa.int/brat/).

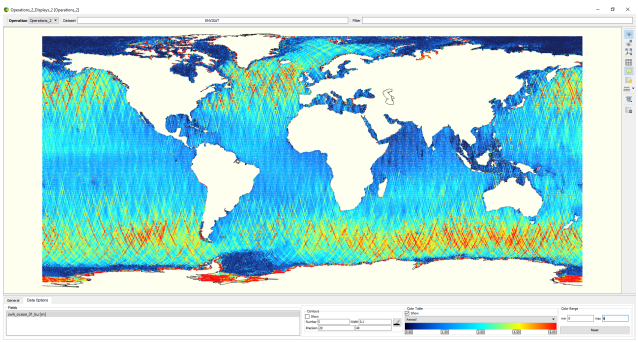


Figure 2: A cycle of Envisat data (Significant Wave Height) opened in BRAT.

6. CONCLUSIONS

To foster a new generation of SAR altimeter specialists and to get prepared for the Scientific Exploitation of Operational Missions (SEOM), a configurable versatile SAR processor has been developed and hosted in the ESA G-POD infrastructure. The G-POD Service coined SARvatore (SAR Versatile Altimetric Toolkit for Ocean Research & Exploitation) is a web platform that provides the capability to process on-line and on-demand CryoSat-2 and Sentinel-3 SAR data, from L1a (FBR) data products until SAR Level-2 geophysical data products, with a suite of selectable configuration parameters. The processing algorithms are the ones used in the Sentinel-3 Ground Segment, which mathematical model, SAMOSA, is described in Ray et al. (2014). By selecting the processing options properly, users can mimic the CryoSat-2 or the Sentinel-3 processing baseline for an easy cross-comparison between missions. Moreover, specific processing options not available in CryoSat-2 and the Sentinel-3 processing baselines have been made available to users along with pre-defined processing configurations. The Broadview Radar Altimeter Toolbox can display the output of SARvatore. The service is open, free of charge and accessible online from everywhere.

7. FURTHER INFORMATION

For any question, bug report and support, please contact us at: altimetry.info@esa.int and eo-gpod@esa.int

SARvatore is available at:
https://gpod.eo.esa.int/services/CRYOSAT_SAR/
https://gpod.eo.esa.int/services/CRYOSAT_SARIN/
https://gpod.eo.esa.int/services/SENTINEL3_SAR/

SARvatore Data Repository is available at:
https://wiki.services.eoportal.org/tiki-index.php?page=SARvatore+Data+Repository&highlight=repository

SARvatore “Join & Share” Forum is available at (GPOD section):
https://wiki.services.eoportal.org/tiki-custom_home.php

BRAT is available at: http://earth.esa.int/brat

8. REFERENCES

Cotton, D. et al., 2008, Development of SAR Altimetry Mode Studies over Ocean, Coastal Zones and Inland Water - State of the Art Assessment, http://www.satoc.eu/projects/samosa/docs/SAMOSA-TN01-V1.0full.pdf/

CryoSat User Workshop proceedings, SP-717, 2014, http://www.spacebooks-online.com/product_info.php?products_id=17581&osCsid=sscldrhw/

Dinardo, S. and J. Benveniste, Guidelines for the SAR (Delay-Doppler) L1b Processing, ESA XCRY-GSEG-EOPS-TN-14-0042, Is. 2.3, 29/05/2013.

Jensen, J. R., and R. K. Raney, "Delay Doppler radar altimeter: Better measurement precision," in Proceedings IEEE Geoscience and Remote Sensing Symposium IGARSS'98. Seattle, WA: IEEE, 1998, pp. 2011-2013.

Martin-Puig, C. et al., SAR Altimetry Applications over Water, ESA SeaSAR Workshop, 21-25 January, SP-656, 2008.

Raney, R. K., "CryoSat SAR-Mode Looks Revisited," Proceedings, ESA Living Planet Symposium, Bergen, Norway, 2010.

Raney, R. K., "CryoSat SAR-Mode Looks Revisited," IEEE Geoscience and Remote Sensing Letters, vol. 9, pp. 393-397, 2012.

Raney, R. K., "Maximizing the intrinsic precision of radar altimetric measurements," IEEE Geoscience and Remote Sensing Letters, vol. 10, pp. 1171-1174, 2013.

Ray, C. et al., SAR Altimeter Backscattered Waveform Model, IEEE Trans. GeoSci. And Rem. Sens., Vol. 53, Iss. 2., pp 911 – 919, 2014. DOI: 10.1109/TGRS.2014.2330423.

Wingham D. J. et al., CryoSat: A Mission to Determine the Fluctuations in Earth's Land and Marine Ice Fields. *Advances in Space Research* 37 (2006) 841-871.

9. UNIVERSAL RESOURCE LOCATORS (URL)

SEOM web site	http://seom.esa.int/
ESA Earth Online	http://eopi.esa.int/
Sentinels Online	http://sentinel.esa.int/
Copernicus	http://www.copernicus.eu
CP4O	http://www.satoc.eu/projects/CP4O/
Coastal Altimetry Workshops	http:////www.coastalaltimetry.org
RADS	http://rads.tudelft.nl
SAMOSA	http://www.satoc.eu/projects/samosa/
AVISO+	http://www.aviso.altimetry.fr/

STANDALONE SOFTWARE FOR DETECTING CHANGES IN SAR AND OPTICAL IMAGES

Behnaz Pirzamanbein and Allan A. Nielsen

Department of Applied Mathematics and Computer Science, Technical University of Denmark

ABSTRACT

Change detection is an important application in remote sensing earth observation which leads to identification of significant environmental events, forest and agricultural land monitoring. In this paper, we introduce a standalone software for two well-known change detection methods, omnibus test and iteratively re-weighted multivariate alteration detection (IR-MAD). Omnibus test deals with synthetic aperture radar (SAR) data and detects changes based on computing a sequence of test statistics of covariance matrices and IR-MAD computes the changes between two time points of optical data. Given the availability of earth observation data from different sources and in large amount, the important role of a free software which can deal with big data is apparent.

1. INTRODUCTION

This paper aims to introduce a standalone software for two automatic and popular change detection methods; Omnibus test [3, Sec 2.1] and iteratively reweighted multivariate alteration detection (IR-MAD) [7, Sec 2.2]. The standalone software is available in two formats i.e. graphical user interface (GUI) app and command-line executable. The app has an interactive user interface while the command-line version can be run directly on console given the specified variables (Sec 3). Moreover, the standalone software can handle different formats of images, such as Georeferenced Tagged Image File Format (GeoTIFF), and ENVI binary image coupled with a header file and it has a special module for big data. Depending on the memory capacity of the computer, users can choose to load images into memory or use line by line processing. Furthermore, to decrease the computation time, the software uses parallel computing techniques. For demonstration, two examples are considered using different formats of the software on SAR and optical images. The software is published on github.com/BehnazP/DataBio.

2. METHODS

In this section both change detection methods are briefly explained [for more detail see 3, 7]. Note that in the standalone software, omnibus test is called WISHARTChange and IR-MAD method is called MADChange.

2.1. WISHARTChange

Omnibus test [3] is a change detection method for a sequence of multi-look polarimetric synthetic aperture radar (SAR) data [9]. The method applies a test statistic to quantify the equality of polarimetric covariance matrices ($\mathbf{\Sigma}_{p\times p}$). Detecting if and when a change has occurred is based on the test statistics' significance level on a per-pixel basis or collections of pixels (segments, patches, fields) .

The equivalent number of looks (ENL) in SAR imagery refers to the number of independent pixels of a surface area that is averaged in order to reduce the effect of speckle. Speckle is a noise-like consequence of the coherent nature of the signal transmitted from the sensor. The main assumption of the WISHARTChange method is that the multiplication of the observed signals ($\mathbf{\Sigma}$) by ENL (n), are complex Wishart distributed, i.e. $\mathbf{X} = n\mathbf{\Sigma} \sim W_C(p, n, \mathbf{\Sigma})$.

The method tests a hypothesis that all pixels from different time points ($t \geq 2$) are characterized by the same $\mathbf{\Sigma}$. Therefore the null hypothesis is $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_T (= \mathbf{\Sigma})$ against the alternative (H_1) that at least one of the $\mathbf{\Sigma}_t$, $t = 1, \ldots, T$, is different, i.e., at least one change has occurred. Since the distributions are known, a likelihood ratio test can be formulated which allows one to decide a desired level of significance whether or not to reject the null hypothesis. The algorithm computes the logarithm of the omnibus likelihood ratio test statistic, Q, for testing H_0 against H_1 (see [3] for more detail).

Furthermore, this test can be factored into a sequence of tests involving hypotheses of the form $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ against $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ against $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 \neq \mathbf{\Sigma}_3$, and so forth. The method computes the likelihood ratio test statistic R_t for testing the hypothesis $H_{0,t} : \mathbf{\Sigma}_t = \mathbf{\Sigma}_1$ against $H_{1,t} : \mathbf{\Sigma}_t \neq \mathbf{\Sigma}_1$. The R_t constitute a factorization of Q such that $Q = \prod_{t=2}^{T} R_t$.

The tests are statistically independent under the null hypothesis. In the event of rejection of the null hypothesis at some point in the test sequence, the procedure is restarted from that point, so that multiple changes within the time series can be identified.

2.2. MADChange

Iteratively reweighted multivariate alteration detection (IR-MAD) algorithm [7, 1] is a statistical approach to detect

changes in optical images. This method utilizes an iteration scheme to identify a better background of no-change against which to detect significant change. The method applies canonical correlation analysis (CCA) [6] to multi-spectral images from two time points. The CCA orders the image bands according to similarity based on correlation, rather than spectral wavelength. The differences between corresponding pairs of canonical variates are called the MAD variates,

$$\mathbf{Z} = \mathbf{a}^T\mathbf{X} - \mathbf{b}^T\mathbf{Y}$$

where $\mathbf{X}$ and $\mathbf{Y}$ represents the m-dimensional images, and $\mathbf{a}$ and $\mathbf{b}$ are the eigenvectors from the CCA. Therefore, $\mathbf{a}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{Y}$ are m uncorrelated canonical variates (CVs) with mean zero and variance one for time points one and two, respectively. The correlation between corresponding pairs of CVs, the canonical correlation ($\boldsymbol{\rho}$) is maximized in CCA, therefore we have m uncorrelated MAD variates with mean zero and variance $2(1-\boldsymbol{\rho})$. Since the MAD variates for the no-change observations are approximately Gaussian and uncorrelated, the sum of their squared values after normalization to unit variance ideally follows a chi squared distribution with m degrees of freedom,

$$C^2 = \sum_{i=1}^{m} \frac{Z_i^2}{2(1-\rho_i)} \sim \chi^2(m).$$

In addition the probability of no-change is approximated by $1 - \mathrm{P}\{\chi^2(m) \leq c^2\}$ where c^2 is the actually observed value of C^2 and used as weight, w. In each iteration the value of each image pixel is weighted by corresponding w which is the current estimate of the no-change probability and the image statistics i.e. mean and covariance matrices are recomputed. Iterations continue until the canonical correlations stop changing according to a pre-defined threshold or a maximum number of iterations is reached.

3. STANDALONE SOFTWARE

The standalone GUI app and command-line program for both WISHARTChange and MADChange methods are developed in MATLAB®. At the time of writing the manuscript, the software packages are available for Windows and Linux operating systems. The Mac version is under development.

Depending on memory capacity of the user's computer, the software can handle small and big images by either fitting them into local memory or reading and treating them line by line. The uniqueness of the standalone software is the ability of dealing with different types of image such as GeoTIFF and ENVI, i.e. a flat-binary raster file with an accompanying ASCII header file. For cloud based software working with Google Earth Engine (GEE) [5] see [2] and for MATLAB based packages for specific image type, see [4] for IR-MAD and see [8] for omnibus test.

3.1. Download and installation

Users can download the GUI app and command-line version of the software from github.com/BehnazP/DataBio and install it on their computer. In order to use the standalone software, first the MATLAB Runtime installer provided with the software should be installed. The installer contains all the required MATLAB functions for running the software *without* having MATLAB installed on the user's computer.

The GUI app can run similar to any other software after installation. For executing the command-line version on Windows, users go to the directory of the saved executable file and call the executable and provide the input variables. Linux users, in addition, have to provide the path to installed MATLAB Runtime following the executable file.

3.2. Input

In WISHARTChange, users are required to provide sequence of SAR images, processing modality to read the images into the memory or using the line by line processing, ENL, p-value significance level for the omnibus test, polarization type, number of time points, polarization names as well as the name used for time sequences, saving directory. In addition, there is an option to select and compute the changes in a region of interest (ROI) by providing a binary mask or by choosing the ROI interactively from the frequency map (see Fig. 2 a.).

In MADChange, users are required to provide the two multi-spectral optical images, the name of the multi spectral bands used in the name of the images, a threshold as a criterion to stop the iteration, processing modality to read the images into the memory or using the line by line processing, and saving directory for the results. In addition, there are two options to do a pre-processing scheme by masking the strongest changes and excluding low values related to dark regions [4].

Note that, if users do not provide any extra information except the input images, the app will apply the method based on default values pre-defined in the software.

3.3. Output

The WISHARTChange software outputs a table containing average no-change probabilities and a figure containing maps of changes. In addition, the table is saved as a text file and the maps are saved as an image with three bands, i.e. first change, last change and frequency of the change in same format as the provided images.

The MADChange software outputs a figure showing the probability of no-change. Moreover, the IR-MAD variates are saved for further analysis as an image with same number of bands and same format as the provided images. In addition, there is an option to save the CVs and chi-square images. The software also provides the canonical correlation convergence plot.

3.4. Illustration

In this section, we include two examples using MADChange and WISHARTChange standalone software for visualisation of results and outputs.

In the first example, the WISHARTChange software is applied to SAR data which are acquired by the fully polarimetric Danish airborne SAR system, i.e EMISAR (see [3] for more detail). The example investigates the changes in crops land and forest area of Foulum in Jutland Denmark for six time points; March 21, April 17, May 20, June 16, July 15, and August 16, 1998. The identified changes can help environmental managers to study the development stage of different crops and forest areas. Fig 1 shows first change, last change and the frequency of the change which occurred in the area. For demonstration, a region of interest (ROI) is chosen interactively (Fig 2 a.) and the probability of no-change is computed and shown in Fig 2 b. The chosen ROI is a peas farm land and from the results one can conclude there has been significant changes between all six points.

In the second example, the MADChange software is applied to GeoTiff data which is obtained from GEE. The example investigates the changes in agricultural land and forest areas of Javier in North Spain for two time points; October 21 2016 and October 11 2017. The identified changes can help forest managers and land owners to identify forest disease and monitor the changes in their lands in early stages. Fig 3 shows the probability of no-change in the Javier region. The first three bands of IR-MAD variates is shown in Fig 4 as a RGB image corresponding to greatest canonical correlation. The figures indicate no-change in forest area, small changes in some of the grass lands around agricultural areas and major changes in agricultural land.

4. CONCLUSION

This paper introduces standalone software for SAR and optical images which can be used in many application areas where analysing and monitoring spatio-temporal dynamics is important. The software packages are called WISHARTChange and MADChange and are based on two automated and popular change detection methods; omnibus test and IR-MAD, respectively. The software is published on github.com/BehnazP/DataBio in two versions; GUI app and command-line executable. The uniqueness of the standalone software is its flexibility to handle different formats of images and also its capability to handle big data.

The outputs of the standalone software are shown in two examples. The results provide insights for detecting changes that might help environmental managers and policy makers.

5. ACKNOWLEDGEMENT

This work is funded by, DataBio (www.databio.eu), the European Unions Horizon 2020 research and innovation programme under grant agreement No 732064.

REFERENCES

[1] M. J. Canty. *Image analysis, classification and change detection in remote sensing: with algorithms for ENVI/IDL and Python*. CRC Press, 2014.

[2] M. J. Canty and A. A. Nielsen. Spatio-temporal analysis of change with sentinel imagery on the Google Earth Engine. In *ESA Conference on Big Data from Space (BiDS)*, pages 126–129, 2017.

[3] K. Conradsen, A. A. Nielsen, and H. Skriver. Determining the points of change in time series of polarimetric SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 54(5):3007–3024, 2016. doi: 10.1109/TGRS.2015.2510160.

[4] N. Falco, P. R. Marpu, and J. A. Benediktsson. A toolbox for unsupervised change detection analysis. *International Journal of Remote Sensing*, 37(7):1505–1526, 2016. doi: 10.1080/01431161.2016.1154226.

[5] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017. doi: 10.1016/j.rse.2017.06.031.

[6] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. doi: 10.2307/2333955.

[7] A. A. Nielsen. The regularized iteratively reweighted mad method for change detection in multi-and hyperspectral data. *IEEE Transactions on Image processing*, 16(2): 463–478, 2007. doi: 10.1109/TIP.2006.888195.

[8] A. A. Nielsen, K. Conradsen, H. Skriver, and M. J. Canty. Visualization of and software for omnibus test-based change detected in a time series of polarimetric SAR data. *Canadian Journal of Remote Sensing*, 43(6): 582–592, 2017. doi: 10.1080/07038992.2017.1394182.

[9] F. T. Ulaby and C. Elachi. *Radar polarimetry for geoscience applications*. Norwood, MA, Artech House, Inc., 1990.

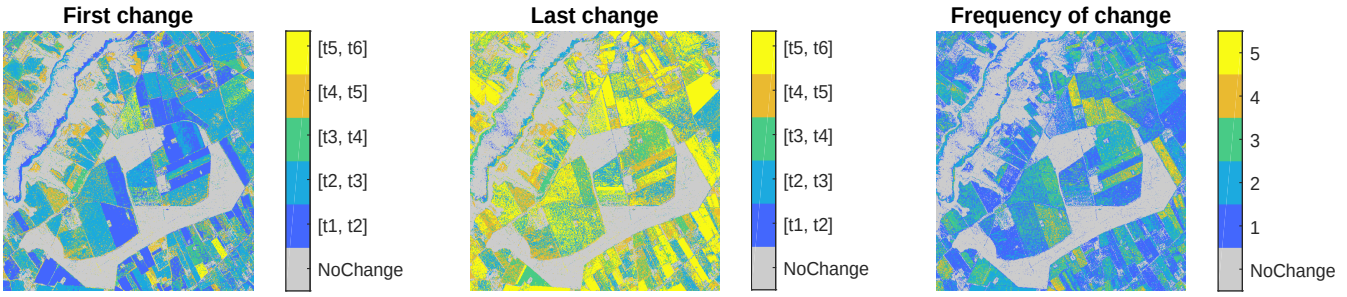
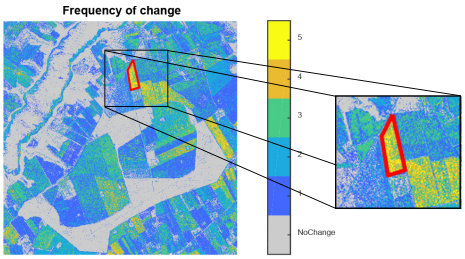


Fig. 1. Output from WISHARTChange software for the Foulum area in Denmark for six time points, on the left is the map of first changes, middle is the map of last changes, and the right shows the frequency of the changes.



a. Peas farm land

P-values

	t1 = ... = t6	t2 = ... = t6	t3 = ... = t6	t4 = ... = t6	t5 = t6
Omnibus	0	0	0	0	1.0000e-05
t1 = t2	0.0008	NaN	NaN	NaN	NaN
t2 = t3	0.0035	0.0036	NaN	NaN	NaN
t3 = t4	0	0	0	NaN	NaN
t4 = t5	0	0	0	0.0050	NaN
t5 = t6	0.0018	0.0015	4.9000e-04	0.0001	1.0000e-05

b. Averaged no-change probabilities

Fig. 2. Output from WISHARTChange software; a. shows the region of interest chosen interactively and b. shows the average no-change probabilities for the peas farm land selected in a).

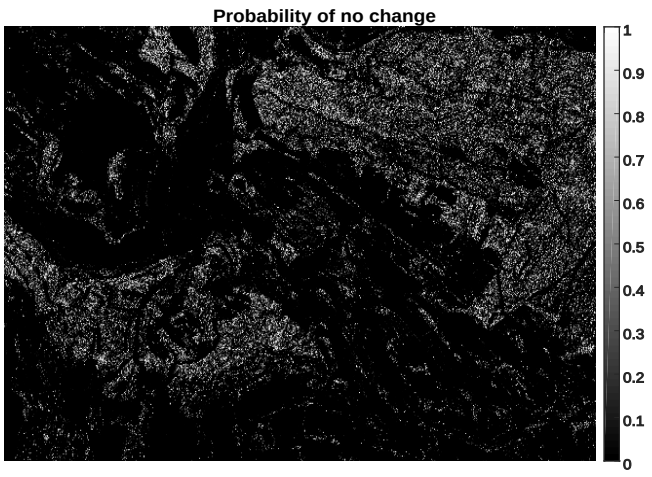


Fig. 3. Output from MADChange software showing probability of no change in Javier region in North Spain between October 21 2016 and October 11 2017.

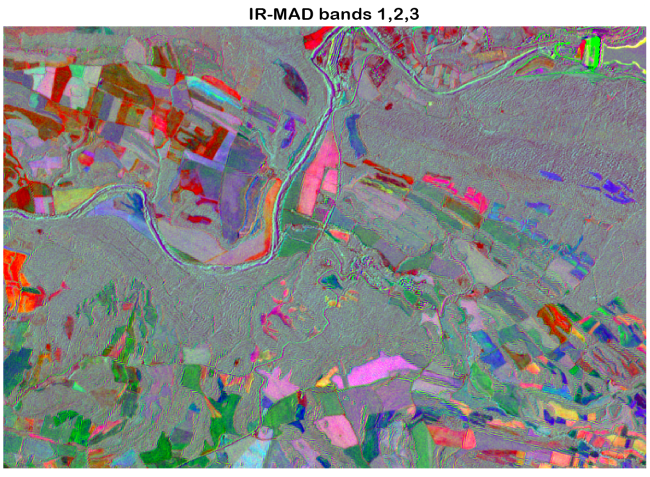


Fig. 4. The image of first three IRMAD variates show different changes in agricultural lands, forest areas and grass lands based on the output from MADChange software.

D-MOSS: AN INTEGRATED DENGUE EARLY WARNING SYSTEM DRIVEN BY EARTH OBSERVATIONS IN VIETNAM

Gina Tsarouchi[1], Iacopo Ferrario[1], Quillon Harpham[1], Alison Hopkin[1] and Darren Lumbroso[1]

[1]HR Wallingford Ltd, Howbery Park, Wallingford, Oxfordshire OX10 8BA, UK

ABSTRACT

D-MOSS, Dengue MOsquito Simulation from Satellites, is a dengue fever early warning system for Vietnam being developed by a project sponsored by the UK Space Agency's International Partnership Programme. It will give beneficiaries several months advance warning of likely outbreaks of dengue fever. Earth Observation datasets are combined with health and water availability information to produce a new integrated dengue forecasting model. The system will also include a water assessment module that will provide the additional benefit of improving water management in Vietnam's transboundary river basins.

Index Terms— Earth Observations, dengue, water availability, forecasting, early warning systems, data integration

1. INTRODUCTION

Before 1970 only nine countries had experienced severe dengue fever epidemics[1]. Today the disease is endemic in 141 countries, affecting 390 million people globally[2]. The total global annual cost of dengue fever has been estimated to be almost US$9 billion per year, which is three times that of cholera and over four times that of gastroenteritis [1]. Since 2000, there has been an increase of over 100% in the number of cases of dengue fever in Vietnam, with ~185,000 cases occurring in 2017[3].

In Vietnam there is currently no system in place to forecast the probability of future dengue outbreaks. Recently the epidemiological situation in Vietnam has been worsened by the failure to maintain adequate control of the *Aedes aegypti* species of mosquito that spreads dengue fever. The D-MOSS project is developing a forecasting system that will allow public health authorities to identify areas of high risk for disease epidemics before an outbreak occurs, in order to target resources to reduce epidemic spreading and increase disease control.

Earth Observation (EO) datasets are combined with health and water availability information to produce a new integrated dengue forecasting model. The model links EO data with weather forecasts and a hydrological model to predict the likelihood of future dengue epidemics up to eight months in advance.

The dengue forecasting tool also includes a water availability module, which will help to improve water management in Vietnam's transboundary river basins where there is a paucity of hydro-meteorological information.

The D-MOSS project is funded by the UK Space Agency's International Partnership Programme and led by HR Wallingford, working with the London School of Hygiene and Tropical Medicine, the UK Met Office and Oxford Policy Management in the UK, and with the following international partners: the United Nations Development Programme, the World Health Organisation, the Vietnamese Institute of Meteorology, Hydrology and Climate Change, the Pasteur Institute Ho Chi Minh City, and the National Institute of Hygiene and Epidemiology in Vietnam.

2. METHODOLOGY

The D-MOSS project is developing a suite of innovative tools that will allow beneficiaries to: issue alerts for dengue fever (with a view to develop the same for Zika virus); and provide assessments of vector-borne disease risk under future climate and land-use change scenarios. In addition, forecasts of water scarcity will be made and incorporated into the dengue early warning tool. The integrated suite combines data from EO-based sources, climate forecasting and land-surface modelling.

D-MOSS integrates multiple stressors such as water availability, land-cover, precipitation and temperature in order to forecast future outbreaks of dengue fever. The approach uses a common spatio-temporal analysis 'grid' with a 'Polygon Series' structure [2] to integrate historical stressor datasets with each other and with historic dengue fever incidents, which are then input into a statistical model which provides forecasts based on future seasonal forecasts of these stressors.

The architecture of the solution relies on open and non-proprietary software, where possible, and on flexible deployment into platforms including cloud-based virtual storage and application processing. D-MOSS makes use of open-source solutions where possible (such as LINUX, POSTGIS and Mongo RDF) and widely known development languages and tools (such as Java, Python, HTML, XML). As such, the reliance on proprietary third party software and the knowledge of such software in the future is reduced. Moreover, the opportunity to replicate the

[1]http://www.searo.who.int/entity/vector_borne_tropical_diseases/data/data_factsheet/en/

[2]http://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[3]https://www.garda.com/crisis24/news-alerts/87956/vietnam-increase-in-dengue-fever-cases-in-2017

doi:10.2760/848593

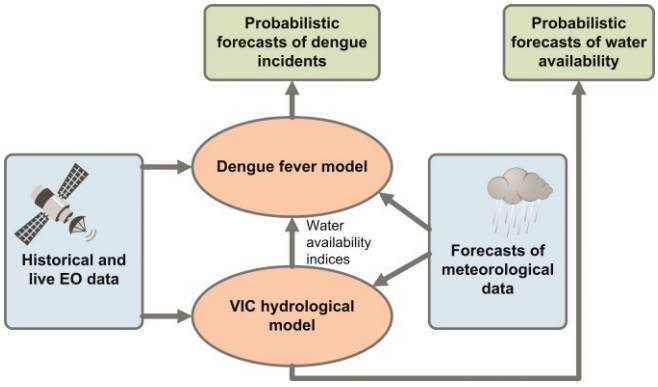


Figure 1: System overview

generic design in other parts of the world and for other diseases is increased. A simplified overview of the system is shown in Figure 1.

2.1. Development of a water availability forecasting system

We are developing an integrated modelling system that combines: EO data, climate forecasting and hydrological modelling. This system: (a) generates water stress assessments; and (b) projects longer-term impacts of multiple stressors on water resources that feed into the dengue prediction tool. UK Met Office seasonal forecasts are used to drive the VIC hydrological model [3], whose outputs are combined with EO datasets and translated to indicators of water stress such as commonly used indicators of drought (e.g. Standardized Precipitation Index). The system includes three EO-based components (i.e. land-use, weather and water resources) and generates monthly forecasts of water stress. The system will be calibrated against historical data. The water availability forecasts are subsequently fed into the dengue prediction tool.

2.2. Development of dengue early warning system

The integrated system described in Section 2.1 provides an assessment of the water resource situation. This feeds into new statistical forecasting models of disease incidence, based on a spatio-temporal Bayesian hierarchical mixed-modelling approach [4] & [5]. The dengue early warning system model integrates the water stress forecast with a range of other covariates important for dengue transmission to forecasts of dengue incidence, up to six months in advance. These forecasts are directly relevant to water management activities. The additional variables such as water stress, related to the surveillance of the *Aedes aegypti* mosquito vectors, are included to strengthen the model in order to enhance its ability to forecast dengue outbreaks and associated morbidity. Data on historical dengue outbreaks will be used as input to the model.

Spatio-temporally explicit models are used to fit non-linear and interactive relationships between each of the covariates and the historical dengue incidence (1998-2018) for all 63 provinces of Vietnam. The model provides predictions of: (a) Monthly dengue incidence including uncertainty; (b) Probability of exceeding the outbreak threshold; (c) Projected incidence under different water management scenarios (as determined following consultation with stakeholders).

3. RESULTS

The D-MOSS project is within its first year, of a three-year programme and is currently focused on the platform development alongside bringing together the key input data streams and engaging with the government in Vietnam to ensure that all components are fit for purpose.

The portrayal system (Figure 2) will be designed to communicate the dengue and water availability forecasts to the Vietnamese Ministries of Health and Natural Resources and Environment, respectively. It will be used within an incident room in Hanoi with other users able to access the website at regional centres around the country. The user interface will also incorporate supporting information on recommended actions to be taken, provided by the decision makers and based on the forecasts and associated uncertainty.

As part collecting, reviewing and integrating data within D-MOSS, we have been focusing on some key variables that have been shown to influence dengue outbreaks. These variables were then assessed for availability, as well as review, integration and processing so that the data is of use before going into the predictive model. Here we present some results on the analysis done for different rainfall products, given here as one of the key, typical hydro-meteorological variables.

The satellite products chosen to provide rainfall estimates are: the TRMM 3B42v7 [6] (hereafter referred to as TRMM) and the Integrated MultisatellitE Retrievals for GPM [7] (GPM IMERG v5, hereafter referred to as GPM). These are multi-satellite-gauge combination products that assimilate data from a core satellite and a constellation of other satellites from partner agencies, equipped with InfraRed and Passive MicroWave sensors.

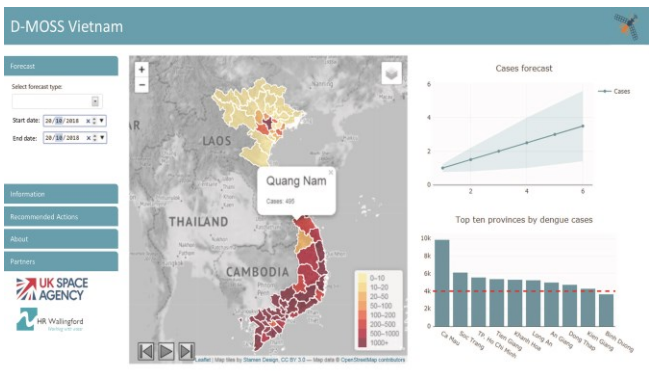


Figure 2: D-MOSS user interface

doi:10.2760/848593

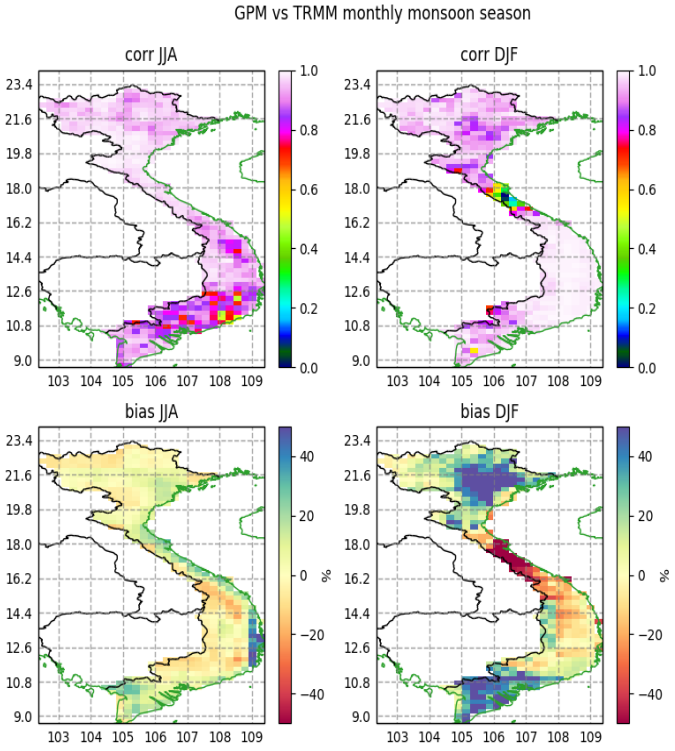


Figure 3: Mean correlation coefficient and mean bias between GPM and TRMM, with the latter taken as reference, for rainfall monthly totals.

To support the validation of the satellite datasets an observational reference dataset of 248 ground rainfall stations was used, provided by the Institute of Meteorology, Hydrology and Climate Change in Vietnam (hereafter referred to as IMHEN-obs).

Figure 3 shows the mean correlation coefficient and the mean bias between GPM and TRMM with the latter taken as reference, for rainfall monthly totals. The comparison is carried out between 12-03-2014 (when GPM starts to be operative) and 31-12-2017, and results are aggregated over two seasons: (a) the wet season corresponding to the central months of the southwest monsoon (June-July-August, JJA); and (b) the dry season corresponding to the northeast monsoon (December-January-February, DJF). Over the selected period the two datasets show an overall good correlation, with different spatial patterns between JJA and DJF. Interestingly there is a region with very low correlation around central Vietnam during the dry season. Regarding the rainfall magnitude (bias plots), during the wet season (JJA) the large differences are found along the coasts, over the mountain regions and over the Mekong river delta in the south. The bias is much larger in the dry season (DJF), however, since in this case the rainfall amount is typically low even small differences could cause large biases.

In Figure 4 the mean performance of GPM and TRMM is assessed against the observations dataset IMHEN-obs, for rainfall monthly totals aggregated by climatic zones. The analysis period is between 12-03-2014 and 31-11-2015. In general, GPM is better in capturing the dynamic of rainfall, with an outstanding exception for the North Central Coast climatic zone where TRMM shows higher correlation with ground stations. However, GPM is found to overestimate the rainfall magnitude over all climatic zones but the Central Highlands.

4. DISCUSSION

Precipitation estimates interact with sensor sensitivity, spatio-temporal resolution, climate, type of rain, topography and the retrieval algorithm; the complex nature of these interactions makes it difficult to define probabilistically the precipitation estimates [8].

For TRMM and GPM the random component of the estimate error is computed following the Huffman (1997) method [9]. However, recognizing the simplistic approach adopted, NASA is currently trying to implement new methodologies to define the random error [8] & [10]. The issue of data quality has been recently addressed in the new GPM products, where a quality flag value has been devised and added to the data [11]. TRMM has no quality index associated.

In the future, provided that a larger number of rainfall ground stations can be made available, the random error may be estimated following the methodology described in Maggioni et al. (2014) [10]. The results from this methodology could inform a bespoke quality index for Vietnam for the period of interest.

5. CONCLUSIONS

EO data can help countries understand the dynamics multiple stressors on the health and water sectors, especially in regions with poor or non-existent ground monitoring. When compared to ground stations, remote-sensing products enable a more accurate representation of the spatial variation of meteorological parameters, which may vary significantly at the local scale, particularly in regions with high elevation variation. EO data also enables scalability of the solution up to national or even international level. However, the associated evidence base is only just emerging and applying this work using remote sensing data is expected to make a significant contribution. The resultant tools also include a water assessment module that will feature the additional benefit of improving water management in Vietnam's transboundary river basins.

This multidisciplinary application of open socio-environmental modelling also extends to on-the-ground practitioners tasked with acting upon the predictions in a way that will best mitigate the risks; particularly in conveying results, changing behaviour in allocating and applying budgets, and responding in advance of potential outbreaks.

The D-MOSS project aims to develop an early warning system to improve dengue epidemic prevention and increase

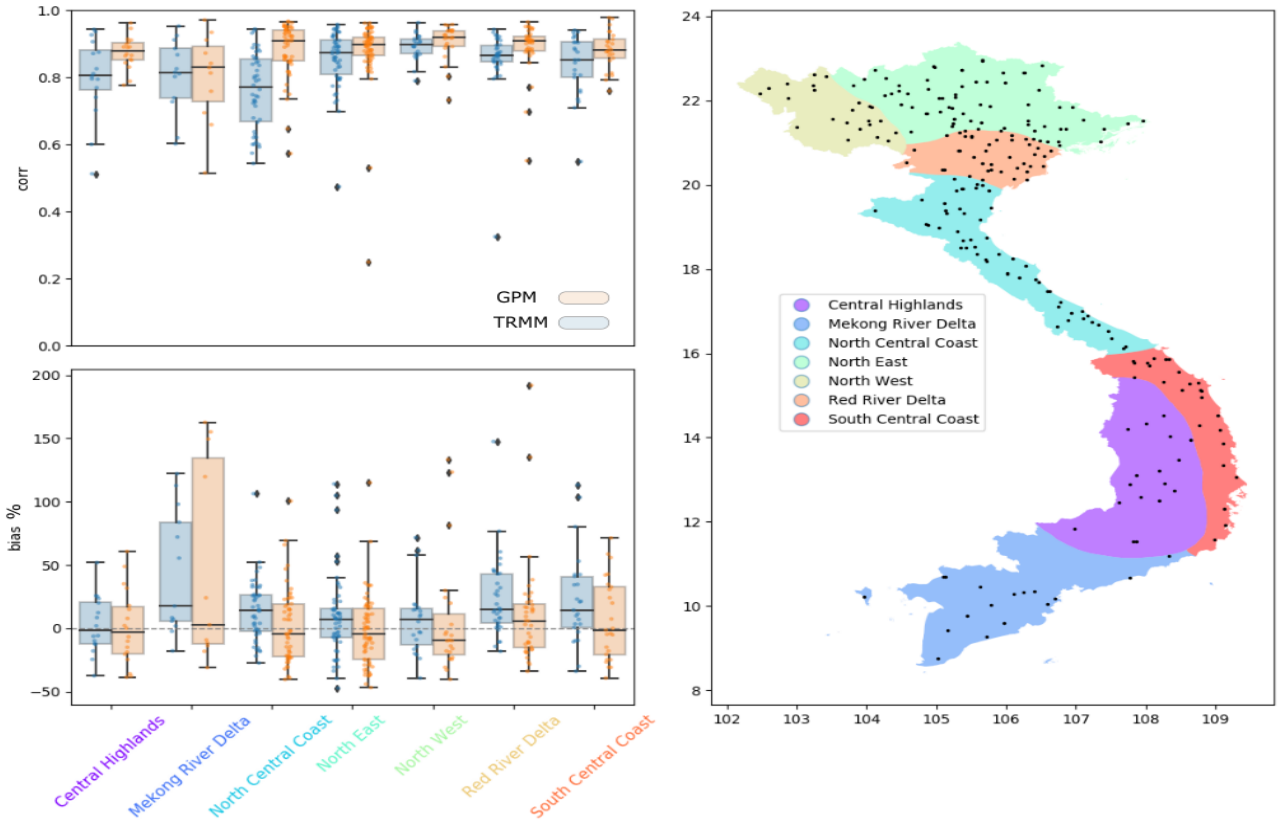


Figure 4: performance of GPM and TRMM against the IMHEN-obs dataset

disease control capacity in Vietnam. The establishment of a dengue forecasting system will assist the Vietnamese public health authorities to identify current areas of high risk of infectious disease epidemics, in order to effectively target resources to ensure effective disease control. From a water resources perspective, given that seven of the nine major river basins that drain to Vietnam are transboundary in nature and are shared between two and five countries, the application of EO-based information will help the Vietnamese Ministry of Natural Resources and Environment to improve their water resources monitoring and management in transboundary river basins.

6. REFERENCES

[1] Shepard, D. S., et al. (2016). The global economic burden of dengue: a systematic analysis. Lancet Infectious Diseases.

[2] Harpham, Q.K. (2018). Using spatio-temporal feature type structures for coupling environmental numerical models to each other and to data sources, under review for publication, Open University, Milton Keynes, UK.

[3] Liang, X., et.al. (1994), A simple hydrologically based model of land surface water and energy fluxes for general circulation models, J. Geophys. Res., 99(D7), 14415–14428.

[4] Lowe R, et al. (2017). Climate services for health: predicting the evolution of the 2016 dengue season in Machala, Ecuador. Lancet Planetary Health

[5] Lowe R, et al. (2014). Dengue outlook for the World Cup in Brazil: an early warning model framework driven by real-time seasonal climate forecasts. Lancet Infectious Diseases, 14(7)

[6] Huffman, G. J., et al. (2007). The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. J. Hydrometeor, 8(1), 38–55.

[7] Huffman, G., et al. (2014). Integrated Multi-satellitE Retrievals for GPM (IMERG), version 4.4. NASA's Precipitation Processing Center.

[8] Kirstetter, P., et al. (2018). Probabilistic precipitation rate estimates with space-based infrared sensors. QJR Meteorology.

[9] Huffman, G. J. (1997). Estimates of root-mean-square random error for finite samples of estimated precipitation. Journal of Applied Meteorology, Volume 36.

[10] Maggioni, V. et al. (2014). An error model for uncertainty quantification in high-time-resolution precipitation products. Journal of Hydrometeorology, Volume 15.

[11] Huffman, G. (2017). IMERG Quality Index. [Online, accessed 28 September 2018]. https://pmm.nasa.gov/sites/default/files/document_files/IMERG_QI-rev.pdf

7. ACKNOWLEDGEMENTS

The authors would like to acknowledge: (a) the UK Space Agency who is funding this project as part of the International Partnerships Programme; (b) ESA, NASA and USGS for access to the input EO data streams ingested into the D-MOSS platform.

AUTOMATIC GENERATION OF SENTINEL-1 DINSAR CO-SEISMIC MAPS

Fernando Monterroso[1,2], Claudio de Luca[2] , Manuela Bonano[2,3], Riccardo Lanari[2] , Michele Manunta[2], Mariarosaria Manzo[2], Giovanni Onorato[2], Ivana Zinno[2], Francesco Casu[2]

1. Univeristy of Naples "Parthenope", Naples, Italy
2. IREA - CNR, Italy.
3. IMAA- CNR, Italy.

ABSTRACT

This paper presents an automatic tool for the generation of co-seismic displacement maps through the satellite Differential Synthetic Aperture Radar Interferometry (DInSAR) technique. By benefiting from the mostly global availability of Sentinel-1 SAR data and the on-line earthquake catalogs, the tool retrieves information about the depth and magnitude of recent earthquakes and triggers, if necessary, the interferometric process over the area affected by the seismic event. The generated DInSAR products will be openly available for the scientific community, to create a global database of interferometric co-seismic deformation maps giving a support to scientific users, especially those non-expert of SAR data processing. Moreover, this tool not only will contribute to expand the use of DInSAR products in the geoscience field, but also can play a key role in the support of the Civil Protection authorities during the seismic crisis.

Index Terms— *Earthquakes; DInSAR; Sentinel-1; Automatic Processing.*

1. INTRODUCTION

In the recent years, the Synthetic Aperture Radar (SAR) data have become more and more popular within the Earth Observation (EO) context and are used in many scientific fields and applications related to both natural (volcanoes, earthquakes, landslides) and man-made (urban and infrastructure monitoring) hazards. Nowadays the scientific community can benefit from the huge space borne SAR archives collected in the last 20 years. Indeed, the data acquired since 1992 to 2011 by both ERS-1/2 and ENVISAT missions, operating at C-band, as well as the current Italian COSMO-SkyMed (CSK) and the German TerraSAR-X (TSX) X-band SAR constellations, have strongly contributed to SAR data diffusion and popularity.

Moreover, a massive and ever increasing data flow is further supplied by the Sentinel-1 SAR mission of Copernicus European Programme, which is composed by two twin satellites operating in C-band that guarantee a repeat pass down to 6 days (in most regions), and is characterized by both a "free and open" access data policy and a global coverage acquisition strategy [1]. The space-borne Differential SAR Interferometry (DInSAR) is one of the most used techniques for the investigation of Earth's surface deformation phenomena from SAR data. Such a technique permits, indeed, to retrieve ground deformation maps with centimetre accuracy [2-3] starting from SAR scenes of the same area of interest from different orbital position and at different epochs.

Especially for seismic events, the satellite SAR systems provide a large coverage on the ground that is essential for estimating the deformation field entails by seismic events. In this context, the Sentinel-1 system, by using the Terrain Observation by Progressive Scans (TOPS) technique [4], has been designed with the specific aim of natural hazards monitoring via SAR Interferometry, indeed, it acquires SAR images using the Interferometric Wide Swath (IWS) mode that guarantees a very large ground coverage of about 250 km. These characteristics sum up with the global coverage policy and an acquisition rate of ten of TByte per day.

According to USGS records [5], from 1992 to 2016, there have been about 3700 earthquakes with significant magnitudes (see Figure 1). More than 200 studies exploited DInSAR data for retrieving the source model of about 130 earthquakes [6]. Supposed that not all the earthquakes can be studied via DInSAR techniques, it anyway appears that the ratio between the number of occurred earthquakes and the number of DInSAR studies is still too low. However, since the launch of Sentinel-1 SAR satellite missions in 2014 and 2016, the availability and accessibility of SAR images dramatically increased, allowing us to obtain co-seismic displacement maps in a short time frame and anywhere in the world.

Considering the relevance of the satellite interferometric analysis for the hazards monitoring, and the availability of new radar systems such as Sentinel-1, which are characterized by a high reliability level, it is possible to implement fully automatic services for the generation of co-seismic DInSAR products.

doi:10.2760/848593

Accordingly, the aim of this work is the development of a systematic tool for the generation of Sentinel-1 DInSAR co-seismic displacement maps via an automatic and fully unsupervised procedure which is activated immediately after the occurrence of an earthquake above of a defined magnitude. The procedure is triggered via an automatic query to the available on line catalogues like those provided by United States Geological Survey (USGS) and National Institute of Geophysics and Volcanology of Italy (INGV). The developed tool relies on widely common IT methods and protocols, making it not specifically tied to a defined architecture, thus implying its portability, in view also of the European Commission Data and Information Access Services (DIAS) [7] where satellite data (mainly Sentinel) and processing facilities are co-located to reduce the transfer time during their processing.

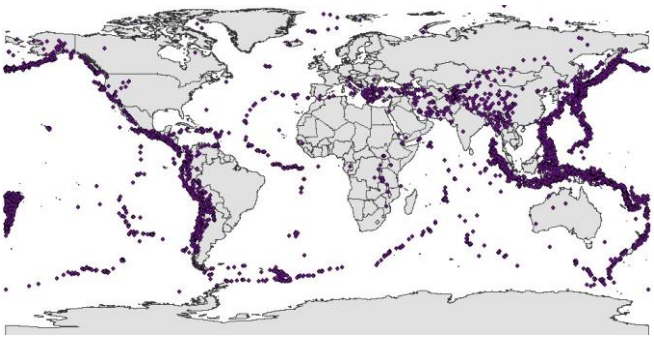


Figure 1: Significant seismic events wold map (1992-2016).

2. AUTOMATIC DINSAR PROCESSING WORKFLOW

In this section we describe the details of the workflow (Figure 2) for the automatic generation of co-seismic displacement maps by using Sentinel 1 images.

First of all, the workflow starts from the extraction of earthquake information, such as Epicentre, Magnitude, Time, which is retrieved from the on-line public available web catalogues, as those provided by main international geophysical institutions (e.g. USGS [8], INGV [9]). Such services systematically provide real-time earthquake information in different standard formats (QuakeML, TXT, geoJSON, ...) and are accessible via subscription feeds that are updated with a defined frequency. The system is not limited to an earthquake catalog interface, for this case we are using a geoJSON format which is a standard format designed to represent simple geographic elements, together with their non-spatial attributes, based on JavaScript Object Notation [10].

The relevant earthquake information is collected, in accordance to an empirical magnitude and depth relation, considering that only high magnitude (> Mw 6.0) and relatively shallow earthquakes (typically < 20 km) very likely induce a surface deformation that is detectable via DInSAR. Among the earthquakes that respect the relation, only those with the epicentre on land (or even on water but that can likely induce detectable deformation on land) are actually processed according to the next steps of the procedure. For the SAR data retrieval, the system starts an automatic query to detect the available SAR Sentinel-1 data acquisitions in the area of interest (that cover the area very likely interested by the earthquake-induced deformation). This query identifies all the tracks from both orbits (ascending and descending passes). The query is performed over an area whose extension depends on the earthquake magnitude and depth (Mw $\geq$ 6 and Depth $\leq$ 20 km). The relation between magnitude, depth and area is derived from theoretical and empirical considerations and is susceptible of further tuning and refinement.

Once the tracks covering the earthquake area have been identified, the system retrieves all the available SAR data up to 30 days before the event (or at least 1 pre-event image even in a larger time span), in order to allow the generation of the co-seismic interferograms. The data retrieval, and accordingly the subsequent DInSAR processing, remains active up to 30 days after the event.

The tool works with the subsequent DInSAR processing, which is carried out by using the Parallel Small BAseline Subset (P-SBAS) processing chain [11-12] implemented at IREA-CNR. Indeed, instead of performing the whole SBAS processing, the P-SBAS chain is exploited up to the interferogram generation step, so that the processing can also benefit of the parallelization strategies implemented within P-SBAS.

The processing of the different tracks can be carried out in parallel, while actually their execution depends on the available computing resources and on the effective temporal acquisition of the SAR data. This strategy allows the processing of a huge amount of data with a significant reduction of the total elapsed time of elaboration, preserving the precision and accuracy of the generated interferometric results.

doi:10.2760/848593

A processing prioritization of the different tracks on the basis of the post-event acquisition time has been implemented (according to a First come-First served policy). The described procedure is coded in Linux Bash, making it highly portable and avoiding the installation of any additional software, tool or library.

The tool provides wrapped interferograms and displacement maps (unwrapped interferograms converted in centimetres) in the satellite Line of Sight (LOS). The elapsed time needed to generate one co seismic displacement map from the availability of the post-seismic acquisition, is of about 30 minutes (plus the data download time).

The output data are provided according to the formats defined within the European Plate Observing System (EPOS) [13] research infrastructure. In particular, the products are provided in geoTiff, while metadata follow the ISO 19115, and will be made openly available through the EPOS portal. As final consideration, it is worth noting that, although tested with Sentinel-1 data, the implemented tool is independent from the exploited SAR acquisitions. The only dependency is on the catalogue interface that may require the implementation of an appropriate wrapper.

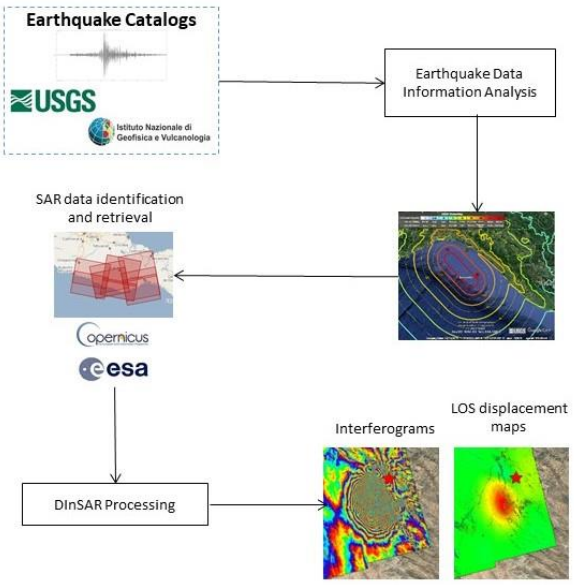


Figure 2: Simplified general workflow of Automatic co-seismic displacement maps triggered by significant earthquakes.

3. PRELIMINARY RESULTS

The system has been implemented on in-house computing facilities and has been tested through a controlled experiment with several significant earthquakes. In total we have processed 15 earthquakes from September 2017 until October 2018, for which we have automatically generated 96 interferograms and displacement maps, using approximately 12 images per earthquake (depending on the availability of data on the investigated region). In the Figure 3 the results are displayed specifically for one earthquake of the seismic sequence that occurred in Indonesia on July and August of 2018.

Indeed, during this period four earthquakes of magnitude larger than 6 occurred in different areas of Western Lesser Sunda Islands close of Mataram City. The first earthquake was on 28 July with a Magnitude of 6.4 and a Depth of 140 km. The second one occurred 7 days later with a Magnitude of 6.9 and a Depth of 34 km. Other two earthquakes occurred in the same day with a Magnitude of 6.9 and 6.3 respectively and 34 and 16 km of depth [8].

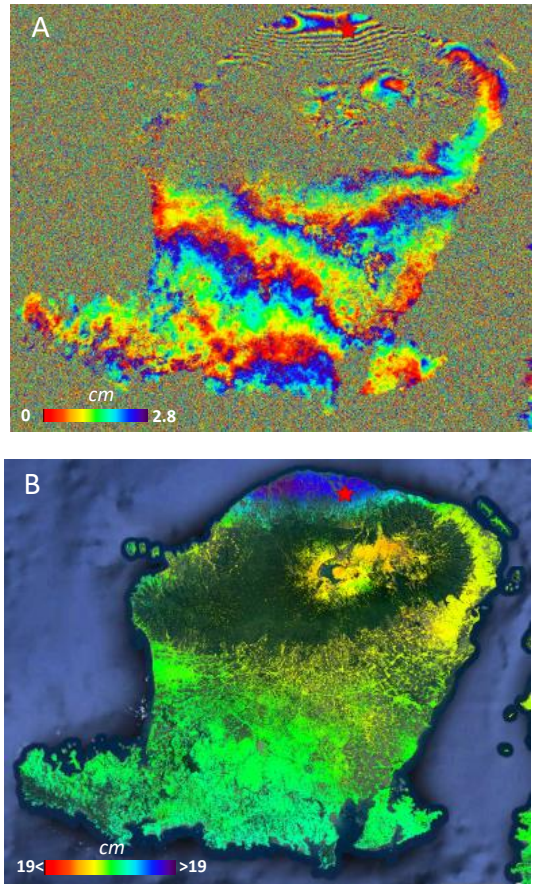


Figure 3: (A) co-seismic interferogram and (B) co-seismic displacement map of Loloan Indonesia Earthquake (August 5th, 2018). The pre and post event data have been acquired on 2018-07-24 and 2018-08-05 from descending orbit.

doi:10.2760/848593

The occurrence of such huge events in these periods represents a very valuable test case to verify and validate the implemented procedure. An example of the obtained results is depicted in Figure 3, which is relevant to Obelobel, Indonesia Earthquake occurred on August 5th;

4. DISCUSSION AND CONCLUSIONS

Thanks to the big amount of available SAR data collected by the Senitnel-1 constellation, we proposed a systematic tool to generate worldwide co-seismic displacement maps in an unsupervised way, being triggered by the earthquake occurrence. Such a tool not only will contribute to expand the use of DInSAR products in the geoscience field, but also can play a key role in the support of the Civil Protection authorities during the seismic crisis.

Moreover, the proposed system can be further improved by including additional post-processing procedures that add value to the generated DInSAR results. It is, for example, the case of tools capable to automatically model the seismic source from the considerable amount of DInSAR displacement maps obtained through the proposed system. These products could contribute to the study of global tectonic earthquake activity by integrating other parameters that allow understanding the seismic source and the behavior of a fault interested by a deformation processes.

Finally, the proposed tool can be used as the basis for the implementation of a massive database of DInSAR-derived co-seismic displacement maps. This database, by exploiting the available SAR data archives acquired by different satellite systems, in theory can be populated starting from 1992, thus providing the scientific community of a huge repository to better investigate the dynamics of surface deformation in the seismic zones around the Earth.

5. ACKNOWLEDGEMENTS

This work was supported by the Italian Civil Protection Department, the ESA's GEP project, the EPOS-IP project of the European Union Horizon 2020 R&I program (grant agreement 676564), the I-AMICA (PONa3_00363) project, and the IREA-CNR/Italian Ministry of Economic Development DGS-UNMIG agreement.

6. REFERENCES

[1] R. Torres, P. Snoeij, D. Geudtner, D. Bibby, M. Davidson, E. Attema, P. Potin, B. Rommen, N. Floury, M. Brown, I. Traver, P. Deghaye, B. Duesmann, B. Rosich, N. Miranda, C. Bruno, M. L'Abbate, R. Croci, A. Pietropaolo, M. Huchler, and F. Rostan, 2012. GMES Sentinel-1 mission. Remote Sens. Environ., vol. 120, pp. 9-24, 2012

[2] Massonnet, D. et al., "The displacement field of the Landers earthquake mapped by radar interferometry," Nature, vol. 364, no. 6433, pp. 138–142, Jul. 1993.

[3] Burgmann, R., Rosen, P.A., Fielding, E.J., 2000. Synthetic aperture radar interferometry to measure Earth's surface topography and its deformation. Annu. Rev. Earth Planet. Sci. 28, 169–209 (May)

[4] De Zan, F.,Monti Guarnieri, A.M., 2006. TOPSAR: terrain observation by progressive scans. IEEE Trans. Geosci. Remote Sens. 44 (9), 2352–2360 (Sept.).

[5] USGS. United States Geological Survey, Earthquakes hazard program, https://earthquake.usgs.gov/earthquakes/browse/

[6] Funning, G., Garcia A. 2018 "A systematic study of earthquake detectability using Sentinel-1 InterferometricWide-Swath data", Geophisical Jornal International, publish in: https://doi.org/10.1093/gji/ggy426

[7] DIAS. The upcoming Copernicus Data and Information Access Services (DIAS), http://copernicus.eu/news/upcoming-copernicus-data-and-information-access-services-dias

[8] USGS. United States Geological Survey, Earthquakes hazard program, https://earthquake.usgs.gov/earthquakes/feed

[9] INGV, National Institute of Geophysics and Volcanology, http://cnt.rm.ingv.it/feed/atom/all_week

[10] GeoJSON. http://geojson.org/

[11] Casu et al. 2014 SBAS-DInSAR Parallel Processing for Deformation Time-Series Computation" in IEEE journal of selected topics in applied earth observations and remote sensing, VOL. 7, NO. 8, AUGUST 2014.

[12] Zinno et al., "National Scale Surface Deformation Time Series Generation through Advanced DInSAR Processing of Sentinel-1 Data within a Cloud Computing Environment," in IEEE Transactions on Big Data, 2018, accepted.

[13] EPOS, European Plate Observing System, https://www.epos-ip.org/tcs/satellite-data

TREE HEALTH ASSESSMENT FOR SATELLITE CALIBRATION AND VALIDATION USING MULTISPECTRAL TERRESTRIAL LIDAR

Junttila, S.[1,2], Vastaranta, M.[2,3], Holopainen, M.[1,2], Lyytikäinen-Saarenmaa, P.[1], Hyyppä, H.[2,4] & Hyyppä, J.[2,5]

[1] Department of Forest Sciences, University of Helsinki, Helsinki, Finland
[2] Centre of Excellence in Laser Scanning Research,
Finnish Geospatial Research Institute (FGI), Masala, Finland
[3] School of Forest Sciences, University of Eastern Finland,
Joensuu, Finland
[4] Department of Built Environment, Aalto University, Aalto, Finland
[5] Department of Remote Sensing and
Photogrammetry, Finnish Geospatial Research Institute (FGI), Masala, Finland

ABSTRACT

Tree and forest health is a global issue in the face of climate change as novel stress is imposed from a variety of biotic and abiotic stresses. Forest health assessments are mainly based on visual assessment that is prone to error and bias. In this paper, we investigated the utilization of terrestrial lidars operating at 905 nm and 1550 nm wavelengths in the objective assessment of tree health condition in a mature forest area subjected to tree decline due to infestation by the European spruce bark beetle (*Ips typographus* L.). We found that multispectral lidar intensity metrics from the canopy and the stem were able to predict tree health, expressed in terms of attack score level, with high accuracy (Adj. $R^2 = 0.87$). We concluded that multispectral intensity metrics have great potential in tree health assessments for the calibration and validation of satellite or other large-area remote sensing methods.

Index Terms— Forest health, multispectral lidar, terrestrial lidar, *Ips typographus*, satellite calibration, lidar intensity.

1. INTRODUCTION

The assessment of the health status of global forests is crucial for the evaluation of forest carbon sinks as climate change is posing new stress on forests [1]. The assessment of tree and forest health has traditionally been based on visual estimation that is prone to error and bias due to the subjective nature of the estimations [2]. Early detection of tree stress is needed for the mitigation of damages and adaption of new forest management strategies in the face of an altered forest operation environment [3]. Multispectral terrestrial lidar produces point clouds consisting of tens of millions of points and has shown promising results in the detection of declined tree health due to various environmental and pathogenic factors [4-6]. However, the studies conducted thus far have been done only in controlled environments and laboratories.

In this paper, we investigated the utilization of terrestrial lidars operating at two wavelengths in the assessment of varying tree health condition in a mature forest environment. The declined tree health was due to an on-going bark beetle infestation in the study area. This study is one the first steps towards an automated and objective tree health classification system that can be used for enhanced collection of reference data for producing satellite-based forest health map products.

2. MATERIAL AND METHODS

The study area is located in SE Finland, in the municipality of Ruokolahti. The test forest is comprised of mainly Norway spruce (*Picea abies*) with a mix of European rowan (*Sorbus aucuparia*), European aspen (*Populus tremula*) and Silver birch (*Betula pendula*). The forest has suffered from infestation by the European spruce bark beetle (*Ips typographus*) since 2011. During the summer of 2017, when the field data was collected, weather was cold and rainy, resulting in low damage due to the bark beetle colonization. Thus, the data is characterized by low and moderate bark beetle damage with many “green attack” trees, which did not exhibit visual symptoms, but the presence of bark beetles was confirmed.

A total of 27 trees were measured using two terrestrial lidar instruments: a FARO X330 operating at 1550 nm wavelength and a Trimble TX5 operating at 905 nm wavelength. The scanning was done consequently from a single location for each tree or plot with several trees. The tree condition was classified by experts in the field according to discoloration

and defoliation of the crown. Bark condition was assessed by classifying resin flows, bark beetle insertion holes and bark structural damage. Each of the five symptoms were classified using a three-class scheme (1-3; 1: no symptoms, 2: moderate symptoms, 3: severe symptoms). An attack score level was calculated by summing the damage values together, i.e. the higher the score level (value range 5-15), the more severe infestation symptoms were observed [7]. A bark damage score was calculated by summing the classes of the bark symptoms. The mean diameter at breast height of the trees was 29.0 cm and the mean height was 25.2 m.

The lidar point clouds at 905 nm and 1550 nm were registered using four external sphere targets. The distance effect on the recorded intensity [8] was corrected using a 10 degree polynomial model formed from an empirical dataset of measured intensities at 2-m intervals [9]. Resulted distance corrected intensity was calibrated with an exponential model based on measured intensities from reflectance panels, with nominal intensities of 5%, 10%, 20%, 40% and 60%, after Kaasalainen et al. [8]. Then, the intensity was normalized using an external reflectance panel with a nominal intensity of 20% to eliminate the effect of environmental factors on the measured intensity [10]. The result of the calibration process is referred as calibrated intensity within this paper.

Each tree was manually segmented from the point clouds resulting in tree point clouds consisting of about 250,000 points each. The points in each point cloud were then classified into stem and needle classes using a multi-dimensionality criterion [11] with the CANUPO software package within CloudCompare software [12]. The classification was based on a training sample of 50,000 manually delineated points. The calibrated and classified point clouds were merged by finding the nearest neighbor in terms of Euclidian distance, and a normalized difference index [4] was calculated for each point. The merged points were filtered to points with equal or less than 1 cm of Euclidian distance to the nearest point

A set of statistical metrics (Table 1) was calculated from each point cloud based on the calibrated intensity of 905 nm and 1550 nm wavelengths and the calculated NDI. In addition, the mean vertical angle in relation to the scanner position of the point clouds was calculated. The metrics were calculated for the classified needle points and stem points (only from 1.6-3.1 m height) separately. Multiple regression models were developed with attack score level as the independent variable and intensity metrics as the explanatory variables. Stepwise algorithm was used to select the variables. The regression models were evaluated using the Aikake Information Criterion, adjusted R^2 and root mean square error (RMSE). To avoid over-fitting and multicollinearity, models exhibiting variables with a variance inflation factor of greater than five were discarded and a model with less variables was chosen. Only models with predictors with a significance of less than $p<0.05$ were approved. All the statistical analysis was performed with the R software package [13].

Metric	Description
*i*_mean	Intensity average value
*i*_std	Standard deviation of intensity
*i*_p10, *i*_p20, … *i*_p90	Percentile 10, 20, 30, 40, 50, 60, 70, 80, and 90 of intensity distribution
*i*_max	Intensity maximum value
*i*_min	Intensity minimum value
*i*_kur	Kurtosis of intensity distribution [14]
*i*_ske	Skewness of intensity distribution [14]
*i*_entropy	Shannon diversity index (entropy) of the intensity distribution [15]
*i*_range	Difference between maximum and minimum intensity values
Angle	Mean vertical angle to the scanner

Table 1. The calculated metrics. N.B. i denotes wavelengths 905 nm, 1550 nm and the calculated NDI.

3. RESULTS

The developed regression models were able to explain 87% (Adj. R^2) of the variation in the attack score level (Table 2). The model was able to predict the attack score well throughout the range of attack scores, including the green-attack trees with no visual symptoms of infestation in the canopy, and a good separation between healthy and infested trees was found (Fig. 1). The predictors in the developed model were based equally on the needle and stem point classes.

Bark damage was predicted with a considerably lower accuracy, but a significant relationship between bark damage and the predictors was found (Adj. $R^2 = 0.47$). The bark damage model was not as successful in separating the healthy and infested trees as the attack level score model (Fig. 2). However, the model shows that the structural changes of the bark and resin flow on the stem affected the calibrated intensity.

Independent variable	Explanatory variables	Model	Adj. R^2	RMSE%
Attack level score	Canopy + stem	$Angle_{stem}^{*}$ + $1550\_ske_{needle}^{**}$ + $NDI\_kur_{needle}^{**}$ + $905\_range_{stem}^{*}$ + $905\_entropy_{needle}^{*}$ + $905\_mean_{stem}^{*}$ + $1550\_max_{needle}^{***}$	0.87	9.1
Bark damage	Stem	$905\_range_{stem}^{*}$ + $Angle_{stem}^{**}$ + $1550\_p90_{stem}^{****}$ + $905\_p40_{stem}^{*}$	0.47	21.2

Table 2. Summary of the regression models. The footnote indicates the source of the variable: needle or stem. The p-value is reported for each of the selected variables of the model. (*p-value = 0; **p-value<0.001; ***p-value<0.01; ****p-value<0.05).

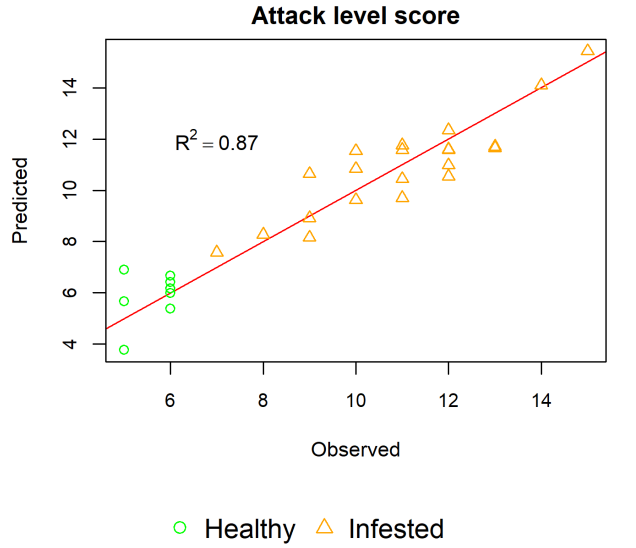


Fig. 1. The predicted vs. observed attack level score. Tree labeled healthy had scores of 5-6 and infested scores of greater than 6.

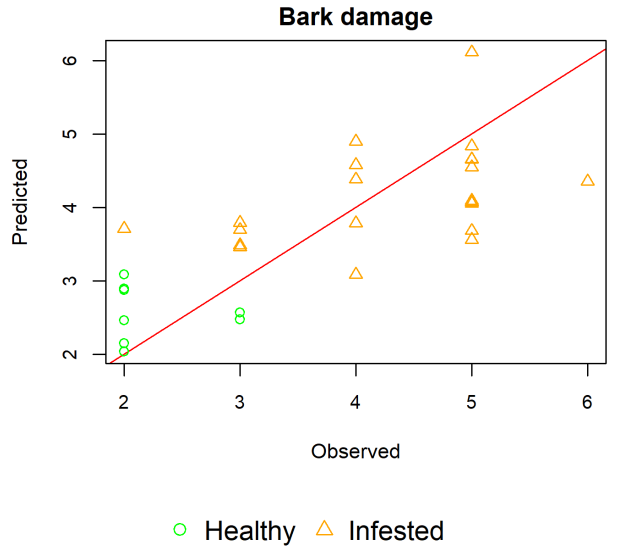


Fig. 2. The predicted vs. observed bark damage. Tree labeled healthy had scores of 5-6 and infested scores of greater than 6.

4. DISCUSSION AND CONCLUSIONS

The developed multiple regression models included a variety of different metrics. Both models had the angle metric of the stem. An incidence angle correction was not performed; thus, the effect of incidence angle was reduced by including the mean angle in the model. The model predicting the attack level score had more metrics that described the shape of the intensity distribution (e.g. kurtosis and skewness) rather than metrics that describe the actual values (e.g. mean or percentiles). This indicates that the tree condition could be at least partially modeled using metrics that are not directly influenced by the absolute values and less affected by the calibration procedure, i.e. how the shape of the measured intensity distribution compares to a normal distribution.

The developed regression models were able to predict the attack level score of each tree with high accuracy using only metrics derived from the intensity distribution of the used wavelengths 905 nm and 1550 nm and the calculated NDI. This study shows the great potential of lidar intensity in detecting varying tree health and tree condition already in the very early stages of tree decline. Since terrestrial lidar captures the 3D structure of trees with high accuracy, the simultaneous collection of tree structural attributes and tree health data is possible with the system described in this paper [16]. However, a study with a larger sample size should be conducted to verify the accuracy of the developed methods for larger areas. Also, other species and other causes of tree decline should be considered to evaluate whether a general model for detecting tree decline is feasible or if the models should be species- or disturbance type specific.

The developed method could be used to collect objective tree health data without the subjective bias imposed by visual inspection. This enables efficient collection of tree health data for the calibration and validation of satellite data products using satellites, such as the Sentinel-2, which has shown potential in detecting the early stages of bark beetle infestation [17]. Airborne multispectral lidar should be investigated for upscaling the terrestrial measurements to larger areas in future studies. The stem metrics were found important for the early detection of *I. typographus*

doi:10.2760/848593

colonization in this study, which is not visible from airborne platforms. On the other hand, the viewing geometry of the airborne lidar is more uniform regarding tree canopies compared to terrestrial lidar, which could enhance the detection of subtle differences.

First spaceborne lidars have been deployed recently. While first experiences of using them are gained, one could expect multispectral spaceborne lidars to emerge in the future. However, as spaceborne lidars do not cover areas as large as satellite imagery, global forest health map products will likely be based on multi- or hyperspectral imagery. The utilization of multispectral lidar from terrestrial to airborne in the development of national scale forest health products could significantly improve their accuracy.

5. REFERENCES

[1] C.D. Allen, A.K. Macalady, H. Chenchouni, D. Bachelet, N. McDowell, M. Vennetier, T. Kitzberger, A. Rigling, D.D. Breshears, E. Hogg, A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests, *For. Ecol. Manage.*, 259, 660-684, 2010.

[2] M. Vastaranta, T. Kantola, P. Lyytikäinen-Saarenmaa, M. Holopainen, V. Kankare, M.A. Wulder, J. Hyyppä, H. Hyyppä, Area-based mapping of defoliation of scots pine stands using airborne scanning LiDAR, *Remote Sens.*, 5, 1220-1234, 2013.

[3] L. Chaerle, D. Van Der Straeten, Imaging techniques and the early detection of plant stress, *Trends Plant Sci.*, 5, 495-501, 2000.

[4] S. Junttila, J. Sugano, M. Vastaranta, R. Linnakoski, H. Kaartinen, A. Kukko, M. Holopainen, H. Hyyppä, J. Hyyppä, Can Leaf Water Content Be Estimated Using Multispectral Terrestrial Laser Scanning? A Case Study With Norway Spruce Seedlings, *Frontiers in plant science*, 9, 299, 2018.

[5] S. Junttila, M. Vastaranta, X. Liang, H. Kaartinen, A. Kukko, S. Kaasalainen, M. Holopainen, H. Hyyppä, J. Hyyppä, Measuring Leaf Water Content with Dual-Wavelength Intensity Data from Terrestrial Laser Scanners, *Remote Sens.*, 9, 2016.

[6] S. Junttila, S. Kaasalainen, M. Vastaranta, T. Hakala, O. Nevalainen, M. Holopainen, Investigating Bi-Temporal Hyperspectral Lidar Measurements from Declined Trees—Experiences from Laboratory Test, *Remote Sens.*, 7, 13863-13877, 2015.

[7] M. Blomqvist, M. Kosunen, M. Starr, T. Kantola, M. Holopainen, P. Lyytikäinen-Saarenmaa, Modelling the predisposition of Norway spruce to Ips typographus L. infestation by means of environmental factors in southern Finland, *Eur. J. For. Res.*, 1-17, 2018.

[8] S. Kaasalainen, A. Jaakkola, M. Kaasalainen, A. Krooks, A. Kukko, Analysis of incidence angle and distance effects on terrestrial laser scanner intensity: Search for correction methods, *Remote Sens.*, 3, 2207-2221, 2011.

[9] K. Tan, X. Cheng, X. Ding, Q. Zhang, Intensity data correction for the distance effect in terrestrial laser scanners, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9, 304-312, 2016.

[10] A.F. Errington, B.L. Daku, Temperature compensation for radiometric correction of terrestrial LiDAR intensity data, *Remote Sens.*, 9, 356, 2017.

[11] N. Brodu, D. Lague, 3D terrestrial lidar data classification of complex natural scenes using a multi-scale dimensionality criterion: Applications in geomorphology, *ISPRS Journal of Photogrammetry and Remote Sensing*, 68, 121-134, 2012.

[12] D. Girardeau-Montaut, Cloudcompare-open source project, *OpenSource Project*, 2011.

[13] R.C. Team, A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[14] O.L. Davies, Statistical methods in research and production, *Statistical methods in research and production.*, 1947.

[15] C.E. Shannon, A mathematical theory of communication, *ACM SIGMOBILE mobile computing and communications review*, 5, 3-55, 2001.

[16] X. Liang, V. Kankare, J. Hyyppä, Y. Wang, A. Kukko, H. Haggrén, X. Yu, H. Kaartinen, A. Jaakkola, F. Guan, Terrestrial laser scanning in forest inventories, *ISPRS Journal of Photogrammetry and Remote Sensing*, 2016.

[17] H. Abdullah, A.K. Skidmore, R. Darvishzadeh, M. Heurich, Sentinel-2 accurately maps green-attack stage of European spruce bark beetle (Ips typographus, L.) compared with Landsat-8, *Remote sensing in ecology and conservation*, 2018.

BIG DATA APPLICATIONS FOR IMPROVED MIGRATION PROGNOSIS

Ipsit Dash, Victor Rijkaart, Gohar Sargsyan

CGI Nederland B.V., Rotterdam, The Netherlands

ABSTRACT

Migration is a complex and multi-faced phenomenon and has many stakeholders involved. Big data sources stemming from space assets and social media can be used to generate insights on both long and short term human migration. The paper describes different big data applications to provide early warning intelligence on migration flows by combining satellite data in the broadest sense with social media data mining. Coordinating the alignment of information from existing migration related data sources with innovative insights from big data sources such as space assets and social media can lead to better timeliness and accuracy for information provisioning for migration prognosis for different stakeholders.

Keywords— Migration, Big Data, Earth Observation(EO), Social Media Analytics, Service Design

1. INTRODUCTION

Human migration is a highly complex phenomenon that transcends cultural, social, economic, geopolitical boundaries. It continues to remain a topic of global relevance with its reflection in the 2030 agenda for Sustainable Development adopted by the United Nations (UN) and in the European Union (EU)[1]. Whilst migration in the broadest term is associated with improved opportunities for the states, business and communities as well as improving human lives in both origin and destination countries, recent years has seen a rather unprecedented growth in forced and irregular migration that are of a consequence of armed conflicts, environmental degradation and nullification of human rights. This calls for coordinated actions from the key global stakeholders with respect to migration prognosis which can be categorized into various stages of emergency preparedness [2] (mitigation, preparedness, response and recovery).

Combining insights from big data sources stemming from Space Assets and Social Media allows multi-faced information provisioning that complements the traditional prognosis methods [3]. Space assets (namely earth observation, GNSS and satellite communications) have global repetitive coverage that provides information from independent (space)-infrastructure. Social Media analytics and web search trends provides local insights and real time information about sentiments, thoughts, actions on certain demography, cluster of users and have potential of information gathering for migration prognosis [4, 5].

2. STAKEHOLDER NEEDS

The stakeholders in the migration prognosis ecosystem are illustrated in Figure 1.

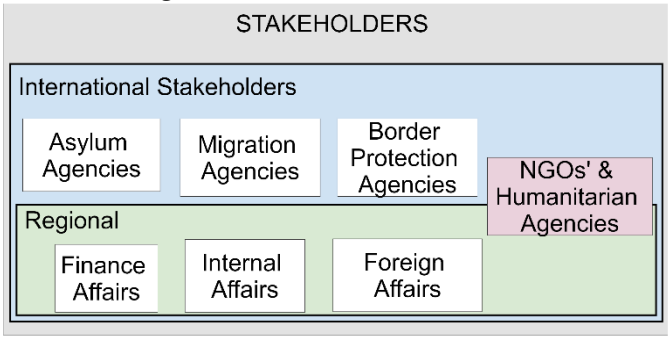


Figure 1 Stakeholders and users in migration prognosis

The following are common observations for information provisioning relevant to the stakeholders dealing with migration related prognosis:

1. There is a need to study and apprehend migration flows for both short-term oriented actions as well as long-term policy making.
2. Different stakeholders have different responsibilities and objectives and thus require different levels of information enrichment. A single product does not suffice the attested demands.
3. Although information about migration flows can be assimilated by a wealth of heterogeneous data sources and types, coordinated alignment of information provisioning to the right stakeholder at the right time is the need of the hour.

3. BIG DATA SOURCES AND APPLICATIONS

3.1. Space Assets

Geospatial datasets such as remote sensing imageries, location information, geodatabases can augment migration related datasets by linking them to a "geographical" attribute and providing contextual spatio-temporal information. For example, a promising application in the context of migration prognosis is "*Spatial-temporal change detection analysis*" [6] wherein, observable changes in geographical area and time periods can be analyzed to generate insights about

doi:10.2760/848593

probable causes of human migration (for instance, growth of temporary settlements, abrupt changes in urbanization). For instance in Figure **2**, growth of tents in Al-Zataari refugee camp from 2012 until 2015 can be discerned through direct interpretation of satellite imageries. Additionally, indirect analysis of information from space assets can be used to determine triggers for long term migration. For example analyzing socio-economic indices and demographic behavior, possibility of food shortage or crop failure. Lastly, information from space data can augment exiting migration related models and also perform their validation using historical datasets.

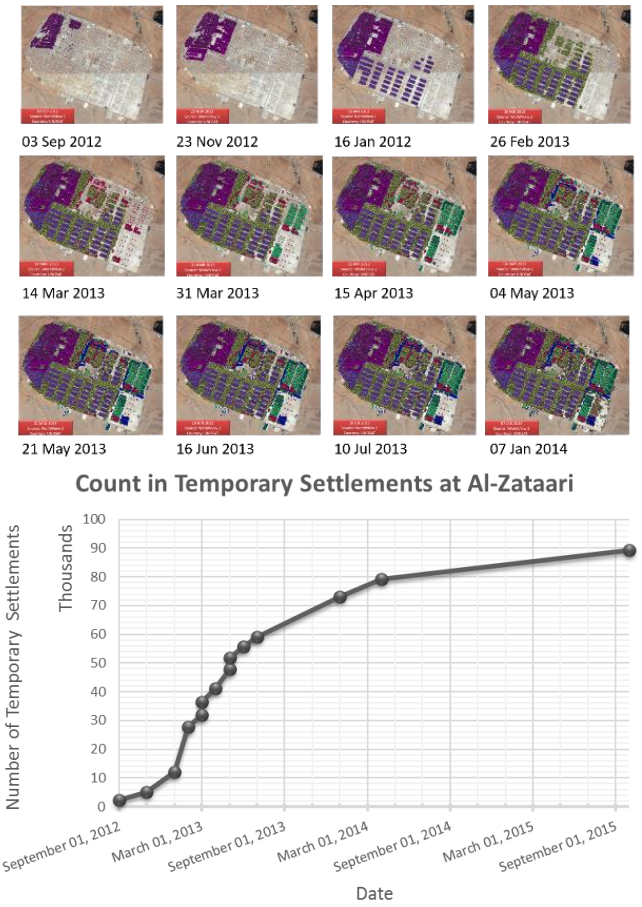


Figure 2 Growth of Al-Zataari camp from Sep 2012- Jan 2014 (source: UnoSAT)

3.2. Social Media Data

Avid usage of online and social media platforms in recent years has offered a wealth of information to better understand local perspectives and provide real time local insights. These insights are of significant value to decision makers involved in the migration prognosis. Such as understanding topic trends, sentiments for a specific population for a geographic area and time. In Figure **3**, sample analysis on tweets from Twitter Public API on topic "Netherlands" for 2 weeks of 2017 are shown. The user group belong to southern Europe. By employing methods such as text based or user centered clustering, relevant events, intentions and sentiments can be discerned which can provide information about migration trends.

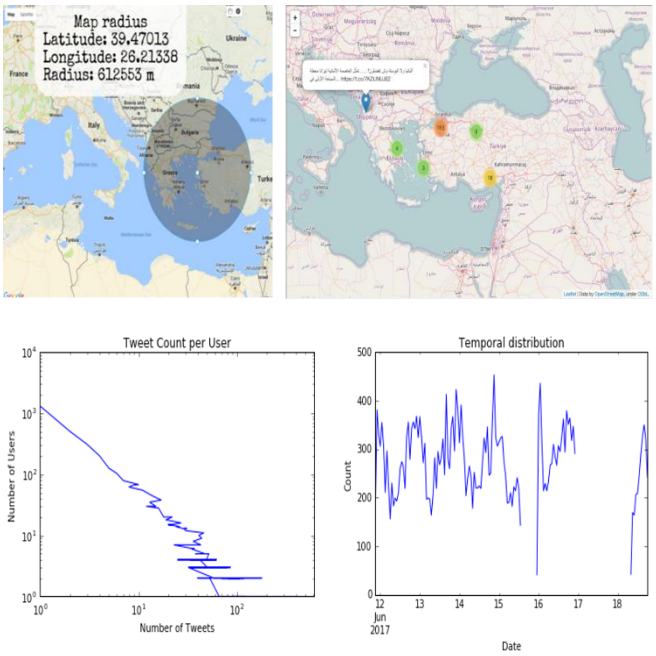


Figure 3 Sample Analysis on Tweets from southern Europe (source: Twitter)

3.3. Services and Applications

Potential big data application services to help migration prognosis are designed by encompassing a theme revolving around the usage and analysis of big data sources such as space assets (EO, Satellite Navigation) and social media data and how insights delivered from these data can help address potential stakeholders' challenges and concerns. Each service description is further subdivided into different applications and the way it is beneficial to the stakeholders. Three different services are identified, namely, Social Media Analytics, Earth Observation Analytics (Table **1**).

Table 1 Big data Applications for Migration

Service Description	**Application Description**	**Usage**
Social Media Analytics	Sentiment Analysis and Multi-lingual Topic Analysis	This gives insights about initiators for migration. The type of initiator, the situation of the sender with respect to his or her social group, possibly desired future status(es) of the sender, or even a timeline of events
	Web Search Analytics	This gives amongst others, insight in planning around

doi:10.2760/848593

		migration included analyses of the searches on possible options to new residencies, as well as routes.
Earth Observation Analytics	Image Classification and Analysis	Observing initiation, growth, demise of settlements through automated image classification combined with intelligence about the type of settlement gives insights in in- or outflow of people in places.
	EO Change Detection Analysis	Combining imagery (growth or demise a new settlement) in an augmented way (do we detect tracks / trucks / left garbage nearby) for different settlements gives insights in routings and flows with respect to migration.
	Socio-economic Forecast Analysis	Space data solutions towards insights in migration routings and initiators can be further enhanced with information from space on food-security. This enables the stakeholders to mitigation the effects of food-security towards migration

4. SERVICE DESIGN AND INTEGRATION

Our service design focused on catering diverse stakeholders with different information enrichment levels through different applications. Thus the architecture allowed customized instances of the solution relevant per user. Additionally, scaling up and ability to be incorporated in existing stakeholder applications were also taken into consideration.

4.1. High level architecture

The architecture consists of three core components, namely Man Machine Interface (MMI), Business Service Interface (BSI) and Space Core Platform (SCP). Datasets from big data sources (Social Media and Space Assets) is be used by the Space Core Platform. End users communicate with the solution using an API or Dashboard. The dashboards are implemented in the MMI. The API is implemented in the Business Service Interface.

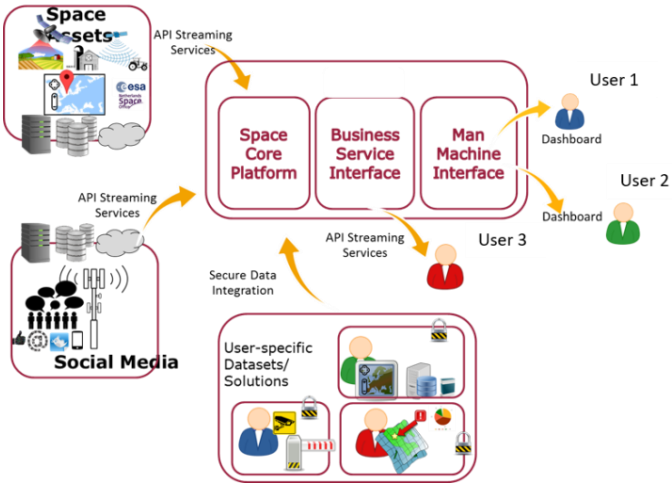


Figure 4 High level Architecture

4.2. Three layered approach

The proposed solution is an online web platform that provides three layers of information products to the end user. These are:

1. Information derived from assembling and cataloguing existing data sources for migration.
2. Providing results from applications evolved from the wealth of space assets and social media data mining as additional layers to existing information.
3. Provisioning of validation of information, which allows verifying results of social media analytics by using results from space imageries or vice versa.

5. EXPERIMENTAL RESULTS

5.1. Test case on Heumensoord, Netherlands

For experimentation purposes, we chose a retrospective case based in Heumensoord in the Netherlands. It is located in south of the Netherlands was the home of refugee and asylum seekers in the Netherlands those affected by the migration crisis that hit Europe in 2015.

5.1.1. Earth Observation Analysis

We performed change detection for high resolution SPOT 6 imageries for the months of August- December 2015 using the following equation.

$$Change\ Detection_{i,j} = {(Band_{3_i}}/{Band_{4_i})} - {(Band_{3_j}}/{Band_{4_j})}$$

Where, i, j corresponds to different months. SPOT 6 Band 3 corresponds to 0.62-0.69 µm and SPOT 6 Band 4 corresponds to 0.76-0.90 µm.

The standard ration ${(Band_{3_i}}/{Band_{4_i})}$ allow the definition of urban areas and human conglomeration uniquely. The histogram of change detection images of different months

doi:10.2760/848593

are shown in Figure **5**. Pixels whose value is less than 0 depict a decrease whereas those greater than 0 depict an increase of urban activities and human conglomeration.

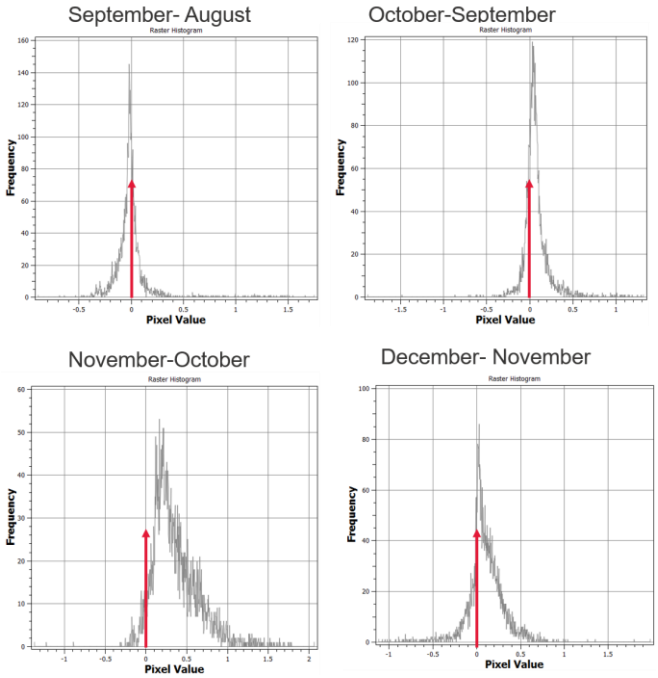


Figure 5 Change detection histograms of image analysis of Heumensoord

5.1.2. Social Media Analytics

Additionally, sentiment analysis was also performed on Twitter and Facebook posts for Heumensoord for same time period (Aug- Dec 2015) involving keywords such as "heumensoord" "migrant" or "migrat" or "vluchte" or "asielzoek". Figure **6** consists of 2 graphs, the upper depicts the volume of social media activity such as posts, mentions of the keywords and the lower graph depicts the sentiments of those posts.

Figure 6 Social media sentiment analysis on Heumnesoord

5.1.3. Inferences

Figure 5 shows increased urban activities in Heumensoord from August- December 2015 which can be seen from the gradual shift of the histogram's median against reference 0 which counts for no change. This also corresponds to increased social media posts and sentiments during the same time period as seen from Figure 6. Thus in a broader sense, trends of peaks and troughs of social media activity can be correlated with those analyzed from satellite imageries.

5.1.4. Limitations

Although the study was carried out using satellite data of monthly frequency, a higher temporal resolution would have led to a better change detection characterization. Furthermore, overview of the characteristics of the social media users would lead to better conclusions on sentiment perceptions.

6. CONCLUSIONS AND FUTURE WORKS

Owing to huge potential of described big data applications towards improving the accuracy and timeliness of migration prognosis, providing early warning intelligence is possible. In the future steps, dissemination of the services to the stakeholders is foreseen. New aspects of insight generation, addition of other big data sources are also envisioned.

7. ACKNOWLEDGMENTS

The authors would like to thank European Space Agency (ARTES Program) for funding the feasibility study [3] which formed the basis of the paper. Special acknowledgments to Statistics Netherlands (CBS) for being part of team for carrying out of the feasibility study.

8. REFERENCES

[1] UNITED NATIONS PUBLICATIONS. (2018). WORLD MIGRATION REPORT 2018. [S.l.]: UNITED NATIONS PUBNS.

[2] Fema.gov. (2018). Mission Areas | FEMA.gov. [online] Available at: https://www.fema.gov/mission-areas [Accessed 14 Oct. 2018].

[3] Business.esa.int. (2018). Migration Radar 2.0 - Big data applications to boost preparedness and response to migration – Feasibility Study | ESA Business Applications. [online] Available at: https://business.esa.int/projects/migration-radar-20 [Accessed 14 Oct. 2018].

[4] Spyratos, Spyridon & Vespe, Michele & Natale, F & Ingmar, W & Zagheni, E & Rango, M. (2018). Migration Data using Social Media: a European Perspective, Publications Office of the European Union. 10.2760/964282.

[5] Rango, Marzia & Vespe, Michele. (2017). Big Data and alternative data sources on migration: from case-studies to policy support - Summary report.

[6] Singh, A. (1989). Review article digital change detection techniques using remotely-sensed data. International journal of remote sensing, 10(6), 989-1003.

EOPEN: OPEN INTEROPERABLE PLATFORM FOR UNIFIED ACCESS AND ANALYSIS OF EARTH OBSERVATION DATA

*Guido Vingione[1], Gabriella Scarpino[1], Laurence Marzell[1], Tudor Pettengell[1], Ilias Gialampoukidis[2], Stelios Andreadis[2], Stefanos Vrochidis[2], Ioannis Kompatsiaris[2], Bernard Valentin[3], Leslie Gale[3], Woo-Kyun Lee[4], Wona Lee[4], Michael Gienger[5], Dennis Hoppe[5], Vasileios Sitokonstantinou[6], Ioannis Papoutsis[6], Charalampos Kontoes[6], Francesco Baruffi[7], Michele Ferri[7], Hoonjoo Yoon[8], Ari Karppinen[9], Ari-Matti Harri[9] **

[1]Serco S.p.A., [2]Information Technologies Institute, Centre for Research and Technology Hellas, [3]Space Applications Services NV/SA, [4]Korea University Environmental GIS/RS Centre, [5]University of Stuttgart – High Performance Computing Center Stuttgart, [6]National Observatory of Athens, [7]Autorità di bacino distrettuale delle Alpi orientali, [8]Sundosoft Limited, [9]Finnish Meteorological Institute*

ABSTRACT

EOPEN (https://eopen-project.eu/) is a project which has received funding from the European Union's Horizon 2020 research and innovation programme under the topic EO Big Data Shift in 2017 and has a duration of 3 years, starting from November 2017. In this work, we present the concept of the project, its objectives and the lessons learnt after almost one year of project lifetime, as a follow-up to our previous project presentation at ESA BiDS'17 in Toulouse.

Index Terms— Earth Observation, Copernicus, data fusion, interoperability, decision making, visual analytics.

1. INTRODUCTION

Earth Observation data access through the Copernicus data distributor systems has paved the way to monitor changes on Earth, using Sentinel data. One of the main objectives of EOPEN [1] is to fuse Sentinel data with multiple, heterogeneous and big data sources, to improve the monitoring capabilities of the future EO downstream sector. Additionally, the involvement of mature ICT solutions in the Earth Observation sector shall address major challenges in effectively handling and disseminating Copernicus-related information to the wider user community, beyond the EU borders. Relevant projects include the openEO [2], BETTER [3], PerceptiveSentinel [4] and CANDELA [5].

To achieve the aforementioned goals, EOPEN fuses Copernicus big data content with other observations from non-EO data, such as weather, environmental and social media information, aiming at interactive, real-time and user-friendly visualisations and decisions from early warning notifications. The fusion is also done at the semantic level, to provide reasoning mechanisms and interoperable solutions, through the semantic linking of information. Processing of large streams of data is based on open-source and scalable algorithms in change detection, event detection, data clustering, which are built on High Performance Computing infrastructures. Alongside this enhanced data fusion, a new innovative, overarching Joint Decision & Information Governance architecture is combined with the technical solution to assist decision making and visual analytics in EOPEN. EOPEN will be demonstrated in real use case scenarios in flood risk monitoring, food security and climate change monitoring, as also shown in Figure 1.

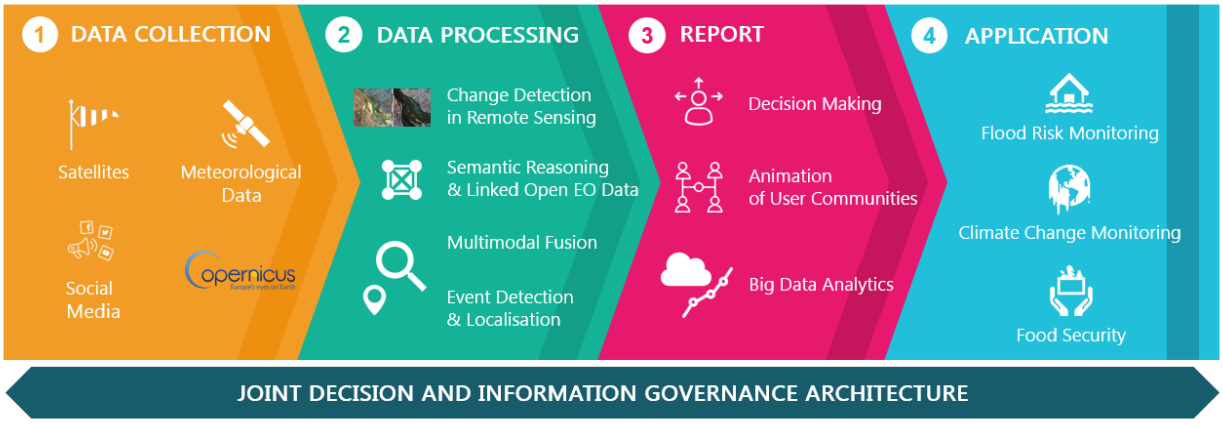


Figure 1: The EOPEN concept

* This work is supported by the EOPEN project, funded by the European Commission, under the grant agreement H2020-776019.

2. APPROACH

The overall objective of EOPEN is to provide a platform targeting non-expert EO data users (non-traditional user communities), experts and SME community that reveals and makes Copernicus data and services easy to use for Big Data applications by providing EO data analytics services, decision making and infrastructure to support the Big Data processing life-cycle allowing the chaining of value adding activities across multiple platforms.

To successfully address the Big Data challenges and to benefit from the services provided by ICT companies for accessing and processing Copernicus data, we are developing the EOPEN platform which delivers Copernicus data to non-traditional user communities, applying data compression and storage of EO and non-EO data (i.e. meteorological data, social media, linked open data), using cloud infrastructure and high performance computing (HPC), in order to fuse data from diverse sources and from different modalities (e.g. visual, textual or spatiotemporal).

Indexing Sentinel high resolution (HR) images will be performed to ensure fast access to their related content and assists pattern recognition and machine learning techniques by boosting their performance when they rely on solid multimedia indexing techniques. Data management techniques, community detection and tracking on the network for visualisation of usage activities and network analytics to reveal the key players (public or private users) and groups of users (communities) in the EO domain will be performed. All these heterogeneous sources of information are combined, through multimodal fusion, to semantically interpret the content of EO data resulting in efficient decision-making and visualisation in line with the proposed Joint Decision and Information Governance Architecture. The overall approach of EOPEN, aligned with the operational timeline is shown in Figure 2.

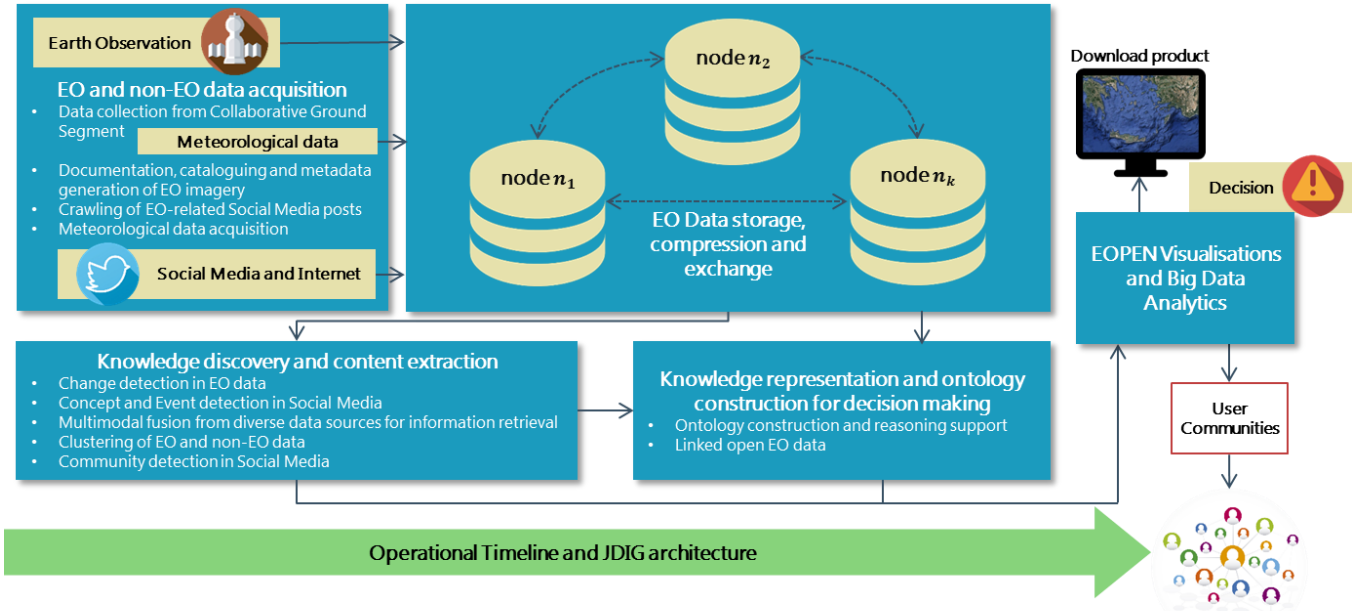


Figure 2: The EOPEN approach

3. APPLICATION

Three use case scenarios are foreseen in EOPEN. In the following, we present the challenge that EOPEN deals with, in each use case considered.

3.1. Flood risk assessment and prevention

The pilot area, within the Italian Eastern Alps river District, comprises all the municipalities of the Local Risk District of Vicenza in Italy. This area is regularly affected by critical flooding from the Bacchiglione River and its tributaries. Planned flood defences remain largely unfinished, and a high risk of flooding therefore persists. Flood in the cities led to high levels of water in the streets, causing many problems such as the drowning of people, building damage and traffic problems. As indicated in the Flood Directive (2007/60/CE) water authorities should plan measures in order to aim at reducing risks by minimizing the possible damages effects and losses that may result.

In the Local Risk District of Vicenza AAWA provide flood forecasts warnings by running its Flood Forecasting System (AMICO) based on traditional meteorological data. The emergency phase is coordinated by the Mayors of the cities with a slow response time due to the lack of a data processing structure able to monitor in real time the evolution of the flooding in the territory (in term of flooded areas evolutions, impacts, damages). Currently these information come to decision-making authority through radio communications by people (civil protection volunteers) distributed in the territory. There is a need to

doi:10.2760/848593

provide faster and more effective emergency responses to extreme weather by increasing the speed of risk analysis.

3.2. Food Security through Earth Observation

"Food Security" is a denomination introduced by the Food and Agriculture Organization (FAO) of the United Nations. The problem is really complex and comprises several different components (food access, distribution, food supply stability, use of food). There are many recent examples that show the problem and more precisely food crises, for instance famine in the horn of Africa (2011) and the critical need to deliver timely food security information to decision-makers. Satellite data have been used to detect and monitor severe agricultural events since 1972 on the occasion of the extreme drought that took place in Russia. Improvements have been made in the spectral, spatial and temporal resolutions of Earth observation (EO) systems since then. Copernicus program and Sentinels' missions are the most ambitious Earth observation initiative and have a great impact and contribution also in the field of food security.

3.3. Monitoring Climate Change through Earth Observation

The climate can be defined as average weather conditions in terms of the mean and variance of temperature, precipitation and wind over a period of time. In climate studies, the averages of these parameters are normally calculated through an averaging filter spanning 30 years. Currently, it is scientifically clear that the climate is changing and the temperature is rising. Precipitation patterns and the frequency of occurrence of storms are changing as well.

The climate change manifests itself most visibly and rapidly at high latitudes. Hence a regional pilot area is established around Finnish borders. The most prominent and clear indicators of climate change are the atmospheric temperature and the average annual snow cover, which is consistent with warmer global temperatures. These parameters have been monitored by the Finnish Meteorological Institute (FMI) for decades, even for more than 100 years. Hence we use the ground-based climate observations of the Finnish Meteorological Institute dating back for more than 100 years (FMI Climate Services Archives) and satellite observations, e.g., by EUMETSAT, ESA, NASA and NOAA. A major project asset is the access to all FMI data which have been made avilable to support open access policy.

4. ILLUSTRATIONS

In this section we present some indicative applications that EOPEN offers to its end user community. EOPEN collects EO data and combines them with weather forecasts and Twitter posts. A social media image is automatically annotated to extract knowledge (concepts), using a technique based on Deep Convolutional Neural Networks (DCNNs), as shown in Figure 3.

Figure 3: Concepts extracted from a social media image

Moreover, the location of a tweet is rarely included (less than 5%) in the stream of posts, hence EOPEN considers automatic estimation of an event location from Twitter content and positioning on a map through an external ontology and linking of data. This functionality supports the situational awareness of an authority that would like to fuse citizen observations with EO products (Figure 4).

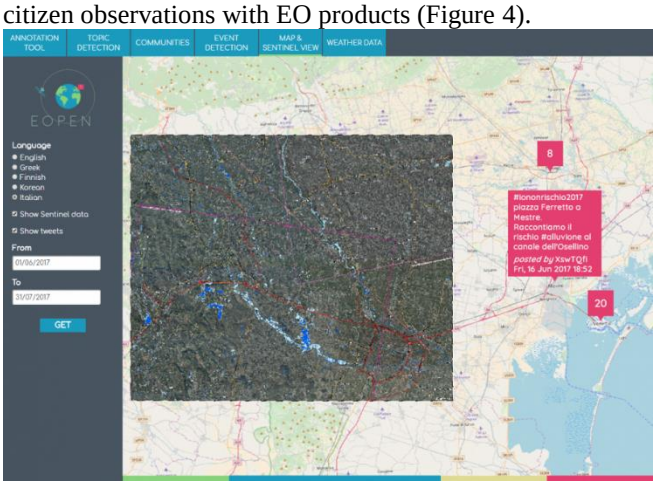


Figure 4: Localised events from social data add value to EO data products, which can both appear in a GIS view.

Satellite images often include a road network or a part of it. In case of an extreme weather event, such as a severe flood event, passing from one part of the road to another is not possible. EOPEN offers the possibility to estimate passable or not passable parts of a road network, with high accuracy, using a Residual Neural Network (ResNet), where some training took place on the annotated dataset of the Multimedia Satellite task of MediaEval2018. The analysed images keep their original georeferenced information, so as to be visualised as a GIS layer.

Figure 5: Various inputs to the road passability service

doi:10.2760/848593

The plan is to employ a Region proposal Neural Network to first detect the road parts (Figure 5) and then apply a binary classification algorithm to infer whether the road is passable or not.

Finally, EOPEN clusters similar content into groups of similar items. Similarity can be defined in many different ways (cluster by concept, location, event, user, etc.) even if refers to EO or non-EO data streams. The management of large and highly heterogeneous content requires scalable techniques that may take advantage of the existing European HPC resources, being also in line with the recent advances in the development of the four Data and Information Access Services (DIAS) and among which the ONDA DIAS service (https://www.onda-dias.eu/cms/) will be used for supporting the project validation. EOPEN develops algorithms using parallel programming techniques and libraries to boost scalability of the user applications and being executed on a HPC infrastructure.

5. IMPACT

The societal, technical and scientific and economic impact of EOPEN is eminent and briefly described as follows.

5.1. Societal Impact

EOPEN provides a means to perform analysis that is not yet available. Three use cases demonstrate how EOPEN can be used to address societal challenges fully addressing EO-2-2017 call's requirement optimising the use of Copernicus data by non-traditional user communities to meet societal challenges. In particular, activities under the societal challenge for climate action, environment, resource efficiency and raw materials focus on GEOSS, much like EOPEN's use cases and services that stimulate past ICT and EO activities coupled with the new Sentinel data and knowledge from various sectors.

5.2. Technical and Scientific Impact

In comparison with the data distribution system of the US, the EU data and products distribution and sharing infrastructure system looks less effective, especially in terms of facilitating small companies to play with data and create marketable services and products out of it. Copernicus lacks a holistic approach to data management because data distribution is based on fragmented data sources. EU Copernicus data and products are spread in many different archives, formats and portals. The Copernicus core service segment (with 6 thematic areas and relative product portfolios), the EUMETSAT Satellite Application Facilities network (with 8 Application facilities and product catalogues) and the ESA Sentinels Open Access Hub portal represent just three examples of different, and sometimes overlapping, large product repositories, hosting huge amounts of information that are difficult to explore and navigate into. The lack of an adequate solution for combining data from multiple sites with non-space derived data has given rise to Thematic Exploitation Platforms and networks of them. The European Space Agency (ESA), on behalf of the European Commission, launched the DIAS services. The DIAS provides a scalable computing and storage environment for third parties. With references to the afore-mentioned architectures, the EOPEN combines state-of-the-art technical solutions from the EO domain with mature ICT technologies, to deliver an efficient orchestration of services and modules, infrastructure agnostic, which are offered to the end user, without his/her need to have experience in downloading and processing Copernicus EO products. In EOPEN, data harmonisation and standardisation have highest priority in order to foster all three use cases covering flood risk management, food risks from environmental factors and climate changes, serving as a worldwide solution inside and outside the EU.

6. CONCLUSION

In this work we briefly present the EOPEN concept and approach, both at the platform level and use case application. The purpose of this work is to demonstrate the current status of the development of EOPEN and to set a solid basis for the next two years of implementation, verification, validation, evaluation and demonstration.

REFERENCES

[1] EOPEN project: https://eopen-project.eu/
[2] openEO project: http://openeo.org/about/
[3] BETTER project: https://www.ec-better.eu/
[4] PerceptiveSentinel: http://www.perceptivesentinel.eu/
[5] CANDELA project: http://www.candela-h2020.eu/
[6] I. Gialampoukidis, A. Moumtzidou, M. G. Scarpino, G. Palumbo, S. Vrochidis, I. Kompatsiaris, F. Zaffanella, D. Norbiato, M. Ferri, & G.Vingione, "Earth Observation and Social Multimedia Data Fusion for Natural Hazards and Water Management: The H2020 EOPEN Project Paradigm", 2nd International Conference Citizen Observatories for natural hazards and Water Management, Venice, 27-30 November 2018
[7] G. Vingione, L. Marzell, E. Cadau., I. Gialampoukidis, S. Vrochidis, I. Kompatsiaris, B. Valentin, M. Melcott, L. Gale, W.-K. Lee, S. Woo, M. Gienger, D. Hoppe, C. Kontoes, I. Papoutsis, M. Ferri, F. Baruffi, J. Yoon, H. Yoon, A. Karppinen, A.-M. Harri, "THE H2020-EO EOPEN PROJECT", INSPIRE Conference 2018, 18-21 September 2018, Antwerp, Belgium
[8] G. Vingione, L. Marzell, E. Cadau., I. Gialampoukidis, S. Vrochidis, I. Kompatsiaris, B. Valentin, M. Melcott, L. Gale, W.-K. Lee, S. Woo, M. Gienger, D. Hoppe, C. Kontoes, I. Papoutsis, M. Ferri, F. Baruffi, J. Yoon, H. Yoon, A. Karppinen, A.-M. Harri. "The H2020-EO EOPEN project". ESA Conference on Big Data from Space 2017, 28-30 November 2017, Toulouse, France.

doi:10.2760/848593

OPENEO – A STANDARDISED CONNECTION TO AND BETWEEN EARTH OBSERVATION SERVICE PROVIDERS

Matthias Schramm[1], Edzer Pebesma[2], Wolfgang Wagner[1], Jan Verbesselt[3], Jeroen Dries[4], Christian Briese[5], Alexander Jacob[6], Matthias Mohr[2], Markus Neteler[7], Thomas Mistelbauer[5], Tomasz Miksa[1], Sören Gebbert[2,4], Bernhard Gößwein[1], Miha Kadunc[8], Pieter Kempeneers[9], Noel Gorelick[10]

[1]Vienna University of Technology (TU Wien), Vienna, Austria; [2]University of Münster, Münster, Germany; [3]Wageningen University and Research, Wageningen, The Netherlands; [4]Vlaamse Interstelling Voor Technologisch Onderzoek N.V., Boeretang, Belgium; [5]Earth Observation Data Centre for Water Resources Monitoring GmbH, Vienna, Austria; [6]Eurac Research, Bozen, Italy; [7]Mundialis GmbH & Co. KG, Bonn, Germany; [8]Sinergise Laboratorij Za Geografske Informacijske Sisteme Doo, Ljubljana, Slovenia; [9]Joint Research Centre, Ispra, Italy; [10]Google Switzerland GmbH, Zurich, Switzerland

ABSTRACT

Developments in Big Data technology during the last decade led to the parallel rise of several independent cloud service providers for Earth Observation (EO) data. The resulting variety of customized solutions of the back-end providers forces users to choose between very different, incompatible interfaces.

openEO offers an alternative, presenting an API which connects to EO service providers and provides standardised access points to users via programming languages as Python, R, and JavaScript. In this ongoing project process catalogues are being built up, covering all aspects of the EO data life cycle and serving as a template for interested service providers and front-end users to connect to the API. Several EO service providers can already be accessed via the API. Use cases, based on openEO will be developed for pilot users to proof its advantages and usability.

Index Terms— Earth Observation, Cloud service providers, Standardisation, Process catalogue

1. INTRODUCTION

Over the last decade important innovations in EO data utilization became possible due to novel sensors which combine a fine geometric with a high temporal resolution. The availability of an unprecedented variety and vastly increased volume of EO data led to a paradigm shift in data managing and processing procedures [1] towards cloud computing approaches that bring the users and their software to the data. New Big Data technologies, capable of dealing with the increased volume, variety, velocity and veracity of the EO data [2], are emerging, paving the way for new market players, service offerings and new user groups. For example, users may access raw and value-added Copernicus and other EO data by connecting to Infrastructure as a Service (IaaS) providers as the ESA's Copernicus Open Access Hub [3]. On the other side, advanced Platform as a Service (PaaS) solutions provide on-demand EO data processing through platform-specific API calls. These new platforms allow the user community to process Big Data decentralized at EO service providers, saving time and reducing internet traffic.

Being a rather recent development, EO platform service providers still offer solutions, customised to their user communities; speed and momentum of this new development prevented service providers from developing generally accepted standards. As a consequence, different data formats, process catalogues and available processing capabilities lead to an insufficient interoperability between the EO platform [4]. A switch between available service providers therefore always also means that users must re-develop already existing process chains.

openEO is an ongoing, user driven API development project to form a standardised access point to various EO service providers, allowing users EO data processing at different cloud platforms via Python, R, or JavaScript [5, 6]. Being language neutral, the openEO API can further be enhanced by implementing APIs for e.g. other programming languages and applications including Quantum GIS or GRASS GIS development, which is under way. The openEO API is currently capable to connect to EO service providers from the EODC / Austria, VITO / Belgium, Eurac Research / Italy, Mundialis / Germany, Sinergise / Slovenia, JRC / Italy, and Google Earth Engine / Switzerland. Also the connection to DIAS platforms is planned. Representing diverse infrastructures, these cloud providers are serving as templates for other back-ends to also connect to openEO.

Utilizing openEO, EO data platforms are able to provide services to the user, reflecting examples from all stages of EO

doi:10.2760/848593

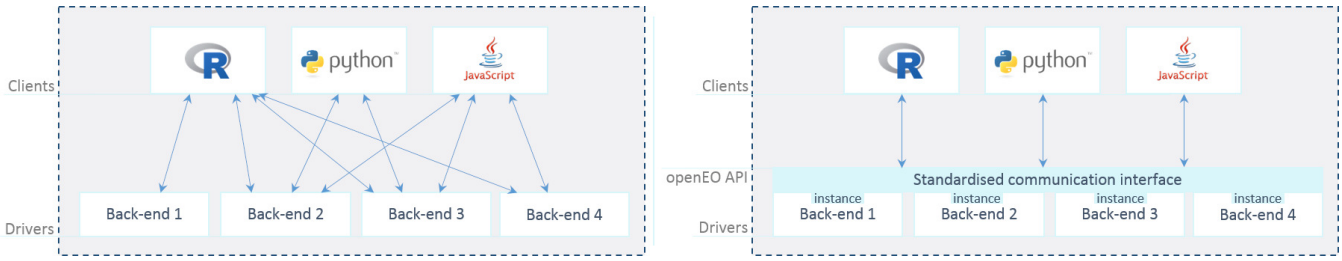


Fig. 1. Communication between clients and back-end drivers. Left: common many–to–many communication; Right: many–to–one communication, using the openEO API.

data processing, such as EO data query, image enhancement, band math operations, image mosaicking and layer stacking, or visualisation and download. Researchers will benefit by having a uniform way of connecting to different back-ends that allow them easily switch between compatible environments with little or no effort. Back-end operators do not need to modify the way back-end work, but need to provide additional interface that translates openEO requests to the software running in the back-end.

Compared to OGC WCPS, which currently sees rather limited uptake in the Earth Observation community, openEO is thought of as more user-oriented, more flexible in functionality, and less restrictive to back-end data representation: it allows for instance image collections that have not been mosaicked to a coverage. WCPS is explored as one of the possible back-ends.

Being a user-oriented open source project, the openEO consortium aims for an extended involvement of the user community while defining and developing relevant EO data processes and user-defined functions (UDFs). Connecting them to workflows for five already defined use cases will help ensuring and validating openEO's usability.

2. OPENEO STRUCTURE

openEO represents a set of standardised contracts between various clients and one well–defined API. Instances of the API can be installed at specific back-end drivers, which are deployed by EO service providers to implement these contracts and thus to establish openEO compatible access points. Contracts, formerly implemented separately for each client and service providers, following a many–to–many communication strategy can now be realised as many–to–one communication (see Fig. 1).

Client APIs are realised by software libraries specific to a given programming language, e.g. Python modules or R packages. This gives the user a way to interact with the openEO API without having to take care of the complexity of building back-end specific HTTP requests. By implementing an openEO instance as access point to a back-end provider, it will be able to process user commands at the back-ends infrastructures and return its results. The process graphs, transferred via JSON requests can be executed at the back-ends in three different ways.

1. A batch job can be submitted, which stays inactive until processing is requested. It will run only once and stores its results after execution.

2. Secondary web services allow web-based access using different protocols such as OGC WMS, OGC WCS or XYZ tiles. The computation runs on demand to allow users to change e.g. the result's viewing extent or level of detail.

3. Lightweight process graphs (e.g. small previews) can be executed synchronously. More costly processes have to expect timeouts for long-polling HTTP requests.

To enable the use of higher level EO process graphs, user-defined functions (UDFs) can be executed within openEO. Meeting extremely specialised demands, their implementation is an ongoing process that will need further input from potential users. While the standardisation of the available processes within openEO is still ongoing, its overall architecture was already specified / developed preliminarily, containing following aspects.

- Query of back-end capabilities regarding e.g. authentication method or UDF compatibility,
- Query of available EO data and processes, depending on metadata,
- User management: e.g. token based user registration / authentication, retrieval of user credits, billing,
- Synchronous / asynchronous job management,
- Data download in different output formats

3. API FUNCTIONALITY

A process catalogue was developed, describing a set of functionalities to be implemented within the project, their I/O data

doi:10.2760/848593

and their exact workflow. Its user-driven development will enable openEO to form widely accepted and used standards, and to build a consistent syntax. Furthermore, the back-ends will be able to translate the commands, standardised to the openEO's syntax to well defined internal processes, guaranteeing a compatibility between the EO service providers. 3rd-party processing platforms are able to use the process catalogue as a guideline to access to openEO.

Core processes of following topics are momentarily defined within an increasing list to be implemented in a standardised way within openEO.

- User data management: Authentication, EO data upload / storage / download / sharing,
- EO meta-database query,
- EO data masking, filtering by logical expressions / metadata / geometry / time range
- Image enhancement: zonal / pixel based math operators, re-projection, contrast enhancement stretching, kernels, predefined math operators (e.g. vegetation indices), geometric and temporal rescaling, spatial / temporal resampling,
- Math operations: binary arithmetic, statistical operations, Boolean operations, zonal statistics, regression,
- Subsetting, mosaicking, layer stacking, expanding / reducing dimensions,
- Sorting and ordering algorithms, searching for specific elements
- Visualizing / saving / downloading EO data,
- Real time interaction: job management, process monitoring

This list, containing all needed processes to realise 5 well–defined use cases (see Section 4), is meant to serve as a template for future enhancement.

Additional to the possible use of the openEO API via R, Python or JavaScript Syntax, a visual process graph builder serves as a GUI for merging single openEO processes to workflows (see Fig. 2).

4. USE CASES

The defined core processes will allow to process first use cases with openEO. The following use cases were chosen, based on existing demands from the user community, to provide processes for broad topics and to serve as seeding points for user discussions and for future development.

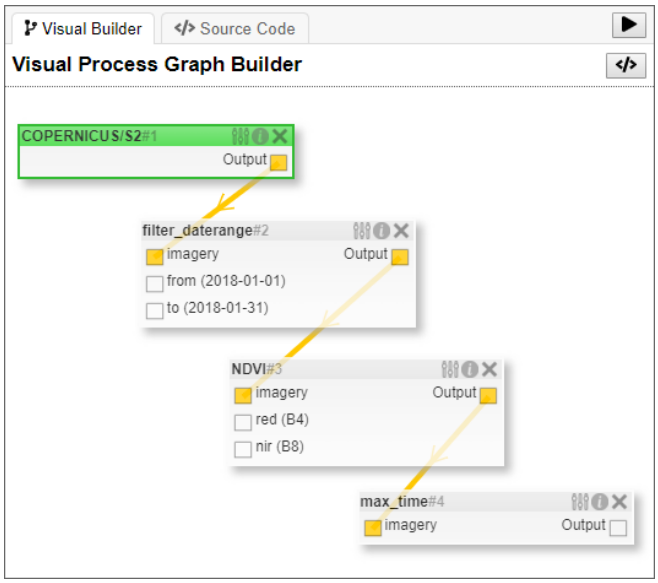


Fig. 2. Example of a visual process graph: NDVI calculation from a Sentinel-2 time series, using openEO.

4.1. Radar image compositing

The process chain produces monthly and seasonal RGB composites of Sentinel-1 backscatter [7]. The composites can be used for further classification and crop monitoring [8, 9] and will be a test case for the basic openEO's functionalities like querying and transforming data, basic statistics, layer stacking and exporting to specific output formats. This work flow will be tested by the Austrian *Federal Ministry Sustainability and Tourism* (BMNT) for Austrian pilot areas.

4.2. Multi-source phenology metrics and data fusion

Already existing data fusion and phenology metrics tools [10] will be ported to openEO to combine Sentinel-2 time series with Sentinel-3 and Proba-V data. Image pre-processing tools will be implemented as well as a product validation with in-situ and other ancillary datasets. These work flows will be used in semi-arid regions in Western Africa for the *Action against Hunger* (ACF) and in the Hindu Kush Region for the *International Centre for Integrated Mountain Development* (ICIMOD).

4.3. Optical-Radar forest monitoring

This use case focuses on the combination of Sentinel-1 and Sentinel-2 time series for a near real-time forest and deforestation monitoring. It provides openEO with various basic statistical and time series algorithms. The work flow will be tested in Latin America for the *Food And Agriculture Organization of the United Nations* (FAO).

4.4. Snow monitoring

This use case focuses on algorithms for detecting snow cover and snow status based on the combination of Sentinel-1 and Sentinel-3 time series. It includes basic user functionalities as e.g. user authentication or data management. This work flow will be tested in South Tyrol by the *Hydrologic office of the Province of Bolzano* to enrich their downstream services.

4.5. Agricultural monitoring

A test case on agriculture monitoring, based on Sentinel-1 and -2 time series will be tested on the JRC back-end (JEODPP). The use case includes functionalities as data extraction and reduction, machine learning methods and interactive data view applications, integrating 3rd party reference datasets.

5. USER INVOLVEMENT

openEO is an open source project and as such depends critically on user input – from remote sensing experts and software developers to back-end providers. Therefore the project consortium has established different communication channels:

1. A website was implemented, explaining in detail the ongoing development of the project: http://openeo.org/.
2. The newest version of the openEO source code is freely available at GitHub. The page also entails maintained user forums: https://github.com/Open-EO.
3. A first documentation and user manual is available at https://open-eo.github.io/openeo-api/.
4. Introduction videos and manuals are available on YouTube. Latest news regarding openEO are published via twitter: https://twitter.com/Open_EO.
5. An email hotline was established to provide users with a direct contacting possibility: `openeo@list.tuwien.ac.at`.

Furthermore, the project consortium hosts hackathons and publishes user questionnaires to enable interested users steering the openEO's path during and after the project's period.

6. ACKNOWLEDGEMENT

This project receives funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 776242). This paper reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, October 2009. [Online]. Available: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

[2] S. Schade, "Big data breaking barriers - first steps on a long trail," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XL-7/W3, pp. 691–697, apr 2015.

[3] "Copernicus open access hub," https://scihub.copernicus.eu/, accessed: 2018-10-22.

[4] "Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions: European cloud initiative – building a competitive data and knowledge economy in europe," apr 2016. [Online]. Available: http://ec.europa.eu/transparency/regdoc/rep/1/2016/EN/1-2016-178-EN-F1-1.PDF

[5] E. Pebesma, W. Wagner, J. Verbesselt, E. Goor, C. Briese, and M. Neteler, "Openeo: a gdal for earth observation analytics," http://r-spatial.org/2016/11/29/openeo.html, accessed: 2018-10-22.

[6] E. Pebesma, W. Wagner, M. Schramm, A. Von Beringe, C. Paulik, M. Neteler, J. Reiche, J. Verbesselt, J. Dries, E. Goor, T. Mistelbauer, C. Briese, C. Notarnicola, R. Monsorno, C. Marin, A. Jacob, P. Kempeneers, and P. Soille, "Openeo – a common, open source interface between earth observation data infrastructures and front-end applications," *Zenodo*, 2017.

[7] D. Sabel, Z. Bartalis, W. Wagner, M. Doubkova, and J.-P. Klein, "Development of a global backscatter model in support to the sentinel-1 mission design," *Remote Sensing of Environment*, vol. 120, pp. 102–112, may 2012.

[8] D. Nguyen, K. Clauss, S. Cao, V. Naeimi, C. Kuenzer, and W. Wagner, "Mapping rice seasonality in the mekong delta with multi-year envisat ASAR WSM data," *Remote Sensing*, vol. 7, no. 12, pp. 15 868–15 893, nov 2015.

[9] V. Naeimi, S. Hasenauer, S. Cao, B. Bauer-Marschallinger, A. Dostalova, S. Schlaffer, and W. Wagner, "Monitoring water resources using big data from sentinel-1 satellites," in *Proceedings of the 2014 conference on Big Data from Space (BiDS'14)*, Nov. 2014.

[10] J. Reiche, S. de Bruin, D. Hoekman, J. Verbesselt, and M. Herold, "A bayesian approach to combine landsat and ALOS PALSAR time series for near real-time deforestation detection," *Remote Sensing*, vol. 7, no. 5, pp. 4973–4996, apr 2015.

doi:10.2760/848593

THE BETTER PROJECT – DELIVERING CONTINUOUS EO BASED DATA STREAMS TO ADDRESS KEY SOCIETAL CHALLENGES

Nuno Grosso(1), Fabrice Brito(2), Pedro Gonçalves(2), Hervé Caumont(2), Simon Scerri(3), Mohammad Nammous(3), Rogério Bonifácio(4), Valentin Pesendorfer(4), Michele Lazzarini(5), Anca Popescu(5), Sergio Albani(5), Andrea Manconi(6), Nikhil Prakash(6), David Petit(7), Vânia Fonseca (1), Nuno Almeida(1), Diego Lozano(8), Nuno Catarino(1)

(1) Deimos Engenharia, (2) Terradue Srl, (3) Fraunhofer Institute for Intelligent Analysis and Information Systems, (4) World Food Programme, (5) European Union Satellite Centre, (6) Swiss Federal Institute of Technology – Zurich, (7) Deimos Space UK, (8) Deimos Space Spain

ABSTRACT

BETTER is a H2020 project under the EO-2-2017 EO Big Data Shift call. Its main objective is to implement an integrated EO Big Data intermediate service layer devoted to harnessing the potential of Copernicus EO data directly from the needs of the users. BETTER aims to go beyond the implementation of generic Big Data tools and incorporate those tools with user experience, expertise and resources. The service layer will deliver customized solutions - Data Pipelines - for large volume EO and non-EO datasets access, processing, analysis and visualisation. The developments are driven by 36 Big Data Challenges set forward by Key Societal Challenges actors (the challenge promoters). WFP, SatCen and ETH – Zurich, working in the areas of Food Security, Geospatial Intelligence *and Geo*-Hazards. Each will introduce 9 challenges over the course of 3 years. The success of BETTER relies on the experience and versatility of the consortium team responsible for service/tool development from DEIMOS and Terradue. This is complemented by Fraunhofer Institute's experience in Big Data systems, which brings transversal knowledge extraction technologies and tools that will help bridge the current gap between the EO and ICT sectors. Today, at the end of the project's first yearly challenge cycle, several pipelines are already developed or under development, and will be made available to other users in a First BETTER Hackathon to take place mid next year.

Index Terms— EO Big Data Shift, Data Challenges, Data Pipelines, Food Security, Geohazards, Geospatial Intelligence

1. INTRODUCTION

In the last decades, the European Space Agency (ESA) and the European Commission (EC), in partnership with other institutions and service operators such as EUMETSAT, have been developing a sustainable ecosystem of Earth Observation (EO) satellite missions, data providers, distributors, storage and processing centres, added value service providers and most importantly end users. Among these initiatives, Copernicus is the European flagship programme, amounting to €8.4bn investment from the EC up to 2020 [1]. Copernicus ecosystem has been consolidated through ground-breaking initiatives such as the launch of the Sentinel satellites constellation, providing operational satellite data, as well as the implementation of the Copernicus Services in different thematic areas, and is complemented by ESA's own efforts in the Thematic Exploitation Platforms and other initiatives, making Europe the worldwide leader in the Earth Observation field.

In parallel to the improved data capabilities available in Europe, the downstream sector has been developing capabilities towards using EO to support decision making processes, and several EO based services are reaching operational status in many business domains, from pipeline infrastructure monitoring to forestry management.

BETTER is implementing a Big Data intermediate service layer focused on user-centric services and tools, and addressing the full data EO data lifecycle to bring more downstream users to the EO market and maximize exploitation of Copernicus data and information services [2]. These customized solutions, denominated as Data Pipelines, are driven by 36 Data Challenges defined by users from key Societal Challenges sectors. The Data Pipelines will facilitate the downstream usage of large volume and heterogeneous datasets, so that users can focus on the analysis of the extraction of the potential knowledge within the data and not on the processing of the data itself. Their main purpose continuous delivery of large volumes of higher level EO products to users, customized to their needs.

This paper presents the BETTER project concept and objectives, the current developments after the first project year, the expected results delivered at the project completion, and its impact on the long term.

2. DATA CHALLENGES

During the BETTER project, challenges are introduced by promoters within the consortium addressing key societal areas, such as Food Security, Geospatial Intelligence

(GEOINT) and Geohazards. A total of 27 challenges are proposed during three yearly challenge cycles, focused on problems such as detection of droughts, assessment of illegal crop cultivations or early warning of significant earthquakes, with an additional 9 brought by external promoters.

In the field of Food Security, the United Nations World Food Programme (WFP) is the main challenge promoter. Through the development of BETTER Pipelines, WFP expects to improve its preparedness in order to better address food security issues in humanitarian crises using Sentinels data, Copernicus Land, Atmosphere and Climate Change Services together with in-house datasets.

The requirements will focus on the development of higher-level products based on those datasets to provide reliable early warning information and support the decision-making process during operational activities. The datasets are related mostly to crop monitoring and meteorology, can be derived from the Copernicus Land, Atmosphere and Climate Change Services. Such possibilities will be explored in the framework of the project.

Main topics in this challenge are related to:

- Multi-temporal EO for Humanitarian Operations;
- Hot-Spot EO Analysis Capacity for Natural Hazard Impact Assessment;
- Dynamic Land Cover Change Detection and Characterization.

In particular, the three challenges addressed in first challenge cycle and currently under development are focused on delivering:

- Sentinel-1 Sigma-0 and coherence time series data to be used by WFP for wetlands/water bodies and multi-temporal flood progression monitoring, extracting qualitative information on soil moisture or as an input for crop classification in specific areas of interest;
- Sentinel-2 and Landsat based vegetation indices and Sentinel-1 backscatter time series so that WFP can extract information on the impact of their restoration activities (irrigation canals, land rehabilitation, dams, roads, etc.) on local environment and communities or the impact of conflicts on agricultural resources in hard-to-access areas (e.g. cropland abandonment, land degradation)
- Global vegetation, precipitation and land temperature time aggregated datasets and respective anomalies to a reference period to support the seasonal monitoring and early warning activities in all WFP regions.

The challenge promoter with respect to Geospatial Intelligence is the European Union Satellite Centre (SatCen). The objective is to enhance the capabilities of the Geospatial Intelligence community in the Space and Security domain through the provision of improved EO products and applications exploiting Big Data methods and techniques. The rapidly increasing amount and variety of data coming from satellites and other sources in the Space and Security domain is raising new issues such as the management and exploitation of extremely large and complex datasets.

Currently, the SatCen supports the decision making and actions of the EU in the field of Common Foreign and Security Policy (CFSP), in particular Common Security and Defence Policy (CSDP), including European Union crisis management missions and operations, by providing products and services resulting from the exploitation of relevant space assets and collateral data, including satellite imagery, aerial imagery, and related services. In particular, the SatCen Research, Technology Development and Innovation (RTDI) Unit has the primary role to assess state-of-the-art technologies as Big Data and deliver innovative geospatial management solutions in order to improve the SatCen operational capabilities to offer EO products and services to Space and Security stakeholders.

In the GEOINT thematic area, BETTER will explore the added value provided by Big Data methods in developing products and applications tackling the following main topics:

- Change Detection and Characterization [3];
- Land Use / Land Cover;
- Thematic Indexes [4].

The first challenge cycle is focused on providing the following data streams for defined areas of interest:

- Sentinel-2 based multitemporal stack of thematic vegetation and water spectral indexes;
- Sentinel-1 multitemporal SLC and coherence stacks to detect changes over man-made and natural structures;
- DVI Maps, multitemporal stack of Sentinel-2 Bottom-of-Atmosphere reflectances and band color composites for the delineation of land cover types, with particular focus on vegetation.

Finally, the field of Geohazards is related to the analysis of global scale long time series EO based information of volcanic activity, earthquakes; Landslides and land subsidence are key in improving forecast and early warning systems for these natural disasters and highly demanding in terms of EO data volume, due to its spatiotemporal scale and the resolution of the required imagery. Currently the Swiss Federal Institute of Technology Zurich (ETHZ), the main challenge promoter in BETTER in Geohazards, is doing research in this field. Their work has been already been integrated in the GeoHazards TEP and its scope fits perfectly into the needs of the Copernicus Emergency Management, mainly in the Risk and Recovery Mapping component. Their work can also influence other related areas from civil protection to the insurance sector.

In this thematic area BETTER will provide data streams to ETHZ to derive information on volcanic activity, earthquakes, landslides and land subsidence to forecast and early warning of these natural disasters. Their research can be integrated in the scope of the Copernicus Emergency Management, mainly in the Risk and Recovery Mapping component. During the project these connections will be further analyzed to connect to additional promoters coming

doi:10.2760/848593

from these sectors and develop higher-level products that can bring additional value to the project.

The main topics in this challenge will be related to:

- Forecasting the impact of earthquakes;
- Rapid generation of landslide inventories;
- Forecasting surface deformation.

The focus of this first cycle will be on the production of a global catalogue of:

- ENVISAT-ASAR interferograms (including coherence maps) before and during the event for earthquakes with a magnitude higher than 5;
- Sentinel-1 interferograms before and after the event for earthquakes with a magnitude higher than 5;
- Sentinel-1 co-seismic deformation change maps before the event and co-seismic;

In total, these thematic challenges will encompass the implementation of sixteen Data Pipelines, delivering to promoters 41 different output data streams, including six global or near global output data streams and other outputs that cover 15 different areas of interest around the world. In addition, 7 new applications are under development to perform user defined on-demand functionalities on top of the provided Data Pipelines.

Additionally, in the second and third cycles, nine challenges will be brought by external challenge promoters. Several stakeholders are already engaged with the project to bring challenges related crop damage estimation related to extreme weather events, dynamic fire risk mapping and sustainable fishing. Other external challenge definition, selection and engagement are already underway and everyone is welcome to take part of this process and propose a challenge for the next early cycles. The next one challenge cycle will start in March 2019.

3. FROM CHALLENGES TO PIPELINES

The design, implementation and delivery of the Data Pipelines that answer the needs of the Data Challenges is driven by the definition of a set of pipeline requirements after the challenge definition process. Those requirements are extracted in a dedicated workshop in the beginning of each challenge cycle. After deriving this set of requirements are defined the BETTER development team composed by Terradue, Deimos and Fraunhofer design and implement the different pipeline components, following an Agile methodology. At each pipeline release, the promoters test and provide feedback to the development team, driving its improvements.

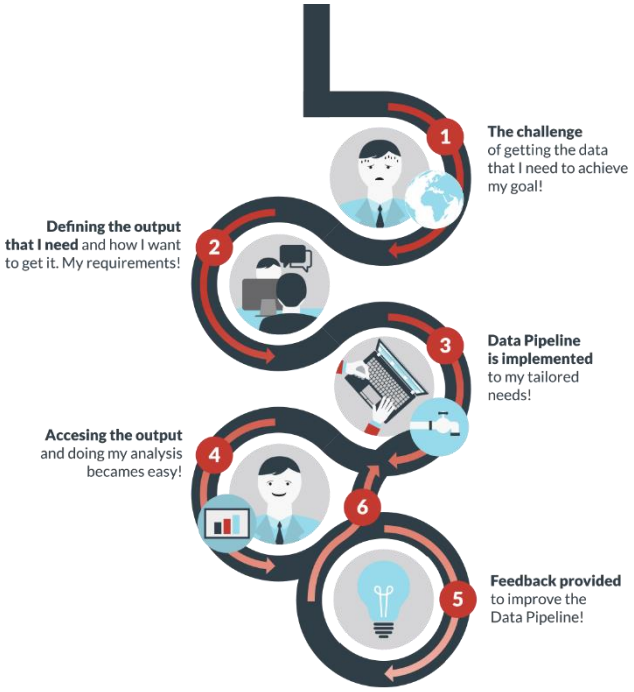


Figure 1 - From challenges to pipelines

The collaborative development environment of the pipelines is based on a Jupyter Lab interface where several workflows archetypes (e.g. SNAP based archetypes, OrfeoToolbox -OTB- archetypes) are available for the development teams. At each release, the different software components will be uploaded to a software repository and enter a continuous integration/deployment environment where it will be merged into the full pipeline, tested with a validation dataset provided by the promoters, built, packaged, dockerized and made available in a production center. In this center it will start producing the data streams continuously and making them available to users for them to test and provide feedback.

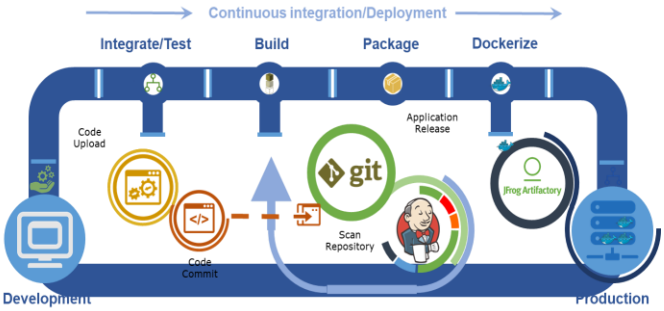


Figure 2 - Pipeline development process

When a pipeline is finalized and answers all the requirements of the promoters, the challenge is considered achieved by the users

doi:10.2760/848593

4. PIPELINE EXPLOITATION

After each pipeline implementation and challenge achievement, the main goal of the project is to promote the developed pipelines to other communities that might be interested in using those data streams.

In addition to communication, dissemination and engagement activities foreseen in the project there will be two hackathons open to everyone, one in the middle of the second challenge cycle and one in the middle of the third. In those hackathons users will be asked to develop additional processing components on top of the current pipelines to extract information customized to their needs. This will allow to: a) showcase the use of the platform and demonstrate how the developed pipelines can support these wider communities, b) to identify additional micro-challenges by interested stakeholders who are interested in using the platform and variations of the pipelines to address their own user stories; c) get the feedback from the people participating in the hackathon on the robustness and usability of the developed pipelines.

Figure 3 - Community Engagement in BETTER

5. CONCLUSIONS

BETTER brings a unique approach to exploit the potential of EO Big Data to help address the top priorities of key societal areas such as Food Security, Geospatial Intelligence and Geohazards, driven by thematic challenges set by promoters that are main stakeholders in their areas. By using this user-centric and flexible approach, BETTER will maximize impact of the built pipelines on the operational and R&D activities of the promoters and reach additional related user communities so that they start reusing those pipelines and building additional processing components on top of them. Currently, BETTER is reaching the end of the first yearly challenge cycle with several pipelines already developed or under development. Those pipelines will be made available to other users in a First BETTER Hackathon to take place in mid 2019.

6. REFERENCES

[1] Copernicus Programme www.copernicus.eu

[2] BETTER Project Website https://www.ec-better.eu/

[3] P. Boccardo, V. Gentile, F. G.Tonolo, D. Grandoni, M. Vassileva, Multitemporal SAR Coherence Analysis: Lava Flow Monitoring Case Study, *Proceedings of IGARSS 2015*, 978-1-4799-7929-5/15

[4] E. Mandanici, G. Bitelli, Preliminary Comparison of Sentinel-2 and Landsat 8 imagery for a Combined Use, *Remote Sensing*, 2016, 8, 1014; doi:10.3390/rs8121014

HIGH RESOLUTION SATELLITE IMAGERY AND POTENTIAL IDENTIFICATION OF INDIVIDUALS

Cristiana Santos, Delphine Miramont and Lucien Rapp

School of Law, University Toulouse1-Capitole, SIRIUS Chair

ABSTRACT

The synergy combining the forthcoming improvements of satellite imagery resolution, real-time space big data, facial recognition technology, and big data analytics might enable, in the near future, to discern more refined details on earth, and to identify individuals. Thus privacy and data protection concerns are being raised by court cases, legal scholars, few stakeholders and media. The intent of this paper is to define what is personal data in big space data, to discuss the possibility of identification of individuals, and to portray mitigation risk approaches for incoming space data policies.

***Index Terms*—** surveillance, data protection, privacy, big space data, big data analytics, identification

1. INTRODUCTION

As of today, the leading-edge imagery resolution commercially available is provided by DigitalGlobe's WorldView-3 satellite constellation, for which each pixel in a captured image corresponds to approximately 31 cm[1]. Notably, there is a tendency for pushing for the resolution restrictions threshold to be lowered to 10 cm. Yet, this does not suffice to directly identify individuals. Indeed there is a need to *demystify* satellite imagery and its powers.

As EO massive constellations of small satellites[2] are being launched in LEO, a bigger influx of high quality imagery and observation capabilities of EO satellites[3] are expected to become more widely available on a timely basis (capturing a single point several times a day) at a much lower cost. Users can plan both the target and frequency, allowing for a more specific analysis in a particular tracking.

Given the growing commercial market of high-res imaging, and the advancements in satellite technology and sensor resolutions, it is most likely that high-res space-based data will improve. And whilst low-cost, highly responsive commercial satellite systems become operational, very high-resolution (VHR) imagery is expected to become a regular attribute for end-user products and services. The *synergy* revolving around foreseeable improvements of satellite imagery resolution, facial recognition technology (and other image recognition software), real-time imaging, and big data analytical software, might enable to discern more refined details and identification of patterns of life in industry and in the environment. *Speculation* revolves around satellite imagery discerning car plates, individuals, and "manholes and mailboxes"[4]. It is claimed that such granular location-based information would only occur within *secondary use-cases*[5] (e.g., data analysis for smart cities, marketing profiles), and not undertaken by first-party uses (e.g. trends, improving geo-aware service)). The ITU-T Study Group 17 (SG17)[6], EO experts and legal scholars foresee that, in concomitance with the growing resolution of remote sensing images, the likelihood of privacy and data protection issues also grow[7]-[8]-[9] [1][5], demanding protection therefrom. One cannot always predict what the downstream forthcoming usages of VHR images will be, given the myriad of mashup tools and technologies available. As satellite images become more ubiquitous, as "god-like views", reflection on how they are created, and the purpose for their use is timely.

While relevant international space law—essentially the Outer Space Treaty, its follow-on treaties developed through COPUOS, and the UNGA Resolution 41/65, containing the Principles on Remote Sensing—do not address privacy concerns of VHR images, analyzing privacy and data protection risks is necessary to inform future regulations and satellite data policies towards General Data Protection Regulation compliance (Regulation (EU) 2016/679, GDPR)[6] [7]. The paper is organized as follows. Section 2 discusses how big data analytics and space data relate.

[1] http://worldview3.digitalglobe.com/

[2] It comprises 300 non-manoeuvrable 3U cubesats, Swiss Re Report "New space, new dimensions, new challenges: how satellite constellations impact space risk", 2018.

[3] Popkin, G."Technology and satellite companies open up a world of data", https://www.nature.com/articles/d41586-018-05268-w

[4] See US lifts restrictions on more detailed satellite images, http://www.bbc.com/news/technology-27868703

[5] Future of Privacy Forum, "Location Data:GPS, Wi-Fi, and Spatial Analytics", https://fpf.org/wp-content/uploads/2018/12/DDF-2-Materials.pdf

[6] https://www.itu.int/en/ITU-T/about/groups/Pages/sg17.aspx

[7] https://theconversation.com/ruling-on-sharper-satellite-images-poses-a-privacy-problem-we-can-no-longer-ignore-28133

[8] http://thescienceexplorer.com/technology/new-satellites-will-detect-your-face-and-phone-space

[9] http://www.digitalethics.org/essays/high-resolution-satellites-are-our-privacy-expectations-too-high; https://www.forbes.com/sites/patrickwwatson/2018/04/26/this-is-the-end-of-privacy-as-we-know-it/#27d88ee96875

doi:10.2760/848593

Section 3 defines personal data in space and argues on the potential identification of individuals and surveillance cases. Section 4 suggests mitigation risk approaches, while Section 5 concludes.

2. BIG DATA ANALYTICS AND SPACE DATA

Two main trends are currently developing in the satellite imagery industry: *(i)* the increasing availability of VHR satellite imagery; and *(ii)* the outsource processing-intensive image analysis tasks to distributed computing. Orbital Insight, SpaceKnow, Descartes Labs, Exogenesis, Remote Sensing Metrics, OmniEarth, DataKind are examples of analytic support companies offering actionable insights or intelligence. Distinctive aspects of big data analytics are briefly mentioned herewith to foresee its (potential) implications [8][9] on privacy and data protection: *(i)* use of large numbers of machine learning (ML) algorithms processing high-res satellite imagery in order to find automatic correlations, inferences from datasets. To note, Target Matching Recognition (TMR) algorithms for satellite images are improving robustness and accuracy, tackling image matching errors and reduce matching recognition time [10]; *(ii)* tendency to collect and analyze *all* the data that is available; *(iii)* repurposing of data for which it was originally collected, as analytics can mine data for new insights and find correlations; and (*iv)* use of new types of data automatically generated and coming from the IOT devices, as sensors. Besides, there is a growing use of *face-based technology systems* in commercial setting and such technology often involves the collection and use of personal data, requiring careful assessment of identifiability and privacy issues. Also, a series of other applications and payloads can also be installed on LEO smallsats, allowing the gathering and processing of personal data and seriously interfering with, and potentially violating citizens' rights to privacy and data protection, like high power zoom, facial recognition, behaviour profiling, movement detection, number plate recognition, thermal sensors, night vision, radar, see-through imaging, Wifi sensors, biometric sensors to process biometric data, GPS systems processing the location of the persons filmed, systems to read IP addresses and track RFID devices, systems to intercept electronic communications, etc.VHR satellite images linked with these significant artifacts carries the potential to increase risks of hampering privacy and data protection.

3. PERSONAL DATA IN BIG SPACE DATA AND IDENTIFICATION OF INDIVIDUALS

The scope of the GDPR on space data (art. 3(1)) conveys that any entity directly or indirectly processing data of EU residents is subject to the GDPR, even if taken from a satellite under the jurisdiction and control of a non-EU country. Therefore it is relevant to assess what is personal data in big space data.

3.1. What is personal data in big space data?

A large proportion of big data is not personal, namely, weather information, satellite imaging, and operational machine data. But some space big data may include elements that link directly to a person, and hence, could be considered personal data, as we shall conclude.

The GDPR regulates the use of multiple data formats – including images – which help to identify, either directly or indirectly, any person. Personal data is therein broadly defined, and includes *all* information related to an identified or identifiable natural person. An identifiable natural person can be directly or indirectly identified, in particular, by reference to other data (art. 4 (1)). Purtova contends that in the age of the Internet of Things (IoT), datafication, advanced data analytics, and data-driven decision-making, any information relates to a person [11], and therefore, it triggers data protection. Such assertion is also conveyed by the Article 29 Working Party[10] (WP29) opinion on the concept of personal data (WP136). The author further refines the three-based elements of personal data: i. any information; ii. relating to; iii. identified or identifiable natural persons. Herewith we shall discuss this third element – the possibility of identifying individuals through VHR satellite imagery. Personal data is broadly defined and includes all information on an identified or identifiable natural person who can be directly or indirectly identified in particular by reference to other data (art. 4 (1) and recital 26). This covers aspects such as name, address, card of phone numbers, IP addresses, etc. But it may also apply to a set of other data that *together* can relate to an identified or identifiable natural person such as, for instance, location data, video footage, public key, signatures, IP addresses, cookies, device identifiers, metadata, etc. The growing number of throughput satellites combined with increasing reliance on satellite technology for connectivity services extends the types of space data to any information that is shared through this means[11].

3.1.1. Potential identification and re-identification of individuals

Personal data includes *all* information related to an identified or identifiable natural person. The attribute “identified” refers to a known person, and “identifiable” is a person who is not identified yet, but identification is possible. One is *directly* identified or identifiable by reference to a name, in combination with additional “direct or unique identifiers”. These “direct and unique identifiers” covers data-types easily referenced and associated with an individual, including descriptors such as a name, ID number or username, location data, card of phone numbers, online

[10] The opinions of the WP29 are not formally binding, but possess “persuasive authority” on this domain..

[11] Conway N, “Why Geospatial Needs to Listen to GDPR”, 2017, https://www.gis-professional.com

doi:10.2760/848593

identifiers, etc. (art. 4 (1)). One is "*indirectly identifiable*" by combinations of indirect (and therefore not unique identifiers) that allow the individual to be singled out; they are less obvious information types which can be related to an individual, such as video footage, public key, signatures, IP addresses, cookies, device identifiers, metadata, and alike. The WP136 and Recital 26 offer for a two-fold standard for the possibility of identification. It establishes a dynamic test of "reasonable likelihood" of identification: i. whether or not all the means of identification are 'reasonably likely to be used to identify an individual; ii. either by the controller or any other person'. To assess such possibility of identification, it is needed to account objective factors such as: cost and time required for identification and the state of art of technology at the time of processing to enable identification. This dignifies that the capacity of (re) identification is increasing at the pace of technology developments. The growing number of throughput satellites with increasing reliance on VHR satellite technology, new analytical technologies extends the types of space data to any information that is processed and shared through these means, and enables identification. For instance, if the footage taken through VHR imaging only shows the top of a person's head and one cannot identify that person without using sophisticated means, it is not personal data. However, if the same photograph is taken in the backyard of a house with additional imaging analytical algorithms that enable identification of the house and/or the owner, that footage would be considered as a personal data. Thus, personal data is very much *context-dependent*. Arguably, a person – as a whole – can be depicted on these pictures, as for the resolution might allow for the identification of a person considering, for example, the person's height, body type and clothing could help in identifying a person on a very high resolution satellite image. Likewise, objects and places (location data) linked to a person could also enable identification of a person via VHR, such as the person's home, cars, boats and others. Identification can also be established through *combinations of data*. In fact, this scenario escalates with the advances of "ultra-high" definition images[12] published online, from commercial satellite companies, and the consequential application of big data analytic tools. It might be possible to identify *indirectly* an individual (and also to depict individual households, etc.), when high resolution images are combined with other spatial and non-spatial datasets.

Thus, while the footage of people may be restricted to "the tops of people's heads", once these images are contextualised by particular landmarks or other information, they may become identifiable. This other information can include "demographically identifiable information" (DII) or "community identifiable information" (CII), which may contain personal information therein, or otherwise transport, administrative, demographic categories, survey data available online, or other imaged information (geo-tagged or otherwise identifiable by location, and crowdsourced geographic information [12]).

3.1.2. Satellite Imagery and Surveillance

Satellite imagery cases have been dealt with in courts. Even if current decisions usually find that these kinds of images are not invasive to personal privacy and they try to find a balance between privacy and the opportunities that satellite technologies can offer, if satellite imagery continues to improve, enabling the identification of individuals due to a more precise resolution, the issue of privacy will become more important. At the European level, several decisions pointed out the dangers of mass surveillance. Satellites collect data as they continually orbit the globe. Since they do not aim a specific targeted surveillance, but collect images of swept areas, the question of surveillance represents an important issue. *Mass surveillance* has been judged to represent a particular serious interference with private life by the Court of Justice of the European Union (CJEU) in the *Digital Rights Ireland* case[13], which lead to the annulment of the so-called "Data Retention Directive"[14], about the storage of communication metadata of every user over a long period of time, without any reasonable suspicion of involvement in some kind of criminal offence. European courts also gave a jurisprudential framework to localization data regarding the right to privacy. Indeed, the use of GPS surveillance in *Uzun* case[15] and the use of data obtained therein in the criminal proceedings against him, breached article 8 (right to privacy) of the European Court of Human Rights (ECHR). In a case about real-time geolocation surveillance measures taken against Mohamed Ben Faiza[16] in a criminal investigation related to drug trafficking, the Court held a violation of the right to privacy. The ECHR pointed out in the *Klass case*[17] that "where a State institutes secret surveillance, the existence of which remains unknown to the persons being controlled (…), Article 8 could to a large extent be reduced to a nullity".

4. MITIGATION RISK APPROACHES

Imagery analysis or dissemination must be consistent with the "Common European Data Space" (SWD(2018)125 final)

[12] https://www.offthegridnews.com/privacy/googles-newest-high-res-satellites-can-monitor-your-every-move-in-real-time/

[13] CJEU, C-293/12, Digital Rights Ireland and others v. Ireland, 8 April 2014, ECLI:EU:C:2014:238.

[14] Directive 2006/24/EC.

[15] ECtHR, Uzun v. Germany, 2 September 2010, application n° 35623/05.

[16] ECtHR, Ben Faiza v. France, 8 March 2018 , application n° 31446/12.

[17] ECtHR, Klass and others v. Germany, 6 September 1978, application n° 5029/71, § 36.

and in furtherance of the GDPR. A privacy-centric taxonomy of identification approaches could be a useful standing to further identify risk scenario missions. Given the rapidly changing technology and the big space data context, it is advisable that privacy issues be considered at every stage of a dataset´s life cycle, and not only to the point of selling. Release of datasets of images that might raise potential privacy issues might call for a special regime of licensing that restrict their use to certain contexts (e.g. noncommercial), or that prohibit activities aimed at re-identification. Nevertheless, such licensing terms would depend on their compliance by its users and on legal action when breaches occur [13]. Information and transparency protocols, both on the missions and the operators, should be devised and implemented, as well as codes of conduct (by industry groups of remote sensing satellite operators) with recommended practices for big space applications, or guiding on the different categories of data that require special care. Data controllers can proactively carry out a data protection and privacy impact assessment processes,[18] notably where there are risks for data protection and privacy, respectively, according to typical VHR scenarios, e.g. this undertakes defining the purpose of the use; choosing the right tools; using the most privacy friendly approaches, or privacy-aware analytics methods; ensuring the security of the data collected, etc. These processes require that before using a privacy-limiting device, means must be in place to limit the impact as far as possible. A dialogue with manufacturers could be envisioned to preemptive implement privacy by design and by default measures and embed data protection requirements in data space applications to ensure compliance from the outset. Remote sensing companies can set up mechanisms to automatically process images by blurring faces, filtering out or obscure identifiable features on, house holding, whenever identification scenarios occur due to forthcoming image improvement. The Remote Sensing Principles contain no specific restrictions on what may be observed, therefore, it could be envisioned an updating of these principles encompassing plausible risks to privacy and data protection.

5. CONCLUSIONS

It is not possible today to directly identify an individual´s face using today's satellites. However, we elaborate upon a forward looking perspective. The opportunities provided by VHR satellite images are inherently linked with significant data analysis rendered by big data analytics and facial recognition technology commercially available which enhances identification and privacy risks. Geoprocessing, spatial analysis and other geo-intelligence tools will need to abide to “geo-privacy” compliance.

[18] https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/

6. REFERENCES

[1] Chun S., Atluri V., Protecting Privacy from Continuous High-Resolution Satellite Surveillance. In: Thuraisingham B., et al. (eds) *Data and Application Security*. IFIP, v. 73. Springer, 2002.

[2] Von Der Dunk, F, “Europe and the 'Resolution Revolution': 'European' Legal Approaches to Privacy and their Relevance for Space Remote Sensing Activities”, *Space and Telecommunications Law Program Faculty Publications*, p. 810, 2009.

[3] Von der Dunk, F., "Outer Space Law Principles and Privacy", in *Evidence from Earth Observation Satellites: Emerging Legal Issues*, Leung D. et al. (eds), Leiden: Brill, 243–258, 2013.

[4] European Space Policy Institute, “*Current Legal Issues for Satellite Earth Observation*”, p. 38, 2010.

[5] Mooney, P, Olteanu-Raimond, et al., “Considerations of Privacy, Ethics and Legal Issues”, in Volunteered Geographic Information. Foody, G, et al. (eds.) *Mapping and the Citizen Sensor*. 119–135. London: Ubiquity Press, 2017.

[6] Stefoudi D., Space Big Data, Small Earth Laws: Overcoming the Regulatory. Barriers to the Use of Space Big. *Proc. of the 2017 conference on Big Data from Space*, EU Publications, 2017.

[7] Cohen, B., Remote Sensing and the New European GDPR, Proc. of the IISL 2017, *Eleven International Publishing*, 2017.

[8] Waterman, K., Bruening, P, “Big Data analytics: risks and responsibilities*”, International Data Privacy Law*, v. 4, Issue 2, 89-95, 2014.

[9] Mantelero A., Vaciago G., *“*The 'Dark Side' of Big Data: Private and Public Interaction in Social Surveillance. How data collections by private entities affect governmental social control and how the EU reform on data protection responds in Social Surveillance*”, CLRI*, v. 14, 161-169, 2013.

[10] Chen, Y., Wei Xu, W. et al., “Target Matching Recognition for Satellite Images Based on the Improved FREAK Algorithm”, *Mathematical Problems in Engineering*, 2016.

[11] Purtova, N., “The law of everything. Broad concept of personal data and future of EU data protection law”, in *Law, Innovation and Technology*, Routledge, 2018.

[12] Mooney, P, Olteanu-Raimond, et al., “Considerations of Privacy, Ethics and Legal Issues” in Volunteered Geographic Information. In: Foody, G, See, L, et al. (eds.) *Mapping and the Citizen Sensor*, 119–135. London: Ubiquity Press, 2017.

[13] Borgesius, F., Gray J., et al, “Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework”, *Berkeley Technology Law Journal*, 30, no. 3, 2073-2130, 2015.

CONTINENT WIDE MONITORING OF GLACIER SURFACE ELEVATION CHANGES AND GLACIER MASS BALANCES

Thorsten Seehaus, Philipp Malz, Christian Sommer, David Farias, Matthias Braun

Institute of Geography, Friedrich-Alexander University of Erlangen-Nuermberg,

91058 Erlangen-Tennenlohe, Germany

ABSTRACT

Significant glacier shrinkage due to climate change has been reported in many mountain regions all around the world. However, continent wide detailed measurements of glacier ice mass losses are missing or quite sparse. Therefore we analyze interferometric SAR data from the SRTM and TanDEM-X mission in order to measure surface elevation changes of glacierized mountain regions world wide. An automatic differential interferometric processing chain, to generate digital elevation models from bistatic TanDEM-X data, and an elaborate referencing and mosaicing algorithm is applied to obtain information on glacier mass balances on continent scale. The procedure was developed and tested using whole South America as a test region. The analysis of the other continents is currently running. The results will deliver fundamental information for glacier mass balance and climate change projections, water resource management plans as well as politicians and decision makers.

1. INTRODUCTION

Glaciers and ice caps outside of the polar regions are strongly affected by climate change and are defined as key indicators for climate change by the Intergovernmental panel on Climate Change (IPCC) [1]. Within the framework of the Global Climate Observing System (GCOS), they are specified as Essential Climate Variables (ECV), due to their importance as fresh water source and storage. In many high mountain and arid regions, glacier melt water is a fundamental water source, but also for downstream communities, like e.g. along the large rivers in Asia, glacier runoff is an important water supply for hydropower, irrigation and wet-land ecology [2]. Thus, the continent wide monitoring of the current and prospective glacier change rates are of high interest for international initiatives but also for regional water resource management plans.

The TanDEM-X mission of the German Space Agency (DLR), provides high resolution bistatic interferometric Synthetic Aperture Radar (SAR) imagery worldwide since 2010 [3]. This data is suitable to derive highly precise digital elevation models (DEM). In combination with the results of the Shuttle Radar Topography Mission (SRTM) in February 2000, whose aim was to generate a consistent digital DEM of the landmasses between 60°N and 56°S, the monitoring

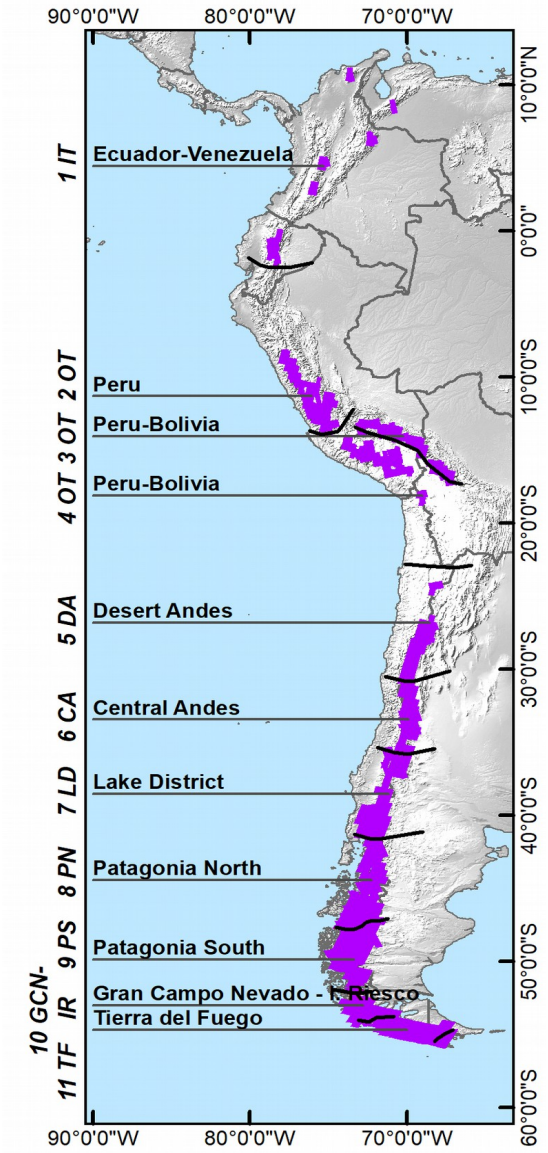


Figure 1: Used TanDEM-X coverage on glaciated areas in South America. Purple polygons indicate the TanDEM-X acquisitions; Black lines are delineations between glacier regions base on their climatic setting.

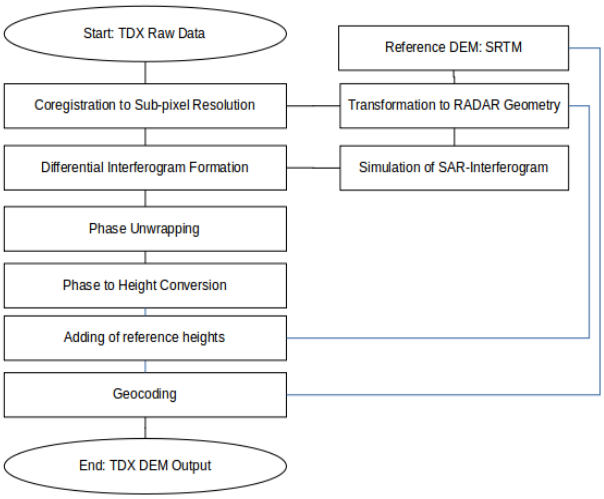


Figure 2: Flow chart of differential interferometric TanDEM-X DEM generation

of glacier surface elevation changes and consequently glacier mass balances since 2000 is feasible.

2. DATA & METHODS

Several products were derived from the bistatic SAR acquisitions of SRTM. For our analysis, we use the void-filled LP DAAC NASA Version 3 SRTM DEM with 1 arcsec (~30 m) ground resolution, which resulted from the C-band radar recordings [4]. The individual SRTM DEM tiles are mosaiced to cover the respective study area and reprojected to UTM projection. TanDEM-X is acquiring data in X-band since 2010. In 2012-2013 during the global DEM mission a nearly complete coverage of the landmasses was obtained, with ascending and descending data takes especially in mountain regions, allowing for continuous region wide analyses. Both sensors used different SAR frequencies, which lead to differences in the SAR signal penetration into snow and ice. The Radar signal penetration into glacier surfaces strongly depends on the surface type, conditions and water content [5]. In order to account for this issue, we selected preferable TanDEM-X scenes from the same season as the SRTM data, to reduce the bias due to differences in the SAR signal penetration, but also due to seasonal surface elevation changes. The spatial coverage of the analyzed TanDEM-X acquisitions of our case study throughout South America is illustrated in Figure 1.

The TanDEM-X data was processed following the approaches of [6], [7] (see Figure 2 for an overview of the processing chain). First acquisitions from the same track and date are concatenated in along track direction if possible. Then, a differential interferogramm is generated using the SRTM DEM as elevation reference. In the next steps the interferogramm is filtered, unwrapped by applying either the branch cut or minimum cost flow algorithm and the differential phase is transferred in to differential elevations. Subsequently, the elevation information of the SRTM DEM is added to obtain absolute height information and the product is finally geocoded. The best results of both phase-

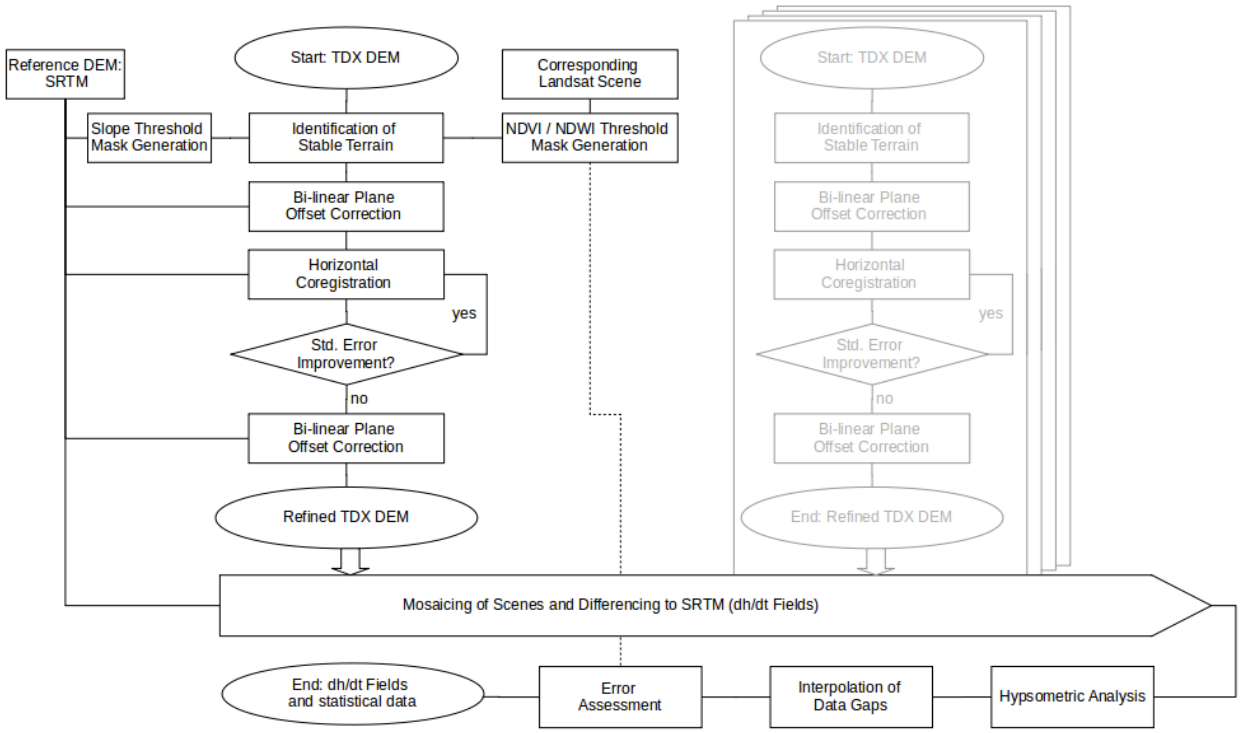


Figure 3: Flow chart of TanDEM-X DEM coregistration, mosaicing and glacier mass balance calculation routine.

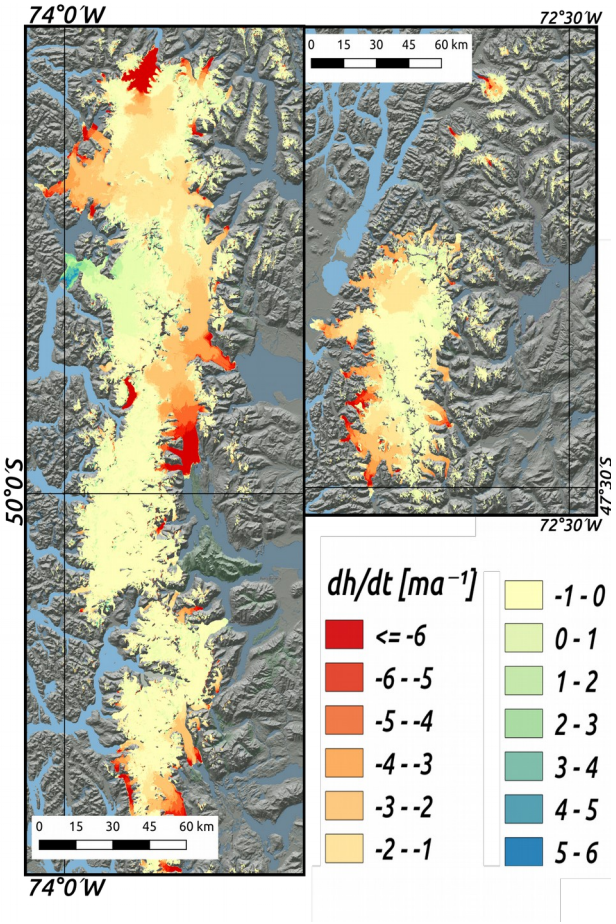


Figure 4: Surface elevation change rates of the Southern (left) and Northern (right) Patagonian icefields.

unwrapping approaches are manually selected for the further post-processing routines.

In order to map highly accurate elevation change data on the glaciated area, the TanDEM-X DEM tiles need to be precisely horizontally and vertically coregistered to the SRTM data (see Figure 3 for an overview of the processing chain). Therefore ice-free areas are defined as stable reference regions, by masking out vegetation (NDVI filter) and water (NDWI filter) and an applying an additional slope threshold of 15°. First, the TanDEM-X tiles are bi-linearly vertically corrected for offsets to the SRTM DEM, measured on this stable regions. Subsequently, a horizontal and vertical coregistration between the SRTM DEM and the TanDEM-X DEMs is carried out following a widely used approach [8]. Afterwards, the TanDEM-X DEMs are again bi-linearly vertically coregistered to the SRTM DEM to reduce still remaining biases. Finally the TanDEM-X DEMs are mosaicked to one regional DEM and a date stamp is added to each grid cell.

To calculate the elevation change rates dh/dt for the study periods, the SRTM DEM and the regional TanDEM-X DEM mosaics are differenced. Since, data voids in the applied SRTM DEM are filled with data from other sources (without date information), the SRTM data voids are filtered out using the masks provided by LP DAAC NASA. Moreover, regions with slopes steeper 50° are also rejected, since major ice aggregation is there quite unlikely (avalanche slopes) and DEMs are less accurate on these steep slopes [9].

The elevation change rates are integrated over the glaciated regions and multiplied by an average ice density, in order to obtain regional geodetic mass balances. To account for data voids in the elevation change fields on glaciated areas, the measured dh/dt values are area weighted base on the hypsometric distribution.

Finally a detailed accuracy estimation is carried out by considering error contributions from:

- DEM registration
- Hypsometric interpolation of data gaps
- SAR signal penetration bias
- Glacier area delineation
- Ice density

3. RESULTS

Here we briefly presents results from South America, our test region for our continent wide glacier surface elevation and mass balance monitoring algorithm. More detailed information can be found in the corresponding publication [10].

The glaciers and icecaps in South America cover an area of ~37751 km² and stretch through various climate zones, from the inner tropics in Venezuela to sub-Antarctic in Tierra del Fuego. Most prominent are the large ice fields in Patagonia. The Northern Patagonian Icefield (NPI, 4653 km²) and the Southern Patagonian Icefield (SPI, 13231 km²). In average we measured glacier elevation changes on 85% of the glaciated areas. (Note: large proportions of the non-map areas are due to voids in the SRTM data especially in the outer tropical regions.) Throughout the study region a mass change rate of -19.43 ± 0.60 Gt/a, corresponding to a specific mass balance rate of -0.61 ± 0.07 m w.e. a^{-1}, is obtained. The major ice losses of about 83% are caused by the Patagonian icefields (see Figure 4), caused by dynamic adjustments of their large outlet glaciers. The glaciers in the tropical regions show surface lowering as well, but at more moderate rates (see Figure 5).

4. CONCLUSIONS AND OUTLOOK

The application of our processing algorithms at the study region reveals good quality results and provides the first continent wide spatially detailed analysis of glacier mass balances throughout South America. There is sufficient suitable TanDEM-X data in other glaciated mountain regions around the world for comparison with SRTM available. Moreover the data amount is continuously increasing, since another global coverage like the first DEM mission is currently ongoing. This will allow for further

doi:10.2760/848593

updated monitoring of ice mass changes worldwide by comparing TanDEM-X to TanDEM-X data.

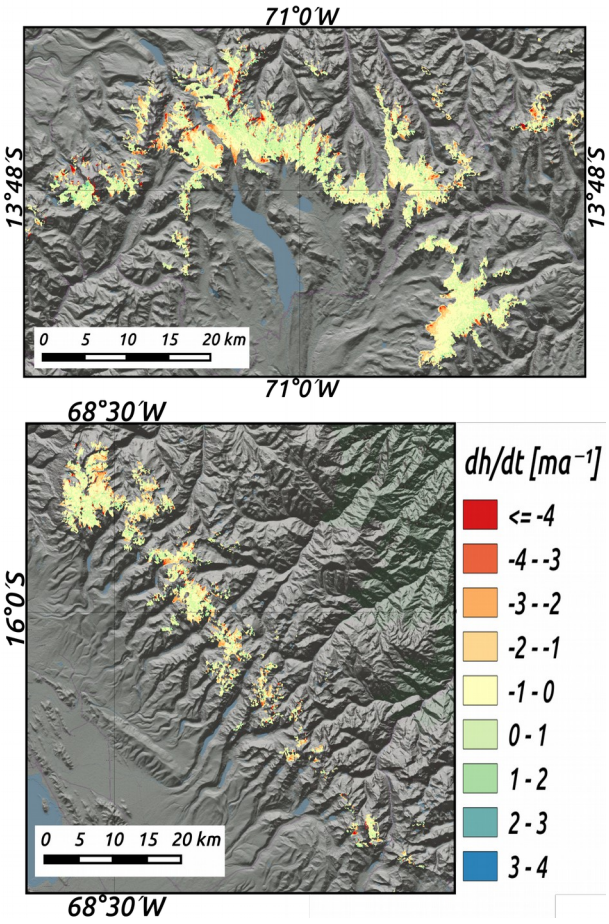


Figure 5: Surface elevation changes rate of glaciers in the outer tropics. Top: Cordillera Vilcanota Peru; Bottom: Cordillera Real, Bolivia

5. ACKNOWLEDGEMENTS

This work is financially supported with the grant BR2105/14-1 within the DFG Priority Program "Regional Sea Level Change and Society" and by grant SA2339/3-1, the BMBF-CONICYT project GABY-VASA (01DN15020, BMBF20140052), the DLR/BMWi grant GEKKO (50EE1544), the HGF Aliance Remote Sensing & Earth System Dynamics and FONDECYT 1161130 and BECAS-Chile.

6. REFERENCES

[1] IPCC, Ed., *Climate Change 2013 - The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press, 2014. http://ebooks.cambridge.org/ref/id/CBO9781107415324

[2] M. Carey, O. C. Molden, M. B. Rasmussen, M. Jackson, A. W. Nolin, and B. G. Mark, "Impacts of Glacier Recession and Declining Meltwater on Mountain Societies," *Annals of the American Association of Geographers*, vol. 107, no. 2, pp. 350–359, Mar. 2017. https://doi.org/10.1080/24694452.2016.1243039

[3] G. Krieger *et al.*, "TanDEM-X: A Satellite Formation for High-Resolution SAR Interferometry," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 11, pp. 3317–3341, Nov. 2007. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4373373

[4] NASA JPL, "NASA Shuttle Radar Topography Mission Global 1 arc second Version 3." NASA LP DAAC, 2013. DOI: 10.5067/MEaSUREs/SRTM/SRTMGL1.003

[5] E. Rignot, K. Echelmeyer, and W. Krabill, "Penetration depth of interferometric synthetic-aperture radar signals in snow and ice," *Geophysical Research Letters*, vol. 28, no. 18, pp. 3501–3504, 2001. http://onlinelibrary.wiley.com/doi/10.1029/2000GL012484/full

[6] P. Malz, W. Meier, G. Casassa, R. Jaña, P. Skvarca, and M. H. Braun, "Elevation and Mass Changes of the Southern Patagonia Icefield Derived from TanDEM-X and SRTM Data," *Remote Sensing*, vol. 10, no. 2, p. 188, Jan. 2018. http://www.mdpi.com/2072-4292/10/2/188

[7] T. Seehaus, S. Marinsek, V. Helm, P. Skvarca, and M. Braun, "Changes in ice dynamics, elevation and mass discharge of Dinsmoor–Bombardier–Edgeworth glacier system, Antarctic Peninsula," *Earth and Planetary Science Letters*, vol. 427, pp. 125–135, Oct. 2015. http://www.sciencedirect.com/science/article/pii/S0012821X15004100

[8] C. Nuth and A. Kääb, "Co-registration and bias corrections of satellite elevation data sets for quantifying glacier thickness change," *The Cryosphere*, vol. 5, no. 1, pp. 271–290, Mar. 2011. http://www.the-cryosphere.net/5/271/2011/

[9] T. Toutin, "Three-dimensional topographic mapping with ASTER stereo data in rugged topography," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 10, pp. 2241–2247, Oct. 2002. DOI: 10.1109/TGRS.2002.802878

[10] M. H. Braun *et al.*, "Constraining glacier elevation and mass changes in South America," *Nature Climate Change*, Jan. 2019. https://www.nature.com/articles/s41558-018-0375-7

ENSURING SPATIAL AND TEMPORAL CONSISTENCIES FOR THE TIME SERIES OF THE COPERNICUS LAND MONITORING PAN-EUROPEAN HIGH RESOLUTION LAYERS

Christophe Sannier, Sophie Villerot, Alexandre Pennec, Alice Lhernould, Clémence Kenner, Antoine Masse

SIRS - CLS Group, 27 rue du Caroussel, 59650 Villeneuve d'Ascq, France

ABSTRACT

Pan-European products assessing the sealed areas, spanning on more than a decade and still in production for 2018, the time series of High Resolution Layer Imperviousness has been making use of multiple sensors, whose data volume is still increasing, in particular with the introduction of Sentinel constellations, at multiple temporal and spatial scales. In this paper, we review the methodologies developed within the HRLs production and enhanced during the H2020 ECoLaSS project to ensure the coherence of this times series, whose updated areas from one date to the next lays within the accuracy specifications.

Index Terms— High Resolution Layers, Sentinel, Copernicus, Change Detection, Time Series, ECoLaSS

1. INTRODUCTION

The urban population in 2014 represented 54% of the overall population, and is expecting to keep rising [1]. According to the World Health Organisation, the global urban population should grow approximately 1.84% per year from 2015 to 2020; this percentage slowly decreasing over the years to reach 1.44% per year between 2025 and 2030.

Despise the impression that the temporal change in urban area does not appear to be significant at the global scale, its impact on the neighbouring forests, agriculture, water systems, through consumption rise, can turned out to be critical. A close monitoring of the urban growth is necessary to ensure a sustainable development [2]. In most developing countries, urban growth is mainly driven by population growth. However, in Europe, population growth no longer increases substantially, but urban areas continue to expand, a phenomenon known as urban sprawl [3].

The Copernicus Land Monitoring (CLMS) is an effort coordinated by the European Environment Agency to produce land cover and land use information, through the CORINE Land Cover (CLC) dataset as well as the five High Resolution Layers (HRL) for each of the specific land cover characteristics: artificial areas, forest areas, grasslands, wetlands and water bodies, that should be soon complemented by a layer of small woody features. The imperviousness (IMP) products quantify the percentage of soil sealing in a status layer for a given year ($\pm$ 1 year) and capture the modifications from the previous status layer to the next into a change layer.

Those raster-based datasets are key to better inform policy makers on the spatial distribution, and extent of urban sprawl in particular for IMP, and are updated every three years.

Thanks to Research Executive Agency (REA) for funding, and to our partners (GAF, UCL, JR, DLR) for their contribution on the H2020 project ECoLaSS.

The H2020 project "Evolution of Copernicus Land Services based on Sentinel data" (ECoLaSS) [4] aims at developing and prototypically demonstrating selected innovative products and methods for future next-generation operational CLMS products of the pan-European and Global Components, based on a multi-temporal and multi-sensors approach. One of its objective is to lay out the feasibility of a higher update frequency for several products - namely, the imperviousness, forest (FOR) and grassland (GRA) layers - of the CLMS continental and global components, for mid-term (2018) and long-term (2020+) evolution. This increased update frequency implies the exploration of new methods to correctly identify the automated changes detected, to ensure the spatial and temporal coherence of those changes along the time series.

2. METHODS AND DATASETS USED TO GENERATE STATUS LAYERS

IMP is the HRL for which the longer time series is available. It consists of a series of 20m and 100m thematic raster status and change products derived from EO data for the 2006, 2009, 2012 and 2015 reference years. Up until 2015, the production of HRL Imperviousness degree (IMD) was based primarily on a combination of SPOT-4 and 5 and IRS LISSIII with a RapidEye coverage introduced in 2012, organized around two separate coverages at least 6 weeks apart during the vegetation growing season.

These coverages were serving multi-purposes: the CLC production as well as the one for 2006 IMD layers and other HRLs from 2012. However, the acquisition of a complete cloud free coverage has been problematic, having to rely on gap filling exercise at a final stage to ensure a near complete coverage. Therefore, the target to achieve complete coverage ($\pm$ 1 year) remains an operational difficulty. The ECoLaSS project will put to test the yearly Sentinel datasets (S1 and S2), through their incorporation in the processing chain.

The HRL IMP production has been largely focused on the creation of a reliable built-up mask, which is then combined with Normalized Difference Vegetation Index (NDVI) data to derive the IMD [5], from 0% to 100%, as displayed in Figure 1. A threshold set at 33% of IMD is then used to create the binary status layer between urban and non-urban areas. Most of the error sources for both layers (status and change) are attributable to the correctness of this input built-up mask.

2.1. Status layer for IMP 2015

Optical datasets from various sources e.g. Landsat-8, SPOT-5, Resourcesat-2, and S-2A, all resampled at a 20m resolution, have been used to generate the 2015 built-up mask. Biophysical variables such as the NDVI and additional parameters such as the Normalized Difference Built-up Index (NDBI) [6] have been computed and time

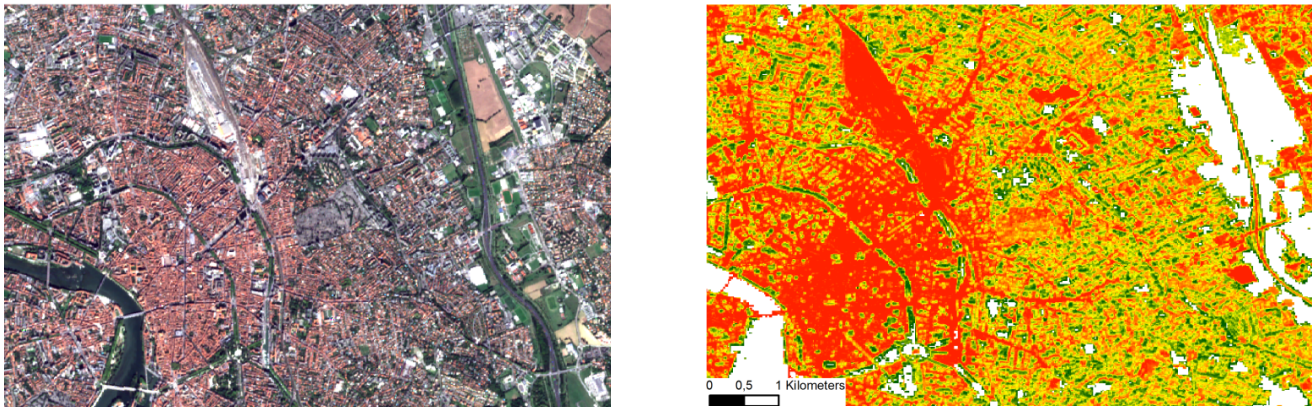
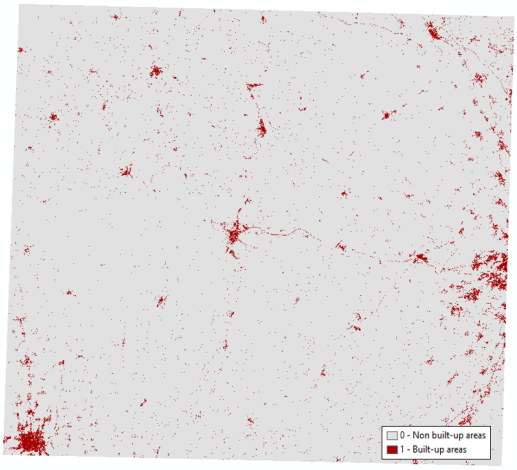
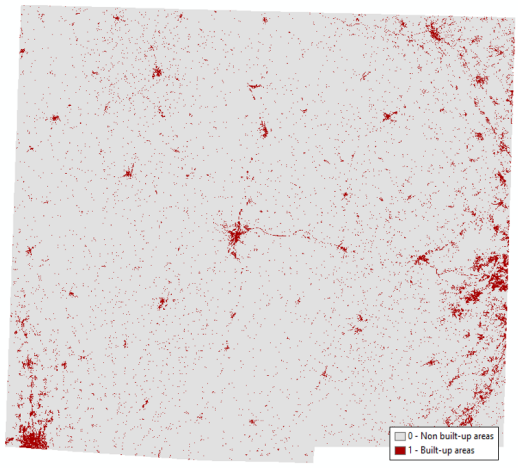


Fig. 1. On the left, S-2 optical image taken above Toulouse, France. On the right, matching HRL IMD layer for 2017, with 100% (fully impervious) in red down to 0% (no sealing) of IMD in green.



Original HRL IMD 2015

Fig. 2. Status layer for HRL IMP 2015, on 2 S-2 tiles in the South-West of France, at a 20m resolutions.



HRL 2017 Input data (t_n)

Fig. 3. Status layer for ECoLaSS prototype HRL IMP 2015, on the same area, deduced from the full dataset of S-2 images, cloud-free, with all the spectral bands available.

series statistics on the seasonal and yearly mean, median, maximum, minimum, standard deviation as well as seasonal and yearly range have been used to take full advantage of the cross-sensor seasonal time series of the top of atmosphere images.

Textural features, e.g. the angular second moment of the gray-level co-occurrence matrix, are also added as classification inputs to highlight the inherent heterogeneity of man-made building-structures [7].

All those variables have then been ingested in a semi-automatic classification approach, based on supervised trees, to identify the 2015 built-up areas, whose resulting classification can be seen on Figure 2.

2.2. Status layer for IMP 2017

In the framework of ECoLaSS, a new status layer for the year 2017, has been generated on a selected testing site (matching S-2 tiles 31TCJ, 30TYP) in South-West of France - yielding a change of spatial resolution for the layer at 10m from the previous 20m.

Classifications have been produced image-by-image by a fully automated processing chain, based on a random forest algorithm applied on a subset of the S-2 optical best scenes (pre-processed to get bottom of atmosphere reflectances) and several spectral and textural indices e.g. NDVI and NDBI. A support vector machine classifier is used to obtain classification results from Sentinel-1 SAR datasets. The resulting stack of classified layers (results from optical and SAR images) has then been merged using a Dempster-Shafer algorithm, with the overall precision as metric.

The final classification accuracies are slightly lower than the actual specifications of the HRL IMP (at 90% user and producer accuracies) but this could be easily improved by manual enhancement. The result can be seen on Figure 3. Please also note that Sentinel-1 classification contribution has been studied and quantified with a 3 points improvement for the global accuracy.

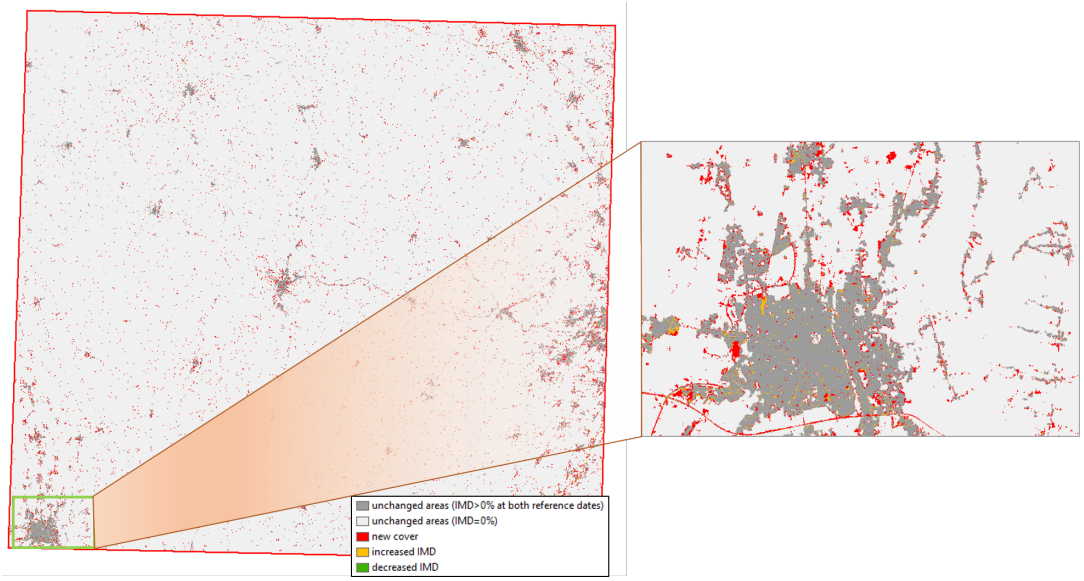


Fig. 4. Final IMP change layer between 2015 and 2017.

Table 1. Classification results for the IMP change layer 2015-2017

Total change areas	For the first calibration	For the second calibration
New built-up 2017	9%	9.64%
Omission: undetected built-up 2015	58%	76.65%
Commission: false built-up 2017	33%	13.71%

3. PRODUCTION OF THE CHANGES LAYER

Specifications only focus on the accuracy of status layers for which the target is set at 90% for both producer and user accuracy, but the results from the previous epochs show that for the IMP change layers, this level of accuracy is still above the expected level of change over the current 3-year period.

The imperviousness change detection relies on two input data described previously: the reference layer (t_0), the HRL IMP 2015, and the new status layer (t_n), the protoype HRL IMP for the year 2017.

The post-classification comparison, to attain full spatial and temporal consistencies, can be decomposed into three main steps:

- A spatial and temporal comparison based on the reference data (t_0) whose purpose is to enforce a geometrical harmonization between the different epochs to prevent problems related to an image-to-image approach;
- A post-processing filtering to remove a significant portion of noise due to small aggregated groups of pixels, which are most likely misclassifications;
- A contextual analysis based on change probability, such as discussed in [8] - this final step will consider the impervious pixels in the 2015 built-up mask to establish a probability map of changes. The analysis describes each pixel's relationship or membership to their neighboring pixels.

The assumption made using this final analysis is that urbanized areas spread more than they appear randomly in the landscape: the resulting urban membership estimates allows the isolation of change areas.

Errors can be present in the reference layer, and new errors could have appeared in the detection of change between two time epochs. They can be linked to:

- Omissions of change new urban areas that appear between 2015 and 2017 were not detected;
- Technical changes due to commission errors added for the new period, i.e. areas falsely flagged as new urban zones, as well as omission errors detected for the previous period, i.e. urban areas, already present in 2015, that were not then detected as such, but have now been flagged as urban areas in the 2017 layer.

A first validation based on a stratified ground truth collection is executed, and the first statistics can be found the second column of the Table 1. The relative magnitude of actual change is then estimated to 9% of the total change areas detected. Thus, errors concerning the remaining 91% of the change areas detected are related to the omission and commission errors detailed above. Regarding the omission errors from the previous epoch, the 2015 production was mostly based on Landsat-8 data whereas the 2017 built-up was produced from S-2 resulting in a nine-fold improvement in spatial resolution, since a Landsat pixel is characterized by nine S-2 pixels, explaining most of the omission errors origin.

This procedure relies heavily on the reference dataset for the statistical calibration of changes described above, which is used to produce statistics from which the estimate areas of each of the three categories in the change stratum will be interfered. Those areas provide then a basis to fine-tune the targets of re-processing, whose objective is to extract the real change areas. This step is achieved by adopting

doi:10.2760/848593

Table 2. Final classification results for HRL TCD change layer 2015-2018 on testing sites near Avignon, France.

	TCD change layer	
	Gain stratum	**Loss stratum**
In 2018	0.5% of gain	17.09% of loss
Omissions	undetected tree: 59% in 2015	undetected tree: 24.62% in 2018
Commissions	false tree detection: 40.5% in 2018	false tree detection: 58.29% in 2015

Table 3. Final classification results for HRL GRA change layer 2015-2018, on testing sites near Arles, France.

	GRA change layer	
	Gain stratum	**Loss stratum**
In 2018	2.5% of gain	14% of loss
Omissions	undetected grassland: 46.5% in 2015	undetected grassland: 24.5% in 2018
Commissions	falsely labeled GRA: 51% in 2018	falsely labeled GRA: 61.5% in 2015

a re-classification approach linking the three categories with suitable training data between 2015 and 2017 imagery.

The statistical computation is then reiterated on the reclassified change stratum, and the results, which can be found in the third column of Table 1, confirm the first rough outcome. Based on the re-processing, of the total area initially detected as changed, only 10% effectively represent new built-up areas while the remaining 90% are mostly omissions undetected in 2015 (76.7%) and new commission errors introduced by the 2017 new built-up mask (13.7%). Most of the omission errors concern small and isolated built-up features and roads, which is mostly attributable to resolution change between Landsat-8 and S-2. Regarding the commissions from 2017, mostly usual errors like small gardens, bare soils in the neighborhood of impervious scattered areas were found. The original change layer represented a nearly 50% increase of the artificial area in the test area which is unrealistic, considering that in fact over 75% of the detected changes were omission from 2015. In the re-classified layer, new built-up areas represent a 4% increase which appear more realistic and already represents a substantial increase over a 2-year period.

This methodology has been successfully implemented on test sites for two other HRLs. It is crucial that no substantial inbalance between omission and commission errors in the HRL change layers remains. Contrary to the IMP change layer, the Tree Cover Density (TCD) change layers and the Grassland (GRA) change layers are composed of two layers each, related to gain and loss, whose results for the reclassification can be found in Tables 2 and 3. The loss of impervious soils being extremely limited, only gain are presented in the change layer for IMP.

Regarding the GRA change strata, they represent together 9.5% of the total study area with losses representing an area 1.5 larger than gains, while the TCD change strata also amount to roughly 10% of the total study area with gains representing an area twice as large as losses. For this HRL, further steps need to be taken to ensure the thematic consistency, regarding the Dominant Leaf Type product, which can be either broadleaved or coniferous, but shouldn't switch between the two typologies over the course of different epochs.

4. CONCLUSION

The slow spatial progression of the sealed areas at European scale present a particular challenge in the frame of Copernicus HRLs. In this paper, the latest developments related to the Sentinel processing as well as the time series reanalysis are presented.

Through ECoLaSS, demonstration has been made of the added values of both Sentinel datasets, from Sentinel-1 (A and B) and Sentinel-2 (A and B) to orient the production toward a yearly release, all the while maintaining the integrity of HRL time series based on heterogeneous datasets coming from multiple sensors and tackling the increase volume of data, using temporal metrics in time-efficient automated algorithms.

New challenges are expected to be the focus of the 2018 production for the HRLs, such as the creation of a building footprint mask, that could be used as a future "backbone", opening new potential tools ensuring the spatial consistency of the time series, as well as in the second phase of the H2020 ECoLaSS project, where new prototypes and their robustness for operational roll-out will be tested.

REFERENCES

[1] UN Habitat, "Global Report on Urban Health equitable, healthier cities for sustainable development", 2016.

[2] Wilson B. and Chakraborty, A., "The Environmental Impacts of Sprawl: Emergent Themes from the Past Decade of Planning Research. Sustainability", Volume 5, pp. 3302-3327, 2013.

[3] "Urban sprawl in Europe, The ignored challenge", 56 pp, ISBN 92-9167-887-2, 2006.

[4] Moser, L., Probeck, M., Ramminger, G., Sannier, C., Descle B., Schardt, M., Gallaun, H., Deutscher, J., Defourny, P., Blaes X., Klein, I., Keil, M., Hirner, A., and Esch, T. "Sentinel-based Evolution of Copernicus Land Services on Continental and Global Scale", 2017, see: https://www.ecolass.eu/

[5] Carlson, T. N. and Arthur S. T., "The impact of land use land cover changes due to urbanization on surface microclimate and hydrology: a satellite perspective" Global and Planetary Change, Volume 25, pp. 49-65, 2000, doi:https://doi.org/10.1016/S0921-8181(00)00021-7.

[6] Zha, Y., Gao, J., and Ni, S., "Use of Normalized Difference Built-Up Index in Automatically Mapping Urban Areas from TM Imagery." International Journal of Remote Sensing, Volume 24, pp. 583-594, 2003.

[7] Pesaresi, M., Ehrlich, D., Caravaggi, I., Kauffmann, M., and Louvrier, C., "Toward global automatic built-up area recognition using optical VHR imagery." Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, Volume 4, pp. 923-934, 2011.

[8] Lefebvre, A., Sannier, C., and Corpetti, T., "Monitoring Urban Areas with Sentinel-2A Data: Application to the Update of the Copernicus High Resolution Layer Imperviousness Degree", Remote Sensing, Volume 8, pp. 1-21, 2016, doi:https://doi.org/10.3390/rs8070606.

SATELLITES MONITORING DATA INSIGHT ANALYSIS THROUGH WAVELETS-BASED METHODS

C. Ciancarelli *, *A. Intelisano, S.G. Neglia*

Thales Alenia Space, Via Saccomuro 24, Rome, Italy

*Corresponding author

ABSTRACT

The definition of reliable algorithms suitable to extract information from housekeeping telemetry data generated by in-flight spacecraft, requires a deep understanding of the time series, e.g. in terms of composition in time-frequency domain, seasonal and noise components. In this context, the investigation about the (hidden) structure of the telemetries is a fundamental step of data analysis and data-mining algorithms design, also taking into account the satellites system mission operational rules. The present paper aims to explore and test some wavelets-based methods for analyzing irregularly-spaced telemetry data time series. The valuable features of wavelet transform are used to achieve information about data structure, with the objective to provide proper input to advanced processing techniques, for improving automatic prediction on anomaly detection.

Index Terms— Satellites monitoring, time series, wavelets analysis, data analytics, Bayesian networks.

1. INTRODUCTION

Nowadays, in the field of satellites monitoring a large number of parameters are measured and saved in control centers databases to be used for on-board units surveillance and historical analysis by satellite engineers. In the classical approach for equipment status monitoring, observable parameters are simply checked to be inside the "green flags" conditions.

The lesson learnt from last decade in satellites systems operation highlights that more sophisticated monitoring systems are required, with the possibility of being used in real time to early detect failures and abnormal behaviors. A way-to-proceed is provided by big data analytics and data-mining algorithms, as Bayesian networks (BN) and neural networks (NN), which allow to inspect insight relations in large amount of telemetry data time series collected during years of in-flight spacecraft operational life. These processing techniques can be used to develop proper methodologies to support real-time monitoring of in-flight spacecraft, specifically for anomaly detection, particularly useful for monitoring of large constellations. Such advanced algorithms exhibit interesting capabilities for both inferring cause-effect relations among different variables and for data prediction, by evidencing potential anomalous behaviors of parameters that, apparently, are in nominal conditions. It is important to highlight that, due to the volume of data generated by satellites constellations throughout several years of in-flight operations, the above problem is addressed in the frame of the big data analytics paradigm.

In this scenario, an innovative approach to retrieve knowledge from satellite telemetry data has been presented in [1], with the aim to study the predictive capability of BN-based algorithms. Anyhow, the analyses showed a great sensitivity of BN network topology, learning algorithms and their convergence time w.r.t. the selected portions of the dataset. Such behavior appears to be related to the characteristics of time series used in the data-driven modeling—i.e. model identified from a source of data.

Motivated by the results of the previous work [1], the goal of the present paper is to analyze the discrete time series of satellite telemetries through wavelets-based methods. Indeed, data preprocessing is one of the main key points involved with data analytics for time series prediction. Wavelets are an important tool for analyzing time series and provide significant properties, such as signal decomposition, multi-resolution analysis, localization and denoising, in both stationary and non-stationary cases. As stated before the wavelets are used in the data preprocessing step, taking advantage of information in time-frequency domain.

Finally, the paper shows some simulation results related to the processing of real in-flight telemetries by using wavelets-based methods.

2. WAVELETS

From their origin (see Morlet and Daubechies [2]) wavelets evolved in many disciplines, particularly in statistics [3] and time series analysis applications [4]. Mathematically, wavelets transform a signal into a different domain, by using a set of *mother wavelets* functions. There is a variety of mother wavelets [3], such as Haar, Daubechies, Meyer, Morlet, which are chosen depending on the characteristics of data. The Daubechies mother wavelets have been increasingly adopted for digital signal processing.

Wavelets main concepts are herein introduced with reference to the *Haar* mother wavelet. The Haar mother

doi:10.2760/848593

wavelet is a very simple mathematical function (see [3]), but it also exhibits many characteristic features of wavelets. One relevant feature is the capability to oscillate and decay fast.

Once a mother wavelet $\psi(x)$ is selected, the operations of dilation and translation are applied to generate a base, i.e. $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, which is orthonormal. Then, a given function $f(x)$ can be decomposed into the following expansion $f(x) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} d_{j,k}\psi_{j,k}(x)$, where the numbers $d_{j,k}$, for $j,k \in \mathbb{N}$, are the *wavelet coefficients* of $f(x)$. The index j is recalled as the scale, and k as the position.

As the telemetries are sequences of data observations, wavelet analysis of sequences rather than functions is used. Let us consider a discrete sequence of data $y = (y_1, y_2, \dots, y_n)$, where each y_i is a real number and $i = 1, \dots n$ (it is assumed $n = 2^J$, with $J \in \mathbb{N}$). From the vector y it is possible to extract multiscale information, i.e. the representation of data at a set of scales simultaneously. In the finer scale (i.e. high resolution scale, index $j = 1$), the discrete Haar wavelets coefficients are provided by $d_{1,k} = (y_{2k} - y_{2k-1})/\sqrt{2}$, in which the new sequence $d_{1,k}$ encodes the difference between successive pairs of observations in the original vector. Similarly, the sum of consecutive pairs of observations provides $c_{1,k} = (y_{2k} + y_{2k-1})/\sqrt{2}$, which are the *father wavelet coefficients*. In the expressions of $d_{1,k}$ and $c_{1,k}$ the factor $1/\sqrt{2}$ is introduced to ensure output sequence to have the same energy of input vector y. Then, the new sequence $d_{1,k}$ can be processed with the same approach to provide a coarser scale-2 sequence $d_{2,k}$, and so on. This step-by-step procedure is a pyramid algorithm (see Mallat [3]), capable to extract local features of the input vector y at different resolution scales. Such algorithm is one kind of discrete wavelet transform (DWT).

In summary, DWT can break down a sequence into many lower resolution components using a given mother wavelet function. Wavelet decomposition produces sequences that may contain important information about the behavior of the original sequence. The decomposition process may be applied iteratively and the level of decomposition applied to a sequence depends on the specific problem to be tackled. Such concepts have been used in analyzing satellite telemetries.

3. TELEMETRIES DATASET

Real set of telemetries have been used in the simulations, derived from the real satellite telemetry data generated by in-flight satellite system, collected and stored during several years of its operational life time.

Telemetries provide a huge amount of information, typically more than ten thousands of parameters are generated by a spacecraft. Then, the volume of data generated by satellites constellations throughout their lifetime (e.g. sampling at 2 s) is handled in the frame of big data paradigm.

To properly select variables to build-up the dataset, the criteria described in [1] has been adopted. The dataset that has been used for the simulations includes a group of 27 variables, constituting a comprehensive set of parameters associated with the satellite attitude control system and with the operative status of related on-board units. Such telemetries have been deeply inspected and they exhibit the following main remarkable features:

- irregularly-spaced data: the sampling of the telemetries is not regular and some portions of data are missing;
- noise component: observables are contaminated with noise, as in most real-world time series;
- data discontinuity: some time series show sharp variations with respect to the "nominal" behavior, in some cases seems to be associated to satellite maneuvers;
- seasonal components: most of time series are clearly affected by periodic phenomena associated to the cyclical orbital movements of the satellite and its specific mission.

Fig. 1 shows an example of irregularly-spaced telemetry time series ($n = 64$) with several missing samples.

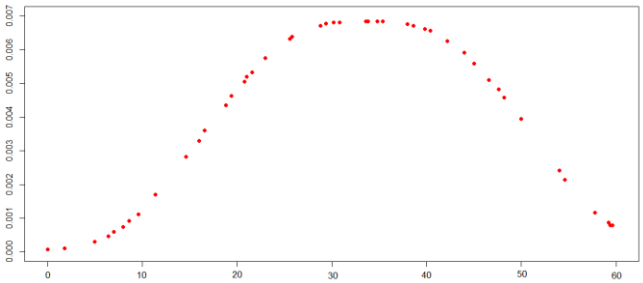


Fig. 1. Irregularly-spaced telemetry time series (example).

Such characteristics, especially the first three ones, have been addressed through the analyses and simulations using wavelets methods, as described afterwards.

4. WAVELETS METHODS AND TOOLS

As stated before, since the sampling of the telemetries is not regular, there is the need to adopt wavelet-based methods suitable to process time series in which samples are not regularly spaced on the time grid.

In order to reduce the noise component (denoising) the wavelet shrinkage approach is applied, which exploits the DWT capability to compress wavelets coefficients in the time-frequency domain. Indeed, the sparsity of wavelets coefficients allows to concentrate the amount of information of original sequence in a reduced number of wavelet coefficients; then, a thresholding is applied to eliminate the coefficients which are mostly carrying noise with negligible information content. The following basic model is adopted

$$y = g + e, \quad (1)$$

doi:10.2760/848593

where the noisy observations y are obtained by the unknown sequence g contaminated by the additive noise sequence e, usually assumed to be white noise.

The preprocessing of the telemetries has been performed using the Kovac & Silverman (KS) wavelet method [3], which possesses good compression and denoising properties for irregularly spaced data. The method adopts the basic model (1), but taking into account the effect of thresholding on the correlation of sequence values.

The idea of the KS method is to take irregularly-spaced and noisy sequence y and interpolate the values y_i, for $i = 1, \ldots n$, to a particular pre-specified regular grid. Then the DWT and the wavelet shrinkage is applied to the interpolated values on the regular grid, with special treatment for the thresholding of the wavelet coefficients, because they are the coefficients of correlated interpolated sequence values, not the assumed independent sequence values themselves. As a result, the wavelets shrinkage allows to estimate sequence g of the basic model (1).

The KS method uses mother wavelets belonging to Daubechies family and implements shrinkage with wavelet decomposition at the finer scale (high resolution). As shown in the results, wavelet decomposition with KS method highlights the local features of the telemetry time series.

The data processing has been implemented through the software environment provided by R tool [8], equipped with the relevant packages implementing the KS wavelet method. The following R main packages have been used: wavethresh, astsa, wmtsa.

Moreover, the same R tools as indicated in [1], implementing BN algorithms, have been used to run a new set of simulations with the dataset preprocessed through DWT. The results are provided in the present paper.

5. WAVELETS PROCESSING RESULTS

The wavelet-based methodology described before has been applied to the telemetries dataset. Fig. 2 shows a sample ($n = 2048$) of an original sequence (red dots) of the telemetries, with the overlapping of the corresponding estimated sequence (blue line) obtained with the processing through KS wavelet method. It is evident the capability of wavelets decomposition, shrinkage and inverse-wavelet transform to estimate the "true" sequence by reducing the noise component.

Fig. 3 shows another sample ($n = 8192$) of an original sequence (red dots) presenting a seasonal component and a sharp variation on the right-hand side of the plot, with rising and descending edges. The estimated sequence (blue line) is obtained through KS wavelet method, evidencing the capability of the method to follow the sharp profile of the original sequence without smoothing effects.

Other interesting results are the wavelet decomposition coefficients of the original sequences, strictly related to local features of the telemetry time series. For instance, Fig. 4 shows the wavelets decomposition coefficients of the sequence previously shown in Fig. 3, ordered from the high resolution level (bottom row) to lower resolution level (top row).

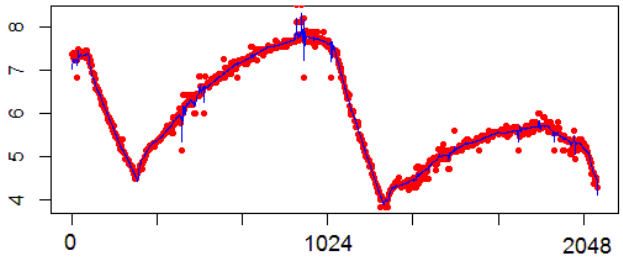


Fig. 2. Example of original sequence (red dots) with the overlapping of the estimated sequence (blue line) obtained through KS method processing.

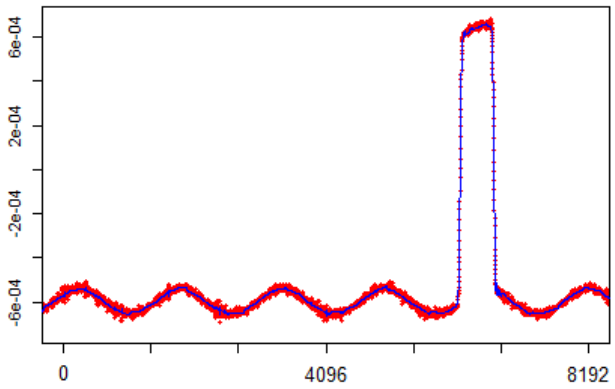


Fig. 3. Example of original sequence (red dots) presenting a sharp variation (on the right side). The estimated sequence (blue line) through KS wavelet method follows the steep profile of the original sequence.

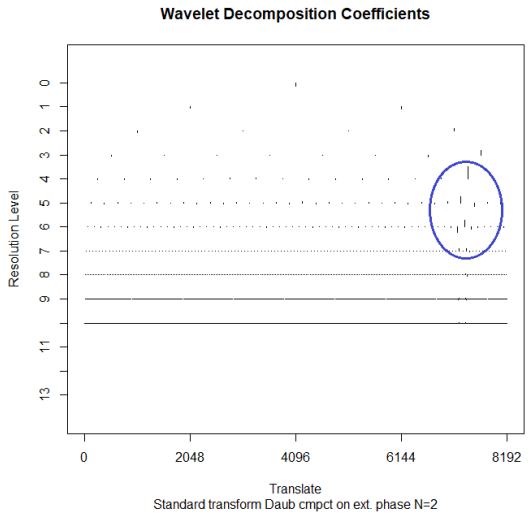


Fig. 4. Wavelet decomposition coefficients of the sequence in Fig. 3. Coefficients in the blue circle are associated to the rising and descending edges visible in the time domain.

doi:10.2760/848593

It is important to note the highest coefficients (see blue circle), which are associated to mother wavelets able to capture detailed information about the local phenomena, in terms of position and amplitude of the coefficients, related to the rising and descending edges in the time domain. On the other hand, after thresholding most of the wavelets coefficients have been removed, as expected, thanks to the sparsity property of wavelet coefficients (see the very small dots in Fig. 4).

Wavelet shrinkage has been applied to all telemetries. It is observed that the percentage of wavelets decomposition coefficients which are removed is more than 90%, confirming the good compression level (i.e. dimension reduction) obtained with the wavelet-based method.

6. BN LEARNING SIMULATIONS RESULTS

As states before, the results obtained in [1] had shown a great sensitivity of BN network topology w.r.t. the selected portions of the dataset. The reasons are related to the intrinsic covariance dependability affecting the Bayesian estimation.

In order to verify how to minimize this effect, a new set of BN learning simulations have been run with the methodology described in [1] (based on [5], [6], [7]), using the dataset preprocessed through the KS wavelet method. The applied wavelet method operates as a nonparametric regression and a time grid regularization on the telemetries. This should reduce the noise effects and the related covariance variability. The results have confirmed that the time dependence on the dataset sample size disappears when using the telemetries preprocessed with DWT, in all parameters related to each BN net type. For example, Table 1 represents the result of a BN learning through a Hill-Climbing learning algorithm. The comparison is performed among four data sets: original dataset with the nominal sample size of the telemetry data @8Hz, the same data set after DWT @8Hz, the two datasets sampled @1Hz. While the original dataset is strongly affected by the sampling time choice, the one processed with WDT shows no dependence on the sample size, as expected.

Table 1. Comparison of BN learning (Hill-Climbing algorithm) using the original dataset and the one obtained after DWT processing.

	Full WDT data set	Reduced WDT data set	Full original data set	Reduced original data set
nodes	27	27	27	27
learning algorithm	Hill-Climbing	Hill-Climbing	Hill-Climbing	Hill-Climbing
directed arcs	183	183	168	130
average markov blanket size	21.48	21.48	21.26	18.74
average neighbourhood size	13.56	13.56	12.44	9.63
average branching factor	6.78	6.78	6.22	4.81
penalization coefficient	4.258597	4.25859	4.258597	1.198948
tests used in the learning procedure	5928	5928	5434	4186

Anyhow, the covariance variability, even if reduced among the different BN learning algorithms, remains too wide to allow the definition of the best BN. It appears that BN definition can be reasonably efficient when a-priori knowledge about causal relation among the variables is used. In other words, the satellite designers and experts knowing how variables interact each other, can be still the most efficient architects for an efficient BN definition.

7. CONCLUSIONS AND FUTURE WORKS

The paper focused on wavelets-based methods for analyzing the irregularly-spaced real telemetry time series generated from in-flight satellites system. The proposed data preprocessing approach provides the means to analyze the insight structure of the telemetries, extracting the multiscale information and highlighting the local features. Compression level and noise reduction capability have been also positively tested. Therefore, the preprocessed telemetries seems to be a suitable input for data analytics methodologies for anomaly detection.

In the case of data-mining based on BN algorithms, the main results have shown how data preprocessing through the KS wavelet method can solve the main problem, i.e. the sensitivity of BN network topology w.r.t. the selected portions of the dataset used for learning. Anyhow, the definition of the best BN is still open, leading to the need to a different approach. In principle, automatic definition of an efficient BN for anomaly detection may be explored through the use of techniques based on neural networks, in order to discriminate the most important variables to be treated and correlated, reducing the state space and the computation time.

8. REFERENCES

[1] C. Ciancarelli, A. Intelisano, S.G. Neglia, *Knowledge Retrieval Strategy for Satellites System Monitoring based on Data Analytics Techniques*, in Proc. of the 2017 conference on Big Data from Space (BiBS'17), Toulouse (France), pp. 394-397. doi: 10.2760/383579.

[2] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[3] G.P. Nason, *Wavelet Methods in Statistics with R*, Springer, 2006.

[4] D.B. Percival, A.T. Walden, *Wavelet Methods for Time Series Analysis*, Cambridge University Press, 2000.

[5] Uffe B. Kjærulff, Anders L. Madsen, *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*, Springer, 2013.

[6] B. De Finetti, *Theory of Probability*, John Wiley & Sons, 1990.

[7] M. Scutari, J.B. Denis, *Bayesian Networks – With Examples in R*, CRC Press, NW, 2015.

[8] R Core Team (2018), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

ROBUST AIRPLANE DETECTOR FOR MULTI-SENSOR SATELLITE IMAGES

Romain Hugues, Amandine Pailloux, Michelle Aubrun, Marc Spigai, Etienne Barritault, Alexandre Scotto di Perrotolo, Alric Gaurier

Thales Alenia Space, 26 av. J.F. Champollion, Toulouse, France
romain.hugues@thalesaleniaspace.com, marc.spigai@thalesaleniaspace.com

ABSTRACT

In the domain of human operators assistance for localizing objects of interest on remote sensing images, one problematic is the robustness and usability of models relative to image resolution. In this research, we intend to start from a model estimated with Deep Learning (DL) technologies to detect planes with, as input, a big database of High Resolution (HR) images representing planes. Given this model estimated with the HR database, the research studies the extension of the detection to the case of a small database of large planes in ESA Sentinel-2 images at 10 meters resolution. Single Shot MultiBox Detector (SSD) neural networks architecture has been used after comparison with You Only Look Once (YOLO). The methodology used for extending HR data and model to data of lower resolution so far have been very encouraging and show a real potential in terms of model engineering robustness and usability.

Index Terms— Object Detection, Robustness, Resolution, Deep Neural Network, Single Shot MultiBox Detector (SSD), You Only Look Once (YOLO), Sentinel-2 Data, Satellite Images.

1. INTRODUCTION

Human operators assistance for object detection in remote sensing is a wide subject of research. If one can find some researches on models for many types of objects detection, the problematic of models robustness and usability relative to image resolution if often not taken into account. We focus in this research on plane detection with Deep learning models given a big database of sub-metric High Resolution (HR) images representing planes and looking how it can be applied to Sentinel-2 images at 10 meters resolution.

Aircraft detection and recognition in remote sensing images is carried out by trained humans because this operation is crucial for military domain. Although this subject has been studied for many years, no efficient algorithm was able to challenge human operators due to the variety of aircrafts (e.g. shape, size and color) and the complexity of the background.

Li et al. [1] propose an airplane detection approach based on visual saliency computation and symmetry detection. Liu et al. [2] applied a coarse-to-fine process integrating the high-level information of a shape prior to detect aircraft. These methods have the advantage not to require computational power, but rely on manually thresholds and assumptions.

These last years, deep learning based on convolutional neural network (CNN) became a very popular method in computer vision because of its ability to learn intrinsic features and the great performance resulting from this method. In image processing, there are three main CNN approaches to detect objects:

- Give the image and its ground truth per pixel to a CNN, which provides a classification per pixel of the image. This approach is, above all, used for not complex and numerous objects. Sherrah [3] applied successfully this method to detect buildings, cars, herbaceous vegetation and shrub vegetation.
- Use a sliding window to navigate in the image and apply a CNN to classify each window. Chen et al. [4] generate two scale sliding windows and sent them to Deep Belief Nets for feature extracting and classification. This approach can detect tiny blurred aircrafts correctly, but is time-consuming and not appropriate for objects of varying size. In order to remedy these drawbacks, Wu et al. [5] use BInarized Normed Gradients (BING) technique to generate a set of candidate object windows.
- Give the image as input to a CNN that detects and classifies objects. Among these algorithms, some use a single neural network to predict bounding boxes and classify them, which allows them to be really fast approaches. For instance, Radovic et al. [6] successfully test a trained "YOLO" network in real-time video.

In this paper, YOLO and SSD have been used to detect aircrafts in images. The SSD combines predictions from multiple feature maps with different resolutions to handle objects of various sizes. Thus, it is proven to be more accurate than YOLO, although slightly slower.

The rest of the paper is organized as follows. In section 2, we describe YOLO and SSD architectures. Then data and results are presented in section 3. Conclusion is Section 4.

2. METHODS

2.1. Networks Architecture

2.1.1. YOLO

YOLO is an open-source object detection and classification algorithm available under the "You Only Look Once" project [7]. We have chosen the version 2 of YOLO which improves the accuracy and speed of the YOLOv1 using batch normalization, higher resolution, initialization of the

doi:10.2760/848593

anchor boxes with the data set boxes, better detection of small objects using a pass-through approach, multi-scale training.

The principle of YOLO is to divide the image in cells, it gives an equally size grid (S x S) . The number of cells or the size of the cells is fixed by the user. Each cell is affiliated to several bounding boxes (in our case, it is 5 bounding boxes), which are described by five parameters:

- the coordinates of the box's center (x, y)
- the width and the height of the box (w, h)
- a confidence score. This score evaluates the probability that the bounding box encloses some object, but provides no statement on the type of object.

And for each bounding box predicted, a probability distribution function is applied over all the classes trained in the model (in our case, there is one class: plane) in order to predict the type of object. The combination of this probability with the confidence score provided a final score, which reflects the probability that the bounding box contains a specific type of object. After a thresholding, only the best scores are kept and then a non-maximum suppression step is applied to remove bounding boxes with high IoU (Intersection over Union).

YOLO is composed of 24 convolutional layers and 2 fully connected layers (see Fig 1). The end of the network is a tensor that contains the bounding boxes coordinates (x,y,w,h), the bounding boxes confidence and the class probabilities for each grid cell. The tensor size is S ×S ×(B ×5+C) where S is the number of cells in one direction, B the number of bounding boxes predicted for one cell and C the class probabilities. In order to detect objects a specific loss function is used, you can find this function in the YOLO paper [9]. As YOLO architecture is huge (millions of parameters), we have tried to change the initial architecture in removing some layers in order to have less parameters to tune during the learning phase. However, the best results have been reached with the full architecture.

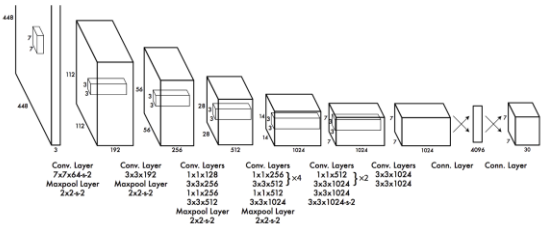


Fig 1 : Yolo architecture

2.1.2. SSD

The SSD [10] is also a fast approach based on a single convolutional neural network that produces a fixed-size set of bounding boxes with their class probabilities. SSD is similar to YOLO in the way to detect objects but SSD is composed of a standard architecture (VGG 16 for example) and an additional structure to detect the objects. This additional structure provides multi-scale feature maps thanks to convolutional layers that decrease in size and allow predictions of detections at multiple scale, unlike YOLO (Fig 2). We have used an open source algorithm to implement SSD in Keras [11].

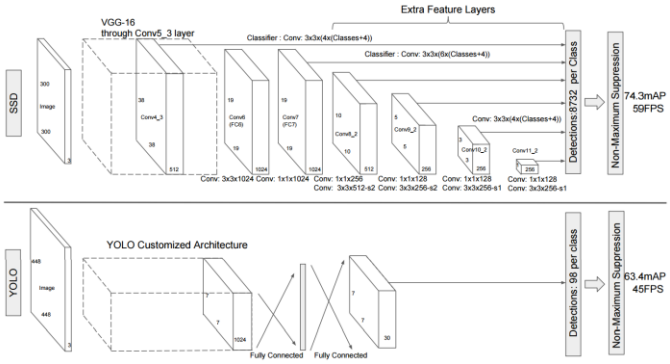


Fig 2 : SSD vs YOLO : SSD model adds feature layers to the end of a standard architecture (VGG-16)

Using feature maps from different layers in a single network allows to handle different object scales. Feature maps from the lower layers can improve semantic segmentation quality (capture more fine details of the input objects) whereas feature maps from the higher layers add more global context and help smooth the segmentation results.

Each feature map is divided into m x n cells. At each cell an offset is predicted relative to default bounding box shapes in this cell as well as the per-class scores. So for each cell, we have k boxes with their c class scores and the 4 offsets relative to the original default box shape. This leads to m x n x k (c + 4) outputs for a feature map of size m x n.

3. DATA, METRICS, RESULTS

3.1. Data

Concerning the High Resolution database, Thales Alenia Space has its own internal database of sub-meter panchromatic images. This database is made up of about 70 airports situated in very different areas all over the world and almost 10 000 planes with different sizes, shapes and colors. It is divided in training and validation data sets which contain respectively about 8 000 and 2 000 samples.

For the Sentinel-2 data (10 meters resolution) the database represents is made up of 20 big airports spread across the world and almost 800 large (width and length above 50 meters) civilian airliners. In this research these data have been only used for validation test phases, the integration in learning phase has not been integrated for the moment due to the small number of examples compared to the High Resolution database.

doi:10.2760/848593

3.2. Metrics

The usual following metrics are used in order to evaluate the results for both algorithms : ROC curve, Intersection over Union (accuracy of the bounding boxes) . Moreover the learning phase has been checked with loss curves to avoid over-fitting. For simplicity reasons we used in the article only ROC curve.

ROC curve: On the x-axis, the number of objects wrongly detected divided by the total number of objects. On the y-axis, the number of objects correctly detected divided by the total number of objects. Note that the curve depends on the number of objects and not the number of pixels in order to obtain a curve easier to understand.

3.3. Selection of a model on HR data

Both algorithms are provided by open source platforms and ready to be used but we have adapted these algorithms for our own purpose and retrained the networks.
First, the influence of hyper-parameters (e.g. learning rate, number of cells, number of boxes generated per cell, …) has been checked. However, we have found that initial hyper-parameters were already well fitted for our application. The other hyper-parameters regarding the scale and aspect ratios of boxes are automatically computed according to the training database used. In order to compare relevant data, each tile has been normalized what did not improve the accuracy but reduced significantly the number of false detection.
In this study, patches with 512x512 pixel size have been used to test the networks. To create the patches, an overlap of 15% has been applied to take in consideration objects that are split between several tiles. The initial weights have been set by the weights already trained on ImageNet.
After several tests, we have found that SSD provides better results than YOLO (see Fig 3), that is why we will focus on SSD results in the following parts concerning the extension to Sentinel-2 data.

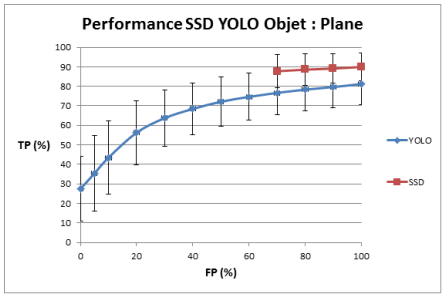


Fig 3 : ROC Curve SSD versus Yolo for plane detection in the HR case

3.4. Extension of learning to S2 data

In order to extend the abilities of the SSD to detect airplanes on images with lower resolution, modifications in the training process have been introduced. First, all the HR satellite images have been divided into tiles of 6 different sizes and resized to 512x512 so as to degrade their resolution. The resulting data sets have resolution ranging from 0.3 to 2.5 meters (0.3, 0.7, 1.0, 1.5, 2.0, 2.5). The training set is then made up of those tiles ensuring around 3000 objects for each of the 6 resolutions. In the end, a Gaussian blur with random sigma ranging from 0 to 4 is added to the input images before entering the network.

3.5. Qualitative Validation Tests

We show here visual results of the detection of planes on Sentinel-2 images. Ground truth is in blue as detection is In red. Fig 4 shows a detection in a “good” case where all the planes are detected in the image. Fig 5 Shows a “bad” case where not all the planes have been detected. To see the limits of the approach we have applied the model to a the smaller French airport Blagnac of Toulouse, Fig 6 (no ground-truth available) shows that the algorithm seems to be promising even for smaller planes. In the same way we have applied the models to urban areas where there are no planes and the number of false alarms is relatively small implying the detector tries to find shape of planes more than white marks.
In a general qualitative point of view the first results of the methodology are very promising.

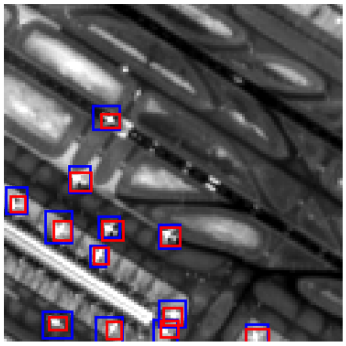


Fig 4 : “Good” case of detection on Sentinel-2

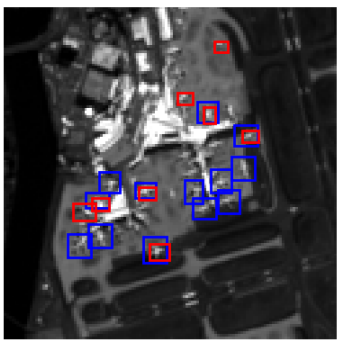


Fig 5 : “Bad” case of detection on Sentinel-2

doi:10.2760/848593

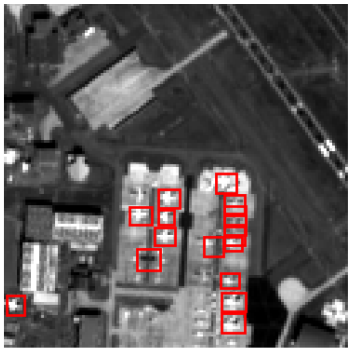


Fig 6 : Test on smaller airport: Toulouse, France -2

3.6. Quantitative Validation Tests

To complete the qualitative analysis, a ROC curve is shown on Fig 7. As preliminary results of this methodology and taking into account the fact that the ground truth is not perfect, the results are acceptable allowing for instance to have a good estimation of an "heat map" representing the density of presence of large planes in airports with Sentinel-2 images

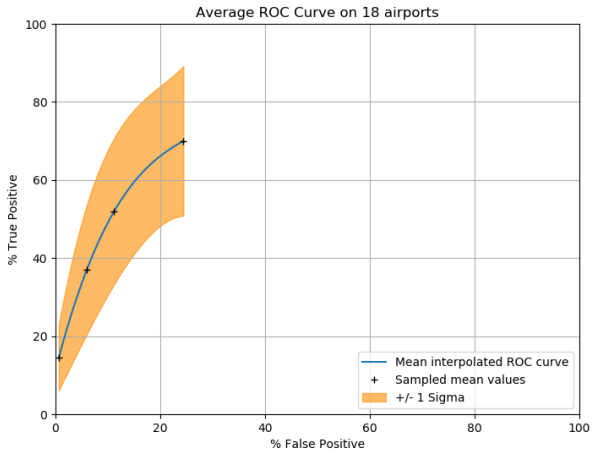


Fig 7 : ROC curve of plane detection on Sentinel-2 images

4. CONCLUSION

The initial objective was: starting from a model estimated with Deep Learning (DL) technologies to detect planes with a big database of High Resolution (sub-meter) images how can this be used to extrapolate to the case of a small database of large planes with ESA Sentinel-2 images at 10 meters resolution. As a summary: an SSD architecture has been selected in HR case and a learning has been processed modifying the HR database to fit Sentinel-2 case. Results so far have been very encouraging and show a real potential in terms of model engineering and robustness of models. Detection performances can potentially be enhanced by enlarging the size of Sentinel-2 database and mixing it with the High Resolution database during learning phase. At present the results allow to have a good estimation of an "heat map" representing the density of presence of large planes in airports with Sentinel-2 images.

REFERENCES

[1] W. Li, S. Xiang, H. Wang, and C. Pan, "Robust airplane detection in satellite images,", in ICIP, 2011, pp. 2821-2824.

[2] G. Liu, X. Sun, K. Fu, and H. Wang, "Aircraft recognition in high-resolution satellite images using coarse-to-fine shape prior," GRSS, vol. 10, no. 3, pp. 573-577, 2013.

[3] Sherrah, J. Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery. 2016.

https://arxiv.org/pdf/1606.02585.pdf (accessed on 8 June 2017).

[4] X. Chen, S. Xiang, C.L. Liu, and C.H. Pan, "Aircraft detection by deep belief nets," in ACPR, 2013, pp. 54-58.

[5] Wu H., Zhang H., Zhang J. Fast aircraft detection in satellite images based on convolutional neural networks; Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP); Quebec City, QC, Canada. 27–30 September 2015.

[6] Radovic M., Adarkwa O., Wang Q. Object recognition in aerial images using convolutional neural networks. J. Imaging. 2017;3:21 doi: 10.3390/jimaging3020021.

[7] You Only Look Once: Unified, Real-Time Object Detection, https://pjreddie.com/darknet/yolo/.

[8] YOLO9000: Better, Faster, Stronger, Joseph Redmon, Ali Farhadi, University of Washington, Allen Institute for AI, https://arxiv.org/pdf/1612.08242.pdf.

[9] You Only Look Once: Unified, Real-Time Object Detection. Joseph Redmon, University of Washington, Santosh Divvala, Allen Institute for Artificial Intelligence, Ross Girshick, Facebook AI Research, Ali Farhadi, University of Washington. https://pjreddie.com/media/files/papers/yolo.pdf

[10] SSD: Single Shot MultiBox Detector. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg.
https://arxiv.org/pdf/1512.02325v5.pdf

[11] A Keras port of Single Shot MultiBox Detector https://github.com/pierluigiferrari/ssd_keras

SATELLITE IMAGE COMPRESSION BASED ON HIGH EFFICIENCY VIDEO CODING STANDARD - AN EXPERIMENTAL COMPARISON WITH JPEG 2000

Miloš Radosavljević [1], *Marko Adamović* [1], *Branko Brkljač* [1], *Željen Trpovski* [1]
Zixiang Xiong [2], *Dejan Vukobratović* [1]

[1]Dept. of Power, Electronics and Comm. Eng., University of Novi Sad, Faculty of Technical Sciences, 21000 Novi Sad, Serbia; [2]Dept. of ECE, Texas A&M University, College Station, TX 77843, USA

{ milos.r, adamovicm, brkljacb, zeljen, dejanv } @uns.ac.rs [1] ; zx@ece.tamu.edu [2]

ABSTRACT

Driven by the rapid growth in the volume of the satellite data, in this work we propose initial effort to explore lossy compression techniques for the satellite data applications. The effectiveness of the compression of satellite images, that typically have 16 b/p resolution and the high quality requrements, has been analyzed on the HEVC vs JPEG 2000 coding performance. This work studies the application of a 16 b/p extension of the HEVC codec to satellite images. Compared to a widely used JPEG 2000 codec, that is currently exploited by the Sentinel-2 image compression, new HEVC based codec significantly improves the PSNR vs. compression ratio performance. Current efforts are focused on the performance improvement by using different configurations and features, reduction in the computational complexity, and exploration on the feasibility of using lossy techniques in the satellite image applications.

Index Terms— HEVC, JPEG 2000, high bit-depth compression, multispectral satellite images, Landsat-8, Sentinel-2

1. INTRODUCTION

Satellites for multispectral imaging of the Earth, such as American Landsat-8 and European Sentinel-2, make huge amounts of high resolution and high bit-depth images on a daily basis. Typically, multispectral images use 16 bits-per-pixel (b/p) to cover wide satellite depth range. Emergence of new imaging technologies, and the big data volumes they produce, considerably increases storage and I/O speed requirements, which results in the higher equipment cost. Apparent need for the efficient compression scheme to reduce storage requirements and processing time is obvious.

Landsat-8 saves images in uncompressed format, while Sentinel-2 uses JPEG 2000 still image compression. With the completion of the High Efficiency Video Coding (HEVC) standard in 2013 [1], and the publication of its second version that supports range extension up to 16 b/p video, [2], one can obtain much better video coding performance than using previous standards such as H.264/AVC or JPEG 2000 [3, 4]. In addition, the main still image profile of HEVC, i.e., HEVC intra coding can be used for efficient still image compression as well.

HEVC coding employs rate-distortion optimized quadtree-based variable block size partition of the image, angular intra prediction or motion driven inter prediction, followed by integer approximation of the DCT (of size 4x4, 8x8, 16x16, or 32x32), uniform quantization, and context-based adaptive binary arithmetic coding (CABAC). It was reported that HEVC intra coding (still image) reduced the average bit rate by 15.8% compared to H.264/AVC, 22.6% compared to JPEG-2000, and 30.0% compared to JPEG-XR, and 43.0% compared to classic JPEG. The margin is even higher using the inter predictive coding. On the other side, JPEG 2000 is still image compression standard [4, 5]. An intra-frame compression scheme that encodes each frame independently has been exploited. Although motion prediction version is provided in later parts of the standard, JPEG 2000 as a still image compression has been used extensively in Sentinel-2 program as a compression scheme. It aims at low-complexity compression of high dynamic range images, and it has been used by industry widely. Compared to the HEVC that is block based, JPEG 2000 uses wavelet decomposition at different scales. It uses simple predictive scheme in order to make dynamic range that is approximately centered around zero (known as DC Level Shifting), discrete wavelet transform, uniform scalar quantization or trellis coded quantization, and arithmetic coding engine to efficiently compress the quantized coefficients in the final bit-stream. One of the main advantages of JPEG 2000 is the high flexibility of the bit-stream formation, giving the possibility to decode image, or just a region of interest, in a variety of ways.

This paper aims to analyze HEVC and JPEG 2000 standards for the use in the compression of multispectral satellite images. More exactly, we propose a novel approach for compression of multispectral images based on HEVC, which can be considered as an initial work towards a flexible and effec-

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, as part of the project III44003. The third author would also like to acknowledge the ERA.Net HARMONIC project.

tive solution that will meet high quality, high resolution, high bit-depth requirements of the satellite image applications. For this purpose, in order to enable reduction of storage and processing requirements of such huge volumes of data, we have analyzed different types of competing techniques in a lossy compression scenario. The set of presented, carefully designed experimental settings provides more insight into performance of different rivaling compression approaches and represents the foundation for further research and exploration of the same subject that will be focused on the application oriented effectiveness of the lossy compression techniques, and their use in particular satellite imaging applications. To the best of our knowledge, this work is the only one that exploits a HEVC codec for this purpose.

The rest of the paper is organized as follows. In Section 2 we describe satellite data sources and their characteristics. Codec description, and experimental setup has been given in the Section 3. Results are analyzed in Section 4, where we engaged in a discussion of the potential feasibility of using lossy techniques for the purpose of satellite image compression. The paper is concluded in Sec. 4, with a reference to the future work.

2. MULTISPECTRAL SATELLITE IMAGES

Satellite imaging provides continuous, large scale observations of Earth and its systems. Although in the past it was limited to a closed community of experts and dedicated professionals, this field has witnessed significant democratization and change in data access policies in recent years. It is exhibiting the same type of information proliferation that is present in other technology sectors that are affected by the big data paradigm, and therefore requires new approaches for more effective data management. One of the aspects of particular interest is the question of effective data compression, which is gaining much more significance with the introduction of modern sensors with improved imaging capabilities. This trend is primarily driven by the higher spatial resolution of the modern spaceborn multispectral imagers, but also with the higher revisit times of the corresponding earth observation missions. In addition, multispectral measurements of the scene's reflectance are highly redundant, which makes them particularly suitable for efficient data compression.

For the purpose of this study we will put the emphasis of our short exposition primarily on the characteristics of the two flagship optical land observation missions that are providing high quality multispectral observations of the world, and which are independently operated by USGS/NASA and ESA. Landsat-8 (L8) [6], represents the most recent extension of the long-term Landsat programme that has provided multispectral observation of Earth's surface for more than forty-five years. It acquires images by using the Operational Light Imager (OLI) instrument [7], which produces eight narrowband spectral channels with 30 m spatial resolution in the visible, NIR and SWIR domain, and a single panchromatic channel with wider spectral response and the spatial resolution of 15 m. In addition, there is also a thermal infrared sensor, however, for the time being, we will limit our considerations to the optical domain. These calibrated measurements are provided at several levels of processing quality, accompanied by additional quality assessment band and the corresponding metadata [8]. In the case of Level-1 image products that were used as a source of uncompressed image data in this study, top-of-atmosphere (TOA) reflectances are stored as 16 bit digital numbers (DN) in georeferenced, Geographic Tagged Image File Format (GeoTIFF). Although compression gain would be the highest in the case of panchromatic channel, which possesses the highest level of spatial details, presented experiments were performed by using only narrowband L8 channels in order to ease experimental comparison and overall computational effort.

In comparison to L8, Sentinel-2 (S2) mission [9], provides data in JPEG 2000 format [10], with support for JPEG 2000 Interactive Protocol (JPIP), where Level-1B and Level-1C TOA products are usually provided by default in the lossless mode. However, the product specification as an option also identifies a lossy compression mode that should have a negligible effect on the image quality. The each of the two satellites in S2 constellation acquires images using Multi-Spectral Instrument (MSI) that is capable of capturing images at medium-high spatial resolution in thirteen spectral bands, mostly in VNIR spectral range, out of which four channels are at 10 m, six at 20 m, and three at 60 m spatial resolution. Since S2 bands are acquired at different spatial resolutions they would be very suitable for the analysis of the compression of images of the same scene that are captured at different spatial scales. However, since MSI bands of S2 mostly have higher spatial resolution than the OLI bands of L8 (larger number of pixels), and since in the case of L8 there are 8 spectral bands of the same spatial resolution, while in the case of S2 there are only up to 6 channels with the same spatial resolution, we have decided to limit our analysis and experimental comparison of HEVC and JPEG 2000 performance only to scenarios with L8 data. This decision was mostly driven by the nature of the presented research that should provide initial insights into expected performance and could be justified by the smaller computational burden of the conducted experiments in the case of L8.

3. EXPERIMENTAL SETUP

Compression methods and standards developed for the common video or image compression, may be applied to the compression of satellite data. Although they have been developed mainly to be used for standard images/video, they have been widely used in many applications such as medical diagnostic imaging, seismic images, remote sensing, etc. [11, 12, 13, 14], due to their widely available resources such as open codes, regular version improvements and extensions, and hardware architecture designs. In order to facilitate

doi:10.2760/848593

the practical implementation of the HEVC and JPEG 2000, HM Test Model and OpenJpeg has been exploited to conduct experiments in our work [15, 16]. We have used HM-15.0+RExt-8.1 and version 2.3.0 of the HM and OpenJpeg, respectively.

As already discussed, L8 data has been used in this work as a test image. More specifically, a region given with *LC08_L1TP_186028_20170414_20170501_01_T1* has been used to conduct the experiments and represent the results. Although it would be more reliable to use different time and spatial representations of the satellite data (e.g. different regions with characteristic soil type, or a region during the year with different vegetation state), to due to page limitation and computational expense of the HM software, we have decided to limit the experiments to use of just one image tile. Also, all optical channels except band-8 (B8), and quality assessment band (BQA) have been used in the compression tests, as discussed in previous section Sec. 2. The given image has the resolution of 7801x7911 pixels.

As a performance measure we have use the PSNR vs compression ratio plots. Compression ratio has been calculated as follows:

$$compression\_ratio = \frac{\sum uncompressed\_band\_size}{\sum compressed\_band\_size} \quad (1)$$

Note, since L8 tile consists of the GeoTIFF images, in order to get uncompressed size of a raw image, GeoTIFF header should be omitted from calculations.

In both reference software, we have use default settings provided within the code. The parameters important for this study have been briefly summarized below.

OpenJpeg parameter set: The parameter that controls quality of the reconstructed image has been varied in the range of 34-134 dB. That gave us the wide range of different qualities and compression ratios. Note here that PSNR control is not exact, hence after the reconstruction precise PSNR value should be calculated based on the difference from the original uncompressed image. Other configuration parameters are set as default.

HM parameter set: In order to explore the effectiveness of using the different parameters set in HM (due to a large number of various features supported in HEVC), we have used three type of configurations in the proposed experiments: i) Intra setup, ii) Inter P setup, and iii) Inter B setup. Intra setup has been based on the *encoder_intra_main_rext* parameter set given within the HM solution. It uses only intra predicted I blocks, and may be considered as a still image compression mode. Inter P and Inter B setups have been based on the *encoder_lowdelay_main_rext* configuration setup of the HM. Those configuration exploits temporal redundancy in the data using motion driven prediction (P and B frames). In this case, we have examined the satellite data (bands within tile) as a sequence of a frames, and thus we have been able to additionally reduce redundancy in the data. Usually, satellite applications require a low coding delay, hence all pictures were coded in a display order. Only the first image is coded as an intra image and all subsequent pictures are temporally predicted only from reference pictures in the past in display order (P setup), or using bi-directional temporal prediction (B setup). Since HM implementation has not provided with the PSNR control scheme, as a quality control parameter we have used quantization parameter Q_p. In HEVC, this parameter goes from 0 to 51. However, due to the extended bit-depth this parameter may vary in the range of -48 to 51 which is not strictly defined in standard. Also, the configuration parameter that specifies the use of the extended bit-depth range must be enabled in the HM in order to use 16 b/p compression.

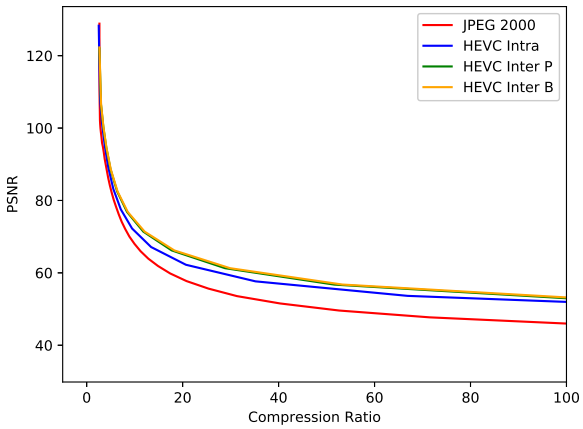


Fig. 1: Performance comparison of new predictive HEVC codec under different configurations vs. JPEG 2000 still image coding scheme in terms of PSNR vs. compression ratio.

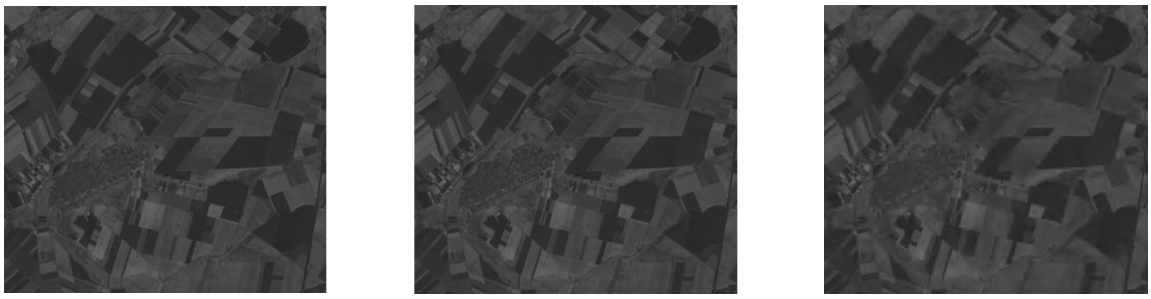
4. RESULTS AND CONCLUSIONS

We first show the performance comparison in terms of the PSNR vs compression ratio between JPEG 2000 and different HEVC setups. On Fig. 1, at low ratios JPEG 2000 and all HEVC setups have almost the same performance. The performance increases in the favor of HEVC on the higher ratios. Also, the difference in the performance between intra and inter HEVC setups has not shown significant gain difference, as it was expected. The reason for this may lie in the fact that, although different spectral bands capture exactly the same area of the land, the measurements they provide may be quite different. The usage of the inter predictive coding has shown performance increase in comparison to the intra coding that exploits only the spatial redundancy, Fig. 1.

Fig. 2 illustrates that even at the same level of PSNR, HEVC can achieve higher compression quality, as perceived by the human interpreter, which is mostly visible in difference between Fig. 2c and Fig 2f. Hence, HEVC's subjective performance is rated as highly satisfactory. This is mainly justified by the use of deblocking and in-loop filters in HEVC. Complexity comparison has not been in the scope of this work. However, it is known that HEVC posses the higher order of the computational complexity. Still, HEVC has

doi:10.2760/848593

(a) HEVC, 100 [dB] (b) HEVC, 60 [dB] (c) HEVC, 40 [dB]



(d) JPEG 2000, 100 [dB] (e) JPEG 2000, 60 [dB] (f) JPEG 2000, 40 [dB]

Fig. 2: Visual comparison of HEVC intra setup and JPEG 2000 image compression with the same PSNR [dB], part of a L8 satellite scene (path 186, row 028) captured by the OLI Near Infrared (NIR) band.

been designed with a care of the increased use of parallel processing architectures and efficient hardware implementation. Therefore, the practical use of the HEVC due to its higher complexity should not be an obstacle. To the best of our knowledge, there is no study focusing on the similar approach as the HEVC inter prediction setup that was presented, which combines image bands and group them for further compression. Notice that multispectral satellite measurements are usually provided as uncompressed data in order to preserve as much of original information. However, many users do not require such high precision of data representation, and depending on particular remote sensing application they are ready to accept some level of information loss. In such case, the best approach is usually to define specific application requirements that will drive the data compression strategy and determine the acceptable loss levels. However, this is not always possible without extensive research aimed towards specific application. Therefore, as the first step in such direction, results presented in this study encourage further research of HEVC based compression of satellite images.

REFERENCES

[1] G. J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand *et al.*, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. on Circ. and Syst. for Video Tech.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] D. Flynn, D. Marpe, M. Naccari *et al.*, "Overview of the range extensions for the HEVC standard: Tools, profiles, and performance," *IEEE Trans. on Circ. and Syst. for Video Tech.*, vol. 26, no. 1, pp. 4–19, 2016.

[3] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H. 264/AVC video coding standard," *IEEE Trans. on Circ. and Syst. for Video Tech.*, vol. 13, no. 7, pp. 560–576, 2003.

[4] D. Taubman and M. Marcellin, *JPEG2000 image compression fundamentals, standards and practice: image compression fundamentals, standards and practice*. Springer, 2012.

[5] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal processing magazine*, vol. 18, no. 5, pp. 36–58, 2001.

[6] D. P. Roy, M. Wulder, T. R. Loveland *et al.*, "Landsat-8: Science and product vision for terrestrial global change research," *Remote Sensing of Environment*, vol. 145, pp. 154–172, 2014.

[7] J. A. Barsi, K. Lee, G. Kvaran, B. L. Markham, and J. A. Pedelty, "The spectral response of the Landsat-8 operational land imager," *Remote Sensing*, vol. 6, no. 10, pp. 10 232–10 251, 2014.

[8] K. Zanter, *Landsat 8 (L8) Data Users Handbook,* Version 2.0.

[9] F. Gascon, C. Bouzinac, O. Thépaut, M. Jung *et al.*, "Copernicus Sentinel-2A calibration and products validation status," *Remote Sensing*, vol. 9, no. 6, p. 584, 2017.

[10] A. Gatti and A. Bertolini, *Sentinel-2 products specification document,* Issue 14.5, 2018.

[11] S.-G. Miaou *et al.*, "A lossless compression method for medical image sequences using JPEG-LS and interframe coding," *IEEE Trans. Inf. Techn. in Biom.*, vol. 13, no. 5, pp. 818–821, 2009.

[12] M. Razaak, M. G. Martini, and K. Savino, "A study on quality assessment for medical ultrasound video compressed via HEVC," *IEEE Jour. of Biom. and Health Inf.*, vol. 18, no. 5, pp. 1552–1559, 2014.

[13] M. Radosavljević, Z. Xiong, L. Lu, and D. Vukobratović, "High bit-depth image compression with application to seismic data," in *Visual Comm. and Image Processing (VCIP), 2016*. IEEE, 2016, pp. 1–4.

[14] M. Radosavljević, Z. Xiong, L. Lu, D. Hohl, and D. Vukobratović, "HEVC-based compression of high bit-depth 3D seismic data," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4028–4032.

[15] "HEVC reference software HM," *Available Online: https://hevc.hhi.fraunhofer.de/svn/svn*.

[16] "JPEG 2000 reference software OpenJPEG," *Available Online: http://www.openjpeg.org/*.

DIMENSIONALITY REDUCTION OF OPTICAL DATA: APPLICATION TO TOTAL OZONE COLUMN RETRIEVAL

Ana del Águila, Víctor Molina García, Dmitry S. Efremenko

Remote Sensing Technology Institute, German Aerospace Center (DLR),
Oberpfaffenhofen, Germany

ABSTRACT

The new generation of atmospheric composition sensors such as TROPOMI, deliver a great amount of data, which is recognized as Big Data. To process the challenging data volumes of spectral information, fast radiative transfer models (RTMs) are required. However, the bottleneck in remote sensing retrieval problems is the computation of the radiative transfer. Thus, the operational processing of remote sensing data requires high-performance RTMs for simulating spectral radiances (level-1 data). In particular, ozone total column retrieval algorithms use the level-1 data in the Huggins band (325-335 nm). Hence, accurate simulation of this absorption band may require several hundreds of monochromatic computations. However, hyper-spectral input data for RTMs has a redundant information, which can be excluded by using the dimensionality reduction techniques. In addition, there is a strong correlation between the input optical data for RTMs and output radiances. Such statistical dependency can be taken into account for accelerating level-1 data simulations using principal component analysis (PCA), thereby providing the performance enhancement for the whole processing chain. In this paper we analyze the efficiency and potential benefits of the optical data dimensionality reduction schemes for simulating the Huggins band and discuss several modifications of this approach.

Index Terms— PCA, data-driven algorithms, trace gas retrieval

1. INTRODUCTION

Atmospheric chemistry observations have received much attention in the last decades because they can improve our understanding of environmental issues such as climate change or stratospheric ozone depletion. In line with that objective, atmospheric composition sensors such as the TROPospheric Ozone Monitoring Instrument (TROPOMI), continuously measure the reflected spectral radiances to finally retrieve the trace gas concentrations. This new generation of sensors like the TROPOMI on board Copernicus Sentinel 5 Precursor (S5P) satellite, delivers challenging data volumes of spectral radiances (level-1 data) that require high-performance computing.

Radiative transfer models (RTMs) are an essential component of retrieval algorithms since they convert optical parameters of the atmosphere into spectral radiances. Usually, the computations of the spectral radiances at different wavelengths are independent and involve a computational loop over wavelengths, as shown in the following pseudo-code:

```
for each wavelength:
      radiance[wavelength] = RT_solver(
          wavelength);
```

However, this line-by-line methodology described in the above code is computationally expensive and inefficient in hyper-spectral remote sensing retrieval applications. This motivates the need for a fast and accurate technique to minimize the spectral mapping computational time. In this regard, Natraj et al. [1] proposed to reduce the dimensionality of the input data due to the strong correlation between atmospheric optical data in the spectral channels by means of an acceleration technique based on the principal component analysis (PCA) to reduce the number of calls to RTMs. To this end, the predictor-corrector approach is proposed as a new method to enhance the performance of the computation of the radiative spectrum for simulating the level-1 data. The predictor stands for a fast approximate radiative transfer model which gives a first estimation for the radiance spectrum. The corrector can be regarded as a post-processing step, in which a correction for the predictor is derived taking into account the redundancy in the input data.

This paper focuses on the reflected spectra of solar radiation in the Huggins band (325-335 nm) that is used to retrieve the total ozone column. The aim of the present work is to evaluate the efficiency of the dimensionality reduction technique of optical parameters in the Hartley-Huggings band regarding the acceleration techniques based on PCA.

2. INPUT DATA STRUCTURE: PRINCIPAL COMPONENT ANALYSIS

The atmospheric state is characterized by the state vector $\mathbf{x}_w$ which consists of a set of optical thickness and single scattering albedo values. For a set of input data vectors

doi:10.2760/848593

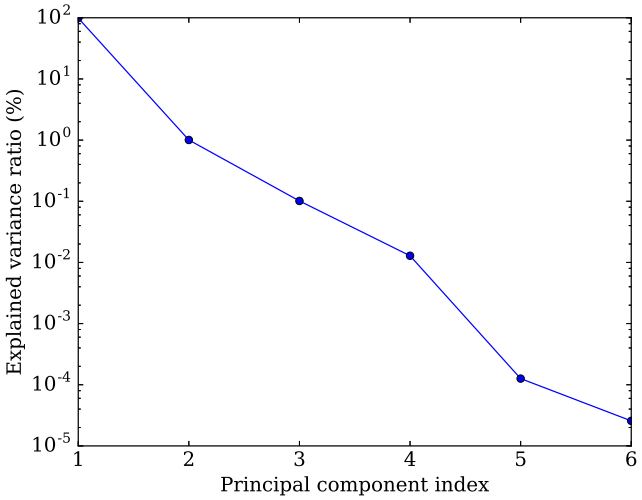


Fig. 1. Explained variance ratio vs. principal component index for the Huggins band simulation.

$\{\mathbf{x}_w\}_{w=1}^{W}$, where $\mathbf{x}_w \in \mathbb{R}^N$, the mean vector is defined as $\bar{\mathbf{x}} = (1/W)\sum_{w=1}^{W}\mathbf{x}_w$ and W is the number of spectral points. By using the principal component analysis, the vector $\mathbf{x}_w$ is represented in a new basis $\{\mathbf{q}_k\}_{k=1}^{M}$ of reduced dimensionality, namely,

$$\mathbf{x}_w \approx \bar{\mathbf{x}} + \sum_{k=1}^{M} y_{wk}\mathbf{q}_k = \bar{\mathbf{x}} + \mathbf{Q}\mathbf{y}_w, \quad w = 1, \ldots, W, \qquad (1)$$

where $\mathbf{Q} = [\mathbf{q}_k]_{k=1}^{M}$ are the matrices of dimension $N \times M$, with columns $\mathbf{q}_k$; y_{wk} is the k-th element of $\mathbf{y}_w \in \mathbb{R}^M$ and N is the $N-$dimensional vector considering a discretization of the atmosphere in L layers, $N = 2L+1$. In the classical principal component analysis (PCA) the basic vectors $[\mathbf{q}_k]_{k=1}^{M}$ are taken as the eigenvectors of the covariance matrix of vectors $\mathbf{x}_w$.

Figure 1 shows the ratio of explained variance depending on the indexes of principal component scores (PCSs), for the input optical data in the Huggins band. The amount of explained variance for the first and second principal component is 98.88 % and 1.00 %, respectively. Thus, not more than two PCSs are sufficient to represent the input data.

3. FAST PREDICTOR MODELS

The fast predictor models consist of radiative transfer models used to compute the radiance as a first estimator. The radiative transfer models used in this study are the multi-stream model, the two-stream model (TS), the single-scattering model (SS) and the model based on the weak absorption Beer-Lambert law. In SS, multiple scattering events are neglected. The multi-stream RTM is based on the discrete ordinate method with matrix exponential (DOME) [2, 3], and uses 16 streams. In the current study, this model is considered as the exact model. Note, that the TS RTM is a two-stream version of DOME, in which the eigenvalues and the eigenvectors of the layer matrix are computed analytically [4]. Essentially, the SS and the Beer-Lambert (BL) models are wrong and directly not applicable for modeling the reflection function in the UV region. However, their results can be correlated with the exact solution and, therefore, can be improved by applying an appropriate corrector.

4. CORRECTION IN THE REDUCED DATA SPACE

We define a correction function $f(\lambda, \boldsymbol{\xi})$ as

$$f(\lambda, \boldsymbol{\xi}) = \ln \frac{L(\lambda, \boldsymbol{\xi})}{L^{\mathrm{P}}(\lambda, \boldsymbol{\xi})}, \qquad (2)$$

here $\boldsymbol{\xi}$ is the state vector of input parameters (practically, a set of monochromatic optical data), L^{P} is the radiance provided by the predictor, $L(\lambda, \boldsymbol{\xi})$ is the radiance simulated by the exact model.

Introducing $\Delta\mathbf{x}_w = \sum_{k=1}^{M} y_{wk}\mathbf{q}_k$, we consider the Taylor expansion of $f(\mathbf{x}_w)$ around $\bar{\mathbf{x}}$:

$$\begin{aligned} &f(\mathbf{x}_w) \approx f(\bar{\mathbf{x}} + \Delta\mathbf{x}_w) \approx \\ &f(\bar{\mathbf{x}}) + \Delta\mathbf{x}_w^T\nabla f(\bar{\mathbf{x}}) + \frac{1}{2}\Delta\mathbf{x}_w^T\nabla^2 f(\bar{\mathbf{x}})\,\Delta\mathbf{x}_w, \end{aligned} \qquad (3)$$

where ∇f and $\nabla^2 f$ are the gradient and Hessian f, respectively. Note, that to estimate them in the initial data space, one would require $N+1$ forward calls for ∇f and N^2+1 forward calls for $\nabla^2 f$. The key point is that the second and third terms in Eq. (3) are estimated in the reduced data space using the formulas of central differences, from which after simplifications, we get

$$\begin{aligned} f(\mathbf{x}_w) \approx & f(\bar{\mathbf{x}}) + \frac{1}{2}\sum_{k=1}^{M}\left(f_k^+ - f_k^-\right) y_{wk} \\ & + \frac{1}{2}\sum_{k=1}^{M}\left(f_k^+ - 2f(\bar{\mathbf{x}}) + f_k^-\right) y_{wk}^2, \end{aligned} \qquad (4)$$

where $f_k^{\pm} = f(\bar{\mathbf{x}} \pm \mathbf{q}_k)$.

Thus, the computations can be organized as shown in the following pseudo-code:

```
for each wavelength:
   approximate_radiance[wavelength] = predictor
       (wavelength);
for each principal_component:
      corrector[wavelength] = corrector(
          wavelength, principal_component);
```


doi:10.2760/848593

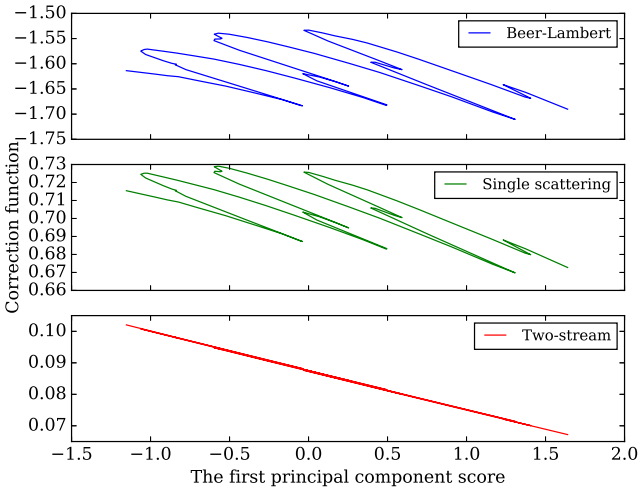


Fig. 2. Corrector vs the first principal component score.

5. PRACTICAL RESULTS

In this section, back-scatter measurements taken by the TROPOMI instrument in the Huggins band are simulated. The wavelength range is 325-335 nm containing $W = 80$. The atmosphere is discretized into 14 layers. The simulations are performed on a personal computer with RAM of 16 GB and processor Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz.

Before proceeding further we analyze the behavior of the corrector as a function of principal component scores (PCS) of the input data. Figure 2 shows that the corrector for the TS model is highly correlated with the first PCS, and almost does not depend on the second PCS. Unlike TS, other predictor models remain correlation for the second PCS. Consequently, keeping one PCS (practically, setting M to 1 in Eq. 4), the computational time for the BL and SS models is reduced by factor of 2, however, at cost of significant lose in accuracy (about 5 times) with respect to the case when $M = 2$ PCSs are preserved. For the TS model, using just one PCS also reduces the computational time by factor of 2 without compromising the accuracy. Thus, for further analysis we set $M = 1$ for the TS model and $M = 2$ for SS and BL models.

In addition, we note, that the dependence of the corrector function is more linear when the TS model is used as predictor. Therefore, the computations of the second order derivatives can be skipped and the last term in Eq. 4 can be neglected. For other models, the dependence is nonlinear and require computations of both the first and second derivatives.

With this setup, in the next section we examine the accuracy and the computational time of the predictor-corrector approach.

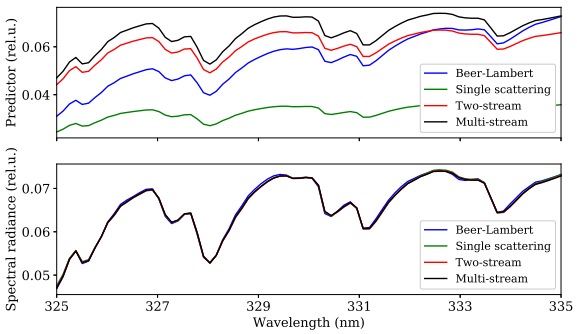


Fig. 3. (upper panel) spectral radiance computed by the predictors; (bottom panel) spectra after applying the corrector.

5.1. Computational time

Table 1 shows the computational time for the models. The computational time required for evaluating the predictor is minimal for the BL model, while it is maximal for the TS model. Such behavior is expected since the BL model is a simplified model (its performance is limited by I/O interfaces).

However, for the corrector function, the results are comparable. The setup, discussed in the previous section, makes the differences even smaller. As a result, the total computational time differs by factor less than two. The PCA requires ~ 0.00075 s for all cases. Note, that the computational time for the exact model is ~ 5.5 s. Thus, the predictor-corrector approach provides the performance enhancement of about 6-10 times, depending on the model used for the predictor. During clear-sky conditions, all computational times are reduced due to the lower number of coefficients in the phase function.

5.2. Accuracy

Table 2 shows the mean and maximum errors in percentage for the three models studied in this paper. The mean relative error for the BL and SS corrected are 83 % and 67 % higher than the TS corrected model, respectively. The similar behavior is also observed for the maximum error, which is lower for the TS followed by SS and finally BL model.

Figure 3 shows the spectral radiances computed by using the predictor models (upper panel) and those after applying the correction function (bottom panel) (all spectra are normalized by the solar irradiance). It can be observed that the higher differences with the exact model are for the BL model followed by the SS and TS models. This result is expected since the BL model is a simplified model and then, the errors are higher. Also, the TS model is a particular case of the exact model with two discrete ordinates, therefore, the higher accuracy is expected. At the same time, the accuracy of the BL model (which is essentially wrong) is significantly improved

Table 1. Computational time for the processes: radiance predicted, radiance corrected and the total time for each model

Predictor model	Computational time (sec)			Acceleration factor
	Predictor	Corrector	Total	
Beer-Lambert	$47 \cdot 10^{-6} \pm 5 \cdot 10^{-6}$	0.470 ± 0.008	0.471 ± 0.008	11.7
Single scattering	0.212 ± 0.008	0.521 ± 0.010	0.734 ± 0.012	7.5
Two-stream	0.531 ± 0.013	0.360 ± 0.012	0.892 ± 0.020	6.2

Table 2. Mean and maximum error for the three predictor models

	BL	SS	TS
Mean error (%)	0.41	0.21	0.07
Max. error (%)	0.87	1.2	0.21

by the correction procedure. This illustrates that the information about the 'exact' spectral radiance is contained in the optical data and can be retrieved by using learning algorithms. In this context, the predictor-corrector algorithm can be considered as a specific machine learning algorithm. But unlike classical machine learning, here we are confronted with the *ad hoc* learning, i.e. the algorithm extracts the most informative part from the data and predicts the correct behavior using the computations in the reduced data space.

6. CONCLUSIONS

Efficiency of the predictor-corrector procedures together with the dimensionality reduction technique has been analyzed for radiance spectra simulations in the Huggins band. The predictor is used to compute the first guess for the exact radiance. Three predictors based on the two-stream model, the single scattering model and the Beer-Lambert absorption law have been included in the computations. The corrector is a function which improves the accuracy of the approximate models. It is computed in the reduced data space, in which the difference between the exact model and the approximate model is estimated. As the input data is reduced, the behavior of the corrector is easily pronounced and therefore can be approximated by a low degree polynomial. Overall, the total performance enhancement due to dimensionality reduction is about one order of magnitude. As shown in [5], the presented approach can be differentiated as a whole or applied directly for differentiated radiances for computing Jacobian matrix in the efficient manner.

It has been shown that two principal components are sufficient to represent the input optical data. For the two-stream model, the corrector is highly correlated with the first principal component, and this dependence is almost linear. That allows to use in the computations only one principal component and the first order Taylor expansion in estimating the correction function. Thus, the results presented in [5] can be improved by reducing number of principal components and computing the correction function only with first order Taylor expansion. Other predictors considered in the paper require two principal components and the second order Taylor expansion. No obviously superior method has emerged in our bench-marking studies (increasingly time-consuming predictors require more sophisticated correctors, and vice versa). However, the best accuracy is obtained with the two-stream model as a predictor. We plan to analise efficiency of the described approach in the framework of GPU computations.

Acknowledgements

This work was funded by the DLR/DAAD Research Fellowships 2018 and 2015 (grants no. 57424731 and 57186656), organized by the German Academic Exchange Service (DAAD) and the German Aerospace Center (DLR).

REFERENCES

[1] V. Natraj, R. Shia, and Y.L. Yung. On the use of principal component analysis to speed up radiative transfer calculations. *J Quant Spectrosc Radiat Transfer*, 111(5):810–816, 2010. doi: 10.1016/j.jqsrt.2009.11.004.

[2] A. Doicu and T. Trautmann. Discrete-ordinate method with matrix exponential for a pseudo-spherical atmosphere: Scalar case. *J Quant Spectrosc Radiat Transfer*, 110(1-2):146–158, 2009. doi: 10.1016/j.jqsrt.2008.09.014.

[3] V. Molina García, S. Sasi, D.S. Efremenko, A.Doicu, and D.Loyola. Radiative transfer models for retrieval of cloud parameters from EPIC/DSCOVR measurements. *J Quant Spectrosc Radiat Transfer*, 213:228–240, 2018. doi: 10.1016/j.jqsrt.2018.03.014.

[4] R. Spurr and V. Natraj. A linearized two-stream radiative transfer code for fast approximation of multiple-scatter fields. *J Quant Spectrosc Radiat Transfer*, 112(16):2630–2637, 2011. doi: 10.1016/j.jqsrt.2011.06.014.

[5] D.S. Efremenko, A. Doicu, D. Loyola, and T. Trautmann. Optical property dimensionality reduction techniques for accelerated radiative transfer performance: Application to remote sensing total ozone retrievals. *J Quant Spectrosc Radiat Transfer*, 133: 128–135, 2014. doi: 10.1016/j.jqsrt.2013.07.023.

SOFTWARE DEVELOPMENT AND VALIDATION UPDATED TO BIG DATA WORLD FOR THE PROCESSING OF GAIA DATA IN CNES

Julie GUIRAUD

CNES, 18, avenue Edouard Belin 31401 TOULOUSE CEDEX 9, France

ABSTRACT

The ESA's Gaia satellite has been launched from Kourou Space Center in December 2013. This mission is the successor of Hipparcos ESA's satellite with also the objective of publishing a stars and objects (galaxies, asteroids, etc.) catalogue but up to 1 billion objects (against 2.5 millions). To achieve this goal, a consortium, called DPAC, was set up to process all the satellite's data composed of more than 400 people mostly in Europe (including scientists and engineers). 9 Coordination Units (CU) corresponding to dedicated thematic and 6 Data Processing Centres (DPC) have created. CNES is in charge of 3 scientific CU (with 7 scientific pipelines) in operations which defines CNES in DPAC as a major DPC. This paper will present the way to update software development and validation to the big data world for the processing of Gaia data in CNES.

Index Terms— Processing Centers, Big Data, Hadoop, MapReduce, software development, software qualification

1. GROUND SEGMENT ORGANISATION

The first catalogue has been released in September 2016, based on the first year of Gaia observations (2014/2015). The second version of the catalogue has been published for the entire scientific community since the 25th of April 2018 (based on observations from 2014 to mid-2016) [1]. The data precision has been improved and all CUs have produced data. The third catalog preparation is ongoing and the publication is scheduled for beginning 2021.

To create these 4 data releases planned, the scientific data processing has been delegated to the Data Processing and Analysis Consortium (DPAC), composed of members of the astronomy community, nationally funded. Following an ESA Announcement of Opportunity, the Data Processing and Analysis Consortium (DPAC) has been created in 2006 and represents now about 450 people, engineers and scientists, from 19 countries across Europe and around the world (Brazil, Algeria…). The yearly workload of the Gaia DPAC is about 250 Full Time Equivalent.
The Gaia DPAC has been divided into 9 Coordination Units (CU) and 6 Data processing Centres (DPC) with an executive committee, the DPACE.

The data processing centres are located across Europe, in charge of processing one or more CU scientific chains [2]:

- CU3: DPCB at Barcelona Supercomputing Centre, Barcelona, Spain
- CU4, CU6, CU8: DPCC at CNES, Toulouse, France
- CU1, CU3 (first processing): DPCE at European Space Astronomy Centre, Villanueva, Spain
- CU7: DPCG at Data Centre for Astrophysics, Geneva, Switzerland
- CU5: DPCI at Institute of Astronomy, Cambridge, England
- CU3: DPCT at Altec, Turin, Italy

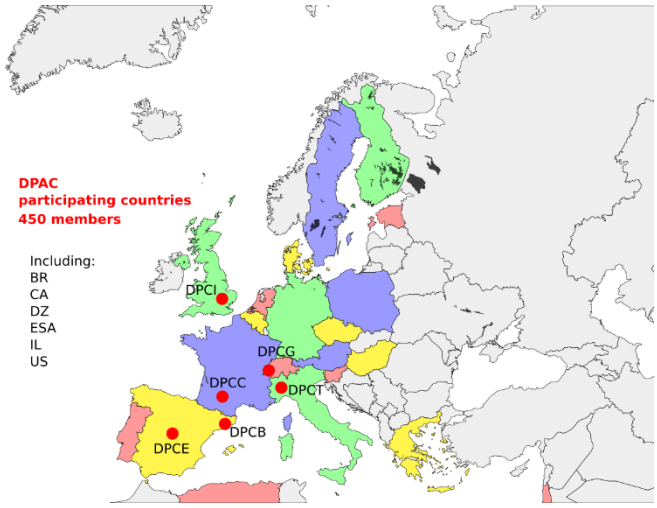


Figure 1. The Gaia DPAC organization (DPAC courtesy)

2. DPCC MAIN CHALLENGES FOR DATA PROCESSING

CNES is responsible for the technical coordination, quality assurance, integration, validation and operations of the scientific developments of Object processing (CU4: Solar System Objects, Non Single Stars, Extended Objects), Spectroscopic processing (CU6), and Astrophysical parameters processing (CU8).
CNES also participates to CU1 as deputy, which is the system architecture unit in charge of the DPAC common

doi:10.2760/848593

tools, definition and management of the interfaces, system tests and operations coordination.
CNES is in charge of the development, validation and operation of the CNES Data Processing Centre (DPCC). The operations are foreseen for 7 years: the 5 years of Gaia mission and 2 years for the final reprocessing of the Gaia catalogue. A total of 4 data releases are planned to the final Gaia catalogue.

The main characteristics that must be taken into account for the management of Gaia CNES project are:
- The huge numbers of contributors working in the DPAC consortium, dealing with heterogeneous ways of working: computer specialists, scientists, managers…
- An organization without contractual relationships, each institution providing its funding and its best effort for the development
- The scientific part of the code developed in each laboratory by the scientists in Java language: code coming from about 80 developers has to be integrated and run in DPCC in a homogeneous way
- The interfaces defined inside the Main Data Base in a collaborative manner by each data producer are difficult to stabilize
- The difficult to manage planning due to large number of people involved and many input data coming from others DPCs

3. DPCC ARCHITECTURE

At DPCC, two platforms have been installed:
- one operational (called OPS platform) to execute all the operational treatments of each chain, with the current following hardware (an upgrade is done each year):
  - 5100 cores
  - 4045 TB
  - 29184 GB of RAM
  - 220 nodes
- one to validate the evolution/correction (called VAL platform) before execution on OPS

Around 110To of cycle 2 data have been received on OPS platform. As VAL platform is smaller than OPS platform, all operational data are not transferred to VAL, only subsets of operational data are transferred. So tests on VAL platform are not always representative.

CNES has developed his own scheduling software called PHOEBUS for “Processing High level Orchestration Engine and BUsiness Service”. The goal of this tool is to schedule, monitor and control the processing. It’s fully automated which allows the data processing on a 24/24 hours and 7/7 days’ basis. CNES choose to use it on several programs including Gaia. DPCC implemented PHOEBUS for GAIA with the Hadoop’s interface and called it SAGA for “System of Accommodation for Gaia Algorithm”.

On each platform, two SAGA instances are installed:
- one of test which can be executed in parallel of the operations (called SAGA TEST)
- one operational (called SAGA OPE)

VAL/TEST is dedicated to the technical validation of pipeline (after delivery from the integration team) but also software (SAGA, Data Delivery Manager…).
VAL/OPE is more representative to OPS platform for pipeline in software configuration point of view and is used to validate patches with smaller dataset of operational data before installation on OPS platform.
OPS/TEST is used for qualification and operations of cyclic pipeline.
OPS/OPE is used for daily pipeline operations.

A workflow is a PHOEBUS term to describe the orchestration and parallelization of steps (which contain modules packaged for PHOEBUS software) creating Hadoop jobs [3].
A chain or a pipeline is a collection of workflows linked, where output of one is the input of the following, with the objective of producing data for the catalogue in a dedicated CU.

4. SOFTWARE DEVELOPMENT AND VALIDATION

The scientific code is developed in java language by the CU members, and then delivered to the corresponding DPCs for integration, system testing and operations.

For the CU4, CU6 and CU8 (operated by DPCC), this scientific code has to be adapted to the big data world that why DPCC supports scientists for developing their code and then to implement it in the Hadoop infrastructure and PHOEBUS software. The CNES procedures of software qualification has been adapted to big data specificities. Finally, the operations have been also adapted to big data environment in Gaia project. This paragraph describes these adaptations on each step of the catalog production at DPCC.

4.1. Development of scientific algorithms

The Hadoop requests are not directly created by the scientific java code. DPCC uses Cascading, a Java library proposing to chain elementary operations which creates complex data processing workflow executed in Hadoop platform [4].

Moreover, the map/reduce algorithm is a powerful tool for processing among of data but the implementation is not natural, as developers are more used to think in a SQL way [5] even if Hadoop is clearly a good tool for processing such among of data in Gaia project.

For example, below, there is a very simple Gaia SQL use case:

doi:10.2760/848593

```
SELECT  astroobservation.data as observation,
        astroelementary.data as aelementary,
        newsource.data as nsource
FROM match
        join newsource using (sourceid)
        join astroobservation using (transitId)
        join astroelementary using (transitid)
WHERE flag =2
```

The figure 2 shows the exact same request in map reduce. The objective of this schema is not to be detailed but to understand how much complicated is a simple SQL request in map/reduce world!

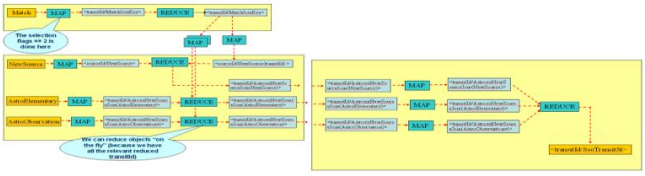


Figure 2. Map/reduce Gaia request

4.2. Encapsulation in Hadoop code

CNES has defined a general organization applicable to all CNES projects: one team is in charge of the development and after a handover, another team is responsible for the operations and exploitation [6].
Furthermore, to support scientists in the big data world, the development team at DPCC has defined the following organization:

- CNES technical coordinator (called CU-T) dedicated to each CU in interface between CNES and scientists;
- CNES technical support (called CU-ST) for each CU responsible of the qualification and integration of the scientific modules;
- Subcontractor integration team in charge of plugging the scientific codes inside the SAGA framework

4.3. Validation and qualification in big data

A qualification in a big data environment implies specificity in software qualification like difficulty of defining dataset for test and the criticality of performance tests to extrapolate the operational processing time.

4.3.1. Data preparation for testing

As it's a Big Data environment, the dataset selection is crucial: not too much to keep it as a test and not operations, but not too few to avoid missing problematic cases. For example, a test is planned with data extracted from each period of data (defined globally by the gaia project).
Each data release corresponds to a cycle for the data processing. The integration of one chain in version N+1 is done before the beginning of cycle N+1 (during cycle N) but with data model N+1, meaning that validation on-site can't start before reception of data model on cycle N+1.
To start the chain's qualification, the operational data have to be available at DPCC.
Cyclic chains are executed on data of cycle N-1 but are compatible of data model cycle N. So the data have to be converted in the new data model before chain's execution or data model changes between version N-1 and N have to be non-breaking.

Some tests are executed on VAL platform and others are executed on OPS platform (for qualification), dataset will be copied on the both platforms.
On VAL platform, the data insertion has to be done on a limited dataset because disk space available on VAL is smaller than on OPS platform.
On OPS platform, the data insertion will be done once for operations but can be used for test purpose as well.

4.3.2. Different type of qualification tests

The first test class is the nominal cases as functional tests. At least one nominal test is played with representative data (which can be operational data or not) which are not the dataset delivered by integration team. This nominal test can be replayed each time a new input data is delivered at DPCC.

One of major concern in Big Data environment is the processing time. For example, 1 billion stars are observed by Gaia, if the processing of each star requires 1 second (which can seem fast), DPCC would need 30 years to process all the stars. During the development and design phase, the performance tests allow DPCC to evaluate the feasibility to operate the pipeline. During the qualification phase, the performance tests are used to extrapolate the elapsed time needed in operations. With all the performances evaluation, DPCC build the operational chronology (where each color/line represents an operational run of a chain).

Figure 3. The DPCC operational chronology

Some performances tests are also used to optimize the technical configuration (number of cores allocated to the

pipeline, with or without other cyclic pipelines…) in order to minimize the processing time.
At each cycle, pipelines are upgraded and new modules are added leading to new chains which are not optimized. When the performance tests start, the new chains can degrade the performance of all the system and need to be tested not simultaneous to execution of operations.

Some degraded cases appear during operations. All these cases cannot be imagined before but during qualification phase, all thought cases have to be anticipated and tested. For example, the following cases have to be checked:

- Use non-consecutive data
- Use the same input data (duplicate output created?)
- Inconsistency of input data (mix of input data coming from different cycles)
- Gaps in the input data
- …

Technical degraded cases are also verified during qualification phase, like deactivation of one module in the pipeline or split of the pipeline…

4.4. Configuration and tuning for operations

During this phase of qualification, the configuration parameters are tested and validated. Using Hadoop system requires a dedicated team of experts to monitor and to tune it for the data processing operations. Each parameter influences the time processing. For example, one CU6 pipeline parameter has been configured to a prime number which leaded to a 6 time reduction of the global processing time.

To allow the operations of several pipelines at the same time, DPCC uses the queue mechanism of Hadoop: each pipeline has a maximum of cores and RAM allocated, depending in the resources needed and available. For example, one CU4 pipeline which uses Cassandra database is configured to 100 cores. In that case, the pipeline is flange because Hadoop is too fast for Cassandra leading to write error due to concurrent access to the database.
To process faster the data, the final computation has not to be done on the final reduce, the CU-ST team has to develop the Hadoop encapsulation to avoid this configuration.

Finally, the input data used by the pipeline are inserted in the Hadoop file system before the operations. The way these files are inserted is crucial to obtain correct processing performances: too big files will saturate the memory at loading by the pipelines, too small files will generate too many temporary files during the reading of input data by the pipeline. In the same way, the files generated by the pipeline (linked to the number of reduce) have to be well managed (which can be contradictory with the low number of reduces to be avoided). Sometimes a post processing is requested to recreate bigger output files (in size but with less number of files).

5. CONCLUSION

After production of two data releases, DPCC implements the lessons learnt of previous cycle operations for the next qualification, for example:

As there is so many input data that the number of errors increases. For example, for cycle 3, the pipelines are robust to duplicate input data, during cycle 2, DPCC faced chain's crashes due to duplicate data in other DPC tables.

The data management is key in big data environment. DPCC has now to implement a way of ordering the output data from operational pipelines to be processed faster by other DPCs consumer.

Finally, even with the best development and qualification, some patches during operations are can't be avoided. There is always unexpected cases, data, field, results… leading to a pipeline issue. Sometimes one patch is not enough et many iterations are made with scientists to find the best solution to a problem. This way of working is close to DevOps method because operational run can least many months and sometimes all tests cannot be re-executed due to time issue before restarting the operations.

6. REFERENCES

[1] Gaia Archive (Data Release 2), https://gea.esac.esa.int/archive/

[2] Data Processing centers and CUs in Gaia Project, https://www.cosmos.esa.int/web/gaia/data-processing-centres and https://www.cosmos.esa.int/web/gaia/data-processing

[3] Hadoop developed by Apache foundation, https://hadoop.apache.org/

[4] Cascading ecosystem, https://www.cascading.org/

[5] map/reduce concept, https://fr.wikipedia.org/wiki/MapReduce

[6] CNES project organisation, https://gns.cnes.fr/en/product-qualification-e-13 and https://gns.cnes.fr/en/project-development (part "6. Phase D and the AR, QR")

PROTOTYPING OF THE DISTRIBUTED DATA PROCESSING CENTER OF LISA

Cécile Cavet[1,*], *Antoine Petiteau*[1], *Maude Le Jeune*[1], *Stanislas Babak*[1], *Michele Vallisneri*[2], *Marc Lilley*[1]

[1] François Arago Centre, APC, Université de Paris, CNRS/IN2P3, CEA/Irfu, Obs. de Paris, 10, rue Alice Domon et Lonie Duquet, 75013 Paris, France
[2] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr, Pasadena, CA 91109 USA

ABSTRACT

The LISA project preparation requires to study and define a new data analysis framework, capable of dealing with highly heterogeneous CPU needs and of exploiting the emergent information technologies. In this context, the mission's **Distributed Data Processing Center (DDPC)** has been initiated. The DDPC is designed to efficiently manage computing constraints and to offer a common infrastructure where the whole collaboration can contribute to development work. A prototype of the DDPC has already been started in order to optimize the detailed design during phase A (initiated in 2018) and to address LISA Consortium needs. This article presents the progress made regarding this collaborative environment in the context of the LISA Data Challenge.

Index Terms— LISA, DDPC, DevOps, Container, Continuous Integration/Deployment, Web application

1. INTRODUCTION

The LISA mission [1] is an ESA Large class mission (L3) which has the goal to study Gravitational Waves (GW) with a space interferometer. The three satellites in formation will be launch in 2034. The LISA Distributed Data Processing Center (DDPC), is the entity that receives calibrated data (level 1 data) from the Science Operations Center (SOC) at ESA, processes them to identify GW sources and their parameters, and sends the results (level 2 and 3 data) back to the SOC. It is also tasked with identifying transient events to provide the outside community with alerts to search for electro-magnetic counterparts with early identification of the transient events and issuing the alarm/warnings to the scientific community at large allowing simultaneous GW and electro-magnetic observations. The DDPC will be delivered by the LISA Consortium under the responsibility of France [1, 3]. Furthermore, the DDPC will have to provide computational resources within the help of Distributed Computing Centers (DCCs) hosted in different countries of the LISA Consortium.

*This activity is supported by the CNES as part of the French contribution to LISA.

In order to provide tools to the LISA Consortium and with the help of the LISA detector definition, the LISA proto-DDPC has been initiated in 2014 as described in [2]. A platform has been released[1] and several tools are ready to be used for the implementation of the data analysis (DA) software, for the development of the LISA simulation pipeline and to enhance the collaborative environment of these tasks.

1.1. Goal

The observations LISA will make will be the first of this kind (i.e. the simultaneous observation of a possibly large number of GW sources) and will explore a new way of observing the Universe. LISA data is expected to be unique in several ways: (i) the data will be signal dominated, it will contain thousands of resolvable signals (ii) many signals are long lived with duration from few weeks to years (iii) we expect very loud GW signals with signal-to-noise ratio of order thousands. Those signals overlap in time and/or in frequency the challenge is disentangle them and characterize. Therefore flexible DA techniques have to be implemented. During the mission operation, the DDPC will have large fluctuations of CPU load due to the need of rapid processing of transient events and the regular iterative in time reprocessing of the data with optimized DA techniques, calibrations, consistency checks, etc. Moreover the event rate for some sources is quite uncertain and can vary from a few events to a few tens thousand events per year. All these factors make the dimensioning of the necessary resources a difficult task.

Some basic methods have been developed during past LISA Data Challenges (LDCs), demonstrating that the LISA DA challenge should be within reach. However the robustness and efficiency of currently available methods have to be studied using LDCs of increasing complexity both for GW sources (with the help of LIGO and Virgo sources) and instrument modeling coming from the recent results of LISAPathFinder and from the implementation of future technological developments.

During the development of the DDPC as well as during

[1] https://elisadpc.in2p3.fr/home

the operation phase, the pipelines will strongly evolve to integrate updated developments in a short loop cycle. For the operation phase, the processing of the data is expected to be done on a daily and weekly basis and should ensure non-regression on short timescales.

Because of all these complexities, new challenges have to be handle using innovative IT technologies such as virtualization and DevOps methods (see [2, 4]). In addition, due to the very long term nature of these activities and rapidly evolving IT solutions, the DDPC has to be kept flexible and easily upgradeable until the end of the mission.

1.2. Why now?

There are a number of reasons pushing for a start of the DDPC during the phase A of the mission:

- a framework to start collaborative work on DA and simulation: needed on the short term;
- an infrastructure to support the DA challenges: the first simulated data has been released under the Radler tag[2];
- an infrastructure to support the end-to-end simulations that will be used to produce realistic data, evaluate mission performances and assess the industrial proposals: needed by 2018;
- a structure hosting the various software used during the development of LISA, in particular performance management tools: necessary in the near future.

The DDPC will be the framework that will support the LISA simulations and the next DA collaborative activities such as LDCs.

2. LDC

The LDC is an open, collaborative effort to tackle unsolved problems in LISA DA, while developing software tools that will form the basis of the future LISA data system. The LDCs are organized by the LISA Consortium's LDC working group. The collaborative effort will provide code and specifications in order to generate challenge datasets, and to work for searching for GW sources and estimate their parameters. Participants can develop their own algorithms or use the one provided by the LDC, and submit their methods and results which will be evaluated and used in designing the most efficient DA pipeline.

2.1. Workflow

The LDC's data release process, shown in Fig. 1, is based on the LDC software including a GW source simulator (LISACode) and the LISA model. Several ingredients are required to simulate LISA data: (i) simulation of instrumental noise; (ii) based on the user request, astrophysical sources are selected; (iii) simulation of gravitational wave signals. The data and metadata in hdf5 format are pushed to the database and published on the LDC web portal. The LDC's

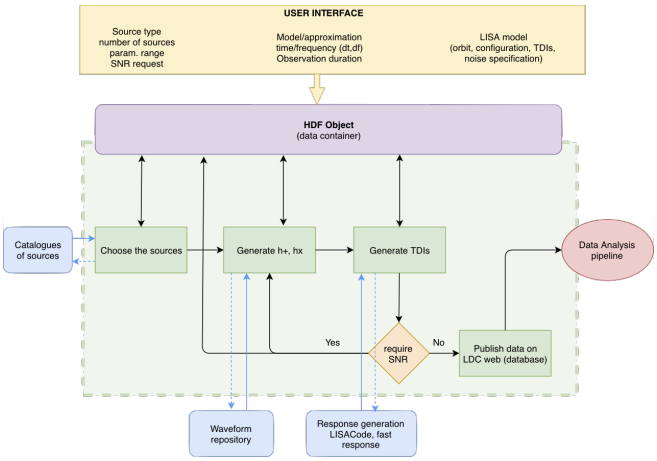


Fig. 1. Scheme of the LDC's data release process: an iterative way is used to publish the simulated data.

complete workflow is run with the following steps:

1. providing of simulated data with the LDC's data release process;
2. team challenge with the analyse of the data to retrieve the source parameters;
3. results submission and evaluation using several metrics: (i) similarity of the proposed solution to the true data, (ii) efficiency and flexibility of the algorithm;
4. increasing complexity of the simulated data.

All these steps are replicated in an iterative way in order to improve the simulated data but at the end, algorithms.

2.2. Running the next LDC within the DDPC

In order to run the next LDC, the team has to provide:

- a common data base associated with a web portal to store the simulation and the team results;
- a configuration management system for the simulator and parameters;
- a dedicated software to compare DA pipeline results and assess their performances;
- an ingestion of DA code into DDPC pipelines to test continuous integration and continuous deployment (CI/CD) concepts and to size CPU needs for constraining DCC resources.

[2] https://lisa-ldc.lal.in2p3.fr/ldc

All these tasks have been started within the prototyping activity of the DDPC. The next section will focus on the web portal implementation and architecture.

3. LDC WEB APPLICATION

As described in the previous sections, innovative and flexible IT technologies have to be selected in the DDPC for handling complexe DA pipelines. The Docker container solution[3] has been chosen as a reference environment for the DDPC. It has been used intensively in all stages of the LDC workflow (simulated data release, web application, packaging of code sharing tools...). For web services, the Docker philosophy is to use containers in a micro-service architecture. This approach enables to deploy smoothly portable system and to upgrade without failure all independent components. The micro-service architecture has been used for the implementation of the application described in this section.

For the new run of the LDC, the DDPC team has developed a web application[4] for the purpose described in Sec. 2.2. This web portal and data base manager, shown in Fig. 2, provides a place to share simulated data and to organise the DA challenge. As described in the next sections, the web portal

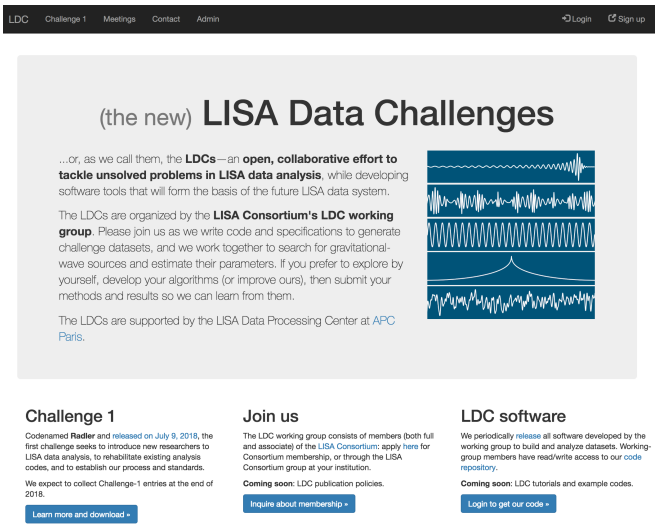


Fig. 2. LDC's web application main page. The application is based on the Django framework and the Bootstrap CSS.

architecture is based on the following items (see also Fig. 3):

- the full website stack is containerized;
- the application is deployed in production in a rolling update mode (health checks during updates) by using a CI/CD platform and a container orchestrator;
- the platform is hosted on an Infrastructure-as-a-Service (IaaS) academic cloud.

The components and technologies that have been used to implement an on-demand service are described below.

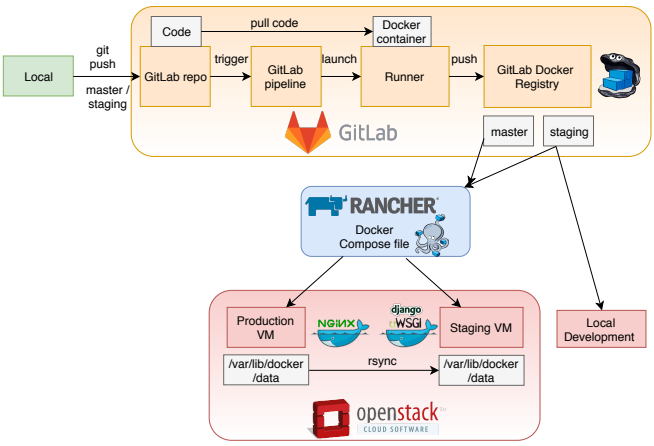


Fig. 3. Scheme of the web application CI/CD: deployment of the micro-service stack. The source code is pushed on the GitLab instance triggering a CI pipeline and producing a Docker image. The CD pipeline notices a Rancher server to deploy micro-services on the production and staging VMs hosted on the OpenStack cloud.

3.1. Web application

3.1.1. Django framework

The web application has been developed in the Django Python framework in order to provide an easy and secure way of sharing scientific data. The Django underlying architecture is the Model-View-Template (MVT) pattern. The data base is built under two data models: the registration model for user account management and the file sharing model.

The authentification system is a customized *UserModel* Django class in order to restrict the data access to registered users. New user accounts are validated by administrators: a unique token identifier is provided via an URL.

The file sharing model is a simple Django model allowing to store URL (instead of data files) in the data base. Furthermore, Django provides an embedded administration interface directly built on the data model and allowing administrators to manage user accounts and data base input.

3.1.2. CI

The application source code is managed by the GitLab instance of the IN2P3 computing center (CC-IN2P3)[5]. The code repository is automatically built with the GitLab pipeline

[3] https://www.docker.com/
[4] https://lisa-ldc.lal.in2p3.fr/
[5] https://gitlab.in2p3.fr

functionality. Indeed, the GitLab-CI instance provides a GitLab runner based on Docker. The runner works with a Docker-in-Docker image (Docker image with privileged mode) providing a Docker daemon inside the container. Each code commits produce a Docker image automatically stored on the GitLab private registry (Docker type) and also registered on the DockerHub account of the DDPC[6].

3.1.3. CD

The continuous deployment has been achieved by using the Rancher container orchestrator[7]. This orchestrator controls running container by monitoring their deployments and providing rollback capability. In the LDC web application, the production and staging environments which correspond to different cloud VMs are registered in the Rancher interface.

Specific stacks are deployed using Docker Compose files as shown in Fig. 4. The Compose YAML file describes micro-services and how they are connected to each others (network, volume, dependance...). The rolling update functionality is permitted by Rancher which give the ability to deploy in production after each code commit. Indeed, the GitLab instance is connected to the Rancher server via a token.

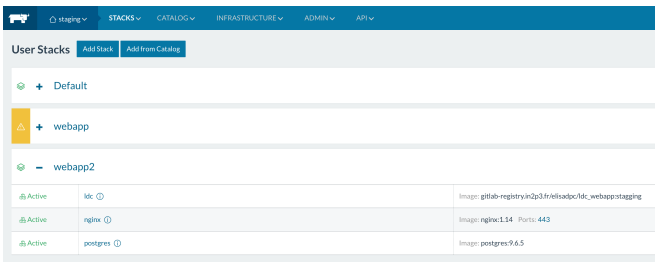


Fig. 4. Rancher web interface: the container orchestrator monitors container deployment and allows rollback. Production and staging deployment are defined in separate environments order to test new features before to push them in production.

3.2. Infrastructure

The infrastructure is hosted on several virtual machines (VMs) of the FG academic cloud[8]. This OpenStack cloud is located in french research laboratories allowing horizontal scaling of the web portal infrastructure (by increasing the number of VM).

The architecture of the LDC web portal is based on three micro-service containers following the Docker philosophy and shown in Fig. 5:

- the Django MVT framework (LTS version) with Bootstrap 3 CSS and an uWSGI gateway;
- a PostgreSQL data base;
- a Nginx web server.

A SMTP server is provided by the cloud infrastructure and allows to send email when new users sign-up.

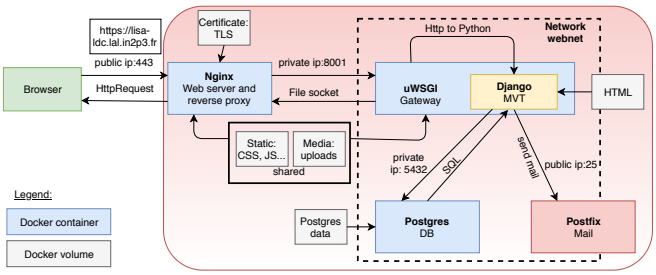


Fig. 5. Scheme of the Web application infrastructure: each container hosts a micro-service.

4. CONCLUSION AND FUTUR DEVELOPMENTS

We have shown that the LDC new run started in the context of the DDPC is an important task for the DA preparation of the LISA mission. Concerning the LDC web application, the micro-service architecture based on container solution is a flexible choice for handling long-term project. Furthermore, futur developments will be provided:

- challenge submission form and result management: the LDC next step after the simulated data diffusion will be the collecting and the comparison of challenger results;
- a REST API and client: API is a standard way of giving a remote access to the application data;
- a data visualization tool: a visualization tool will be implemented to provide a dynamic way of interacting with data.

REFERENCES

[1] Amaro-Seoane, P. et al., Laser Interferometer Space Antenna, arXiv:1702.00786 (2017)

[2] C. Cavet, A. Petiteau, M. Le Jeune, E. Plagnol, E. Marin-Martholaz, J-B. Bayle, A proto-Data Processing Center for LISA, Journal of Physics: Conference Series, Volume 840, conference 1 (2017)

[3] CNES, "NGO Ground segment - Phase 0 report", CNES Report, DCT/ME/EU - 2014.0015466 (2016)

[4] M. Poncet, T. Faure, C. Cavet, A. Petiteau, P.-M. Brunet, E. Keryell-Even, S. Gadioux, M. Burgaud, Enabling collaboration between space agencies using private and cloud based clusters, BiDS'16 (2016)

[6] https://hub.docker.com/r/lisaddpc/
[7] https://rancher.com/
[8] http://www.france-grilles.fr/services-catalogue/fg-cloud/

OPEN SOURCE MULTI-CLOUD EO FRAMEWORK

Sébastien Dorgan, Adrien Oyono, Pierre Crumeyrolle, Audrey Paccini, Vincent Gaudissard

CS SI, 5 rue Brindejonc des Moulinais, 31500 Toulouse, France

ABSTRACT

Earth Observation (EO) satellite missions provide routine, frequent, and high resolution monitoring of our environment at the global scale, delivering an unprecedented amount of data.

Managing these huge data volumes has naturally led the EO world to develop EO cloud platforms – such as DIAS - that provides the necessary resources to store, to process and disseminate such a large amount of data.

To support actors of the EO world CS SI has developed innovative and powerful open source tools: **SafeScale,** an open source multicloud DevOps solution to seamlessly deploy cloud applications on any combination of cloud platforms and **EODAG** (Earth Observation Data Access Gateway), a Python software development kit for searching, aggregating and downloading remote sensed images using a unique API for any EO data sources.

Index Terms— Multicloud, DevOps, Interoperability, Cloud Security, Block Chain, Cryptographic Hash Function, Asymmetric Encryption, Erasure Coding.

1. INTRODUCTION

Today, EO players have the opportunity to access a wide range of cloud platforms for the implementation of their EO data services. For example the five DIAS platforms (CREODIAS, MUNDI, ONDA, SOBLOO and WEKEO) created to foster the use of Copernicus data, but also CloudSigma, ERBC, Amazon Web Services and Google Cloud Platform which offers free access to a large set of EO data: Sentinel 2, Landsat 8, SRTM….

These platforms offer unprecedented operational capabilities for global monitoring from space and an ever more relevant analysis of the state of our planet. However, all these platforms are managed independently and the offer of computing and data resources is heterogeneous. This heterogeneity is a sometimes an impassable difficulty for those who would like to benefit from all these resources.

SafeScale and EODAG have been initiated as CS SI internal R&D projects and put in production for RUS project. Today RUS manages more than 1000 hundreds Copernicus data analytics cloud environments spread over 3 public clouds and it is ready to start operating on the 4 operational DIAS platforms. SafeScale is also used to explore cloud agnostic Payload Digital Ground System efficiency in the frame of the Phase-A of the CNES CO3D Mission and to manage the infrastructure of the Flood Supervisor of the French city of Nimes

This diversity of offers is a fantastic innovation booster but it also constitutes a real technical challenge for the creation of multi-platform services. Indeed, each of these platforms offers different APIs and service levels to access their computing resources and the EO data they distribute.

In this article, we will look at how SafeScale and EODAG tools can overcome these heterogeneities and enable added value services to take the most of the resources offered by EO cloud providers.

2. SAFESCALE

SafeScale is designed in the form of 3 tools: SafeScale **broker**, SafeScale **deploy** and SafeScale **security**. In the DevOps spirit - pushing the automation to infrastructure creation, deployment and maintenance tasks - these 3 tools are accessible via command line interfaces or via software development kits available in various programming languages Golang, Python, Java, Ruby, C#....

2.1. SafeScale broker

SafeScale broker is the pillar of the SafeScale building, it offers simple and generic interfaces to manipulate and configure cloud computing resources. The objective of broker is to create automation scripts to manipulate computing resources without modification regardless of the supplier targeted. To realize such a system we analyze the IaaS (Infrastructure as a Service) API's of the main cloud providers and exhibits the major heterogeneities:

✓ Virtual machines (VM) are assigned using template names specific to each cloud provider. These template names characterize the sizing of the virtual machine in terms of CPU type, number of CPUs, number of GPUs, RAM size, and disk space.

✓ When a user allocates a VM an OS have to be chosen be there is no standard for OS naming, thus, inevitably, OS names differ from one cloud provider to another.
✓ The levels of network services offered by cloud providers can vary between services from Layer 2 to Layer 7 of the OSI model …
✓ Finally, it is necessary to provide users with an aggregated monitoring system allowing them to analyze the overall state of their system.

To overcome these heterogeneities, SafeScale broker is designed around a plugin mechanism. To simplify as much as possible the creation of these plugins, we analyzed the fundamental set of services necessary to implement all the functionalities offered by SafeScale:
✓ List the available VM templates
✓ List the available OS images
✓ Creation and destruction of a network
✓ Creation and destruction of a VM
✓ Creation and destruction of a Block Storage
✓ Assign/unassigned an public IP to a VM

These features are always available regardless of the cloud providers used and easy to emulate on a HPC cluster or simply on a set of interconnected servers using common virtualization tools. As a result, SafeScale allows creating hybrid infrastructure without the need to deploy a private cloud stack on premise.

The following command illustrates how SafeScale broker works and solve the heterogeneity issues mentioned here above.

```
>> broker network create cluster-net --cidr 192.168.2.0/24 |
   --gwname cluster-front |
   –os "Ubuntu 16.04" |
   --cpu 8 --ram 30 --cpu-freq 2.5
```

This command creates:
✓ a network named *cluster-net* with the "*192.168.2.0/24"* CIDR
✓ a gateway which will plays the role of external router and security bastion for the VMs of the network cluster-net

The gateway is a VM under Ubuntu 16.04 OS, with at least 8 cores, 30 GB RAM and a CPU frequency of at least 2.5 GHz. The gateway makes it possible to overcome the heterogeneity of the network layers, because not all cloud providers offer routing services. To be able to find the OS image name corresponding to *Ubuntu 16.04* among the images proposed, SafeScale broker uses the Jaro-Winkler[2] algorithm which provides a distance measurement between character strings. SafeScale broker selects the closest image name to *Ubuntu 16.04* in the sense of Jaro-Winkler. For example with Flexible Engine cloud provider SafeScale broker will select the image named *OBS_U_Ubuntu 16.04*. To find a VM template corresponding to a set of resource SafeScale broker uses the Dominant Resource Fairness[1] (DRF) algorithm to select the VM that best match. In our example the template named *C2-30* will be selected if the cloud provider used is OVH and the template *m5.2xlarge* if the cloud provider is Amazon.

As you can note a user of SafeScale broker is capable in one line of code to create a network, and a VMs without having any knowledge of the underlying cloud provider specificities. VM created with SafeScale are accessible using 'broker ssh' tool or a full web remote desktop build on Apache Guacamole.

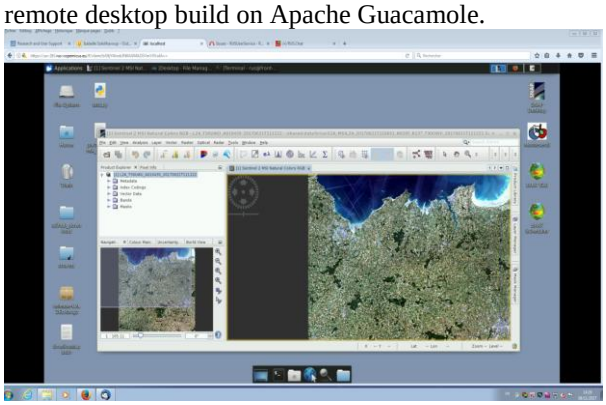


FIGURE 1: SAFE SCALE REMOTE DESKTOP

To offer a global infrastructure monitoring solution SafeScale broker relies on Elastic Stack open source solution. The operating principle of the monitoring system is simple. On all VMs created by SafeScale broker, the Elastic Metric Beat probe is installed. The latter transmits to ElasticSearch the CPU, memory, the file system, the disk IO, and the network IO usage statistics, as well as top-like statistics for every process running on your systems. The Kibana dashboard creation system is used to create ergonomic dashboards.

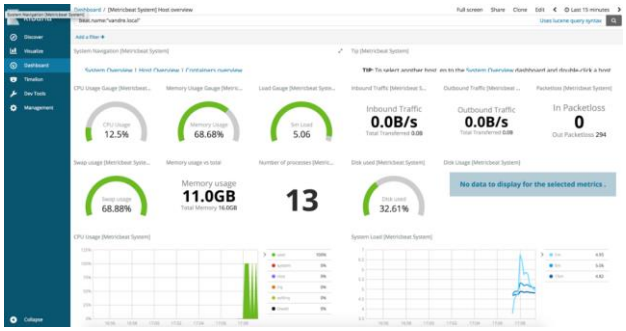


FIGURE 2: KIBANA DASHBOARD EXAMPLE

To date SafeScale broker has the necessary plugins to use the main European cloud providers offering EO data: OVH, Flexible Engine, Open Telekom Cloud, and CloudFerro. It is therefore compatible with all DIAS operational DIAS platforms.

2.2. SafeScale deploy

Although SafeScale broker already greatly simplifies the automation of cloud infrastructure management, the deployment and management of container orchestration systems, Big Data frameworks, and AI frameworks stay very challenging. SafeScale deploy allows creating very easily a wide variety of clusters: Ansible, Kubernetes, DCOS, Spark, HPC.

The following command illustrates how SafeScale deploy works

```
>> deploy cluster dcos-cluster create -F DCOS -C Normal
```

This command creates a DCOS cluster named *dcos-cluster* of *Normal* complexity. SafeScale deploy offers three 3 levels of complexity Small, Normal and Big. Small complexity is intended for test purpose because no high availability (HA) mechanism is provided. In production *Normal* and *Big* complexity offer HA for very large cluster: hundreds of nodes for *Normal* complexity and thousands for *Big* complexity. The master nodes of the cluster are accessible using the full web remote desktop solution.

2.3. SafeScale security

SafeScale security covers 3 facets of security:
- Security of services
- Data security
- Detection of attacks and intrusions

2.3.1. Security of services

SafeScale security provides a gateway system that offers encryption and identity management. To illustrate this mechanism, let's suppose that a SafeScale user created a service named *usr-service* accessible via using and HTTP interface available at port *8080* of a VM named *cluster-node-1* created in the virtual network named *cluster-net*. In this case the user can protect this service running the 2 following commands:

```
>> gateway create gw cluster-net
>> gateway protect gw usr-service cluster-node-1 -p 8080
```

The first command creates a security gateway named *gw* and returns the public IP of this gateway (*ip-gateway*) and the credential to access the Identity and Access Management system as administrator. The second command adds the service to the security gateway. Running these commands the service is now accessible via the URL https://*ip-gateway/usr-service/*, the traffic between the user and the gateway is encrypted and the gateway is connected to an Identity and Management system build on open source software KeyCloak.

2.3.2. Data Security

Cloud Object Storage services are very convenient to store large amount of data but they do not provide the security services necessary to store sensitive data. Furthermore storing large amount on data inside a unique cloud provider make you locked to this cloud provider.

SafeScale security bypasses these downsides by drawing inspiration from the concepts of the Block Chain and combining powerful technologies: Cryptographic Hash, Hybrid Encryption[3] and Erasure coding[4] algorithms. SafeScale security splits the data to be stored in blocks, creates a parity block using 2 or more data blocks using erasure coding, encrypts each block and distributes them over several buckets on different clouds. To encrypt blocks, SafeScale uses "hybrid" encryption. Hybrid encryption combines the convenience of an asymmetric-key cryptosystem with the efficiency of a symmetric-key cryptosystem. Symmetric encryption is used to encrypt data efficiently; asymmetric encryption is used to encrypt symmetric keys. The encrypted symmetric key and the footprint of the block are added to the header of each block.

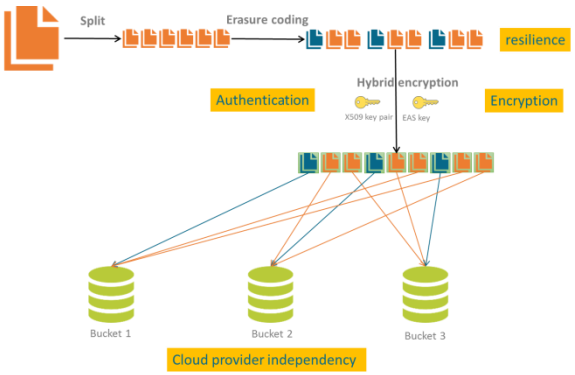


FIGURE 3: DATA SECURITY WORKFLOW

To retrieve a data file all the blocks available for this file are downloaded. For each block the footprint is verified, the symmetric key is decrypted using asymmetric public key and the data are decrypted using asymmetric key. If there is missing or corrupted blocks (wrong footprint) they are reconstructed using the erasure coding system.

SafeScale security uses SHA256 hash function to compute footprints, RSA asymmetric-key encryption, AES symmetric-key encryption and Reed Solomon erasure codes.

2.3.3. Detection of attacks and intrusions

Each VM created by SafeScale broker are equipped with a Suricata probe. The information collected by the Suricata probes can then be collected and analyzed by

Prelude SIEM deployed on the security gateway. Suricata is an Open Source Software for intrusion detection (IDS), intrusion prevention (IPS), and network security monitoring (NSM). It is developed by the OISF (Open Information Security Foundation). Prelude is an Open Source Security Information and Event Management (SIEM) system developed and improved by CS SI for more than 20 years.

3. EODAG

As we have seen in the previous chapters, SafeScale allows managing efficiently virtual infrastructure and computing frameworks in a multi cloud environment.
To enable application developers to create truly portable applications on all EO cloud providers it is necessary to offer a unified data access tools. For this purpose CS SI has developed EODAG (Earth Observation Data Access Gateway). EODAG is a command line tool and a Python framework for searching, and accessing remote sensed images using a unified API regardless of the data provider. It provides 3 main functionalities: List available product types, Search products by types, geographical extent time interval and metadata and load products.

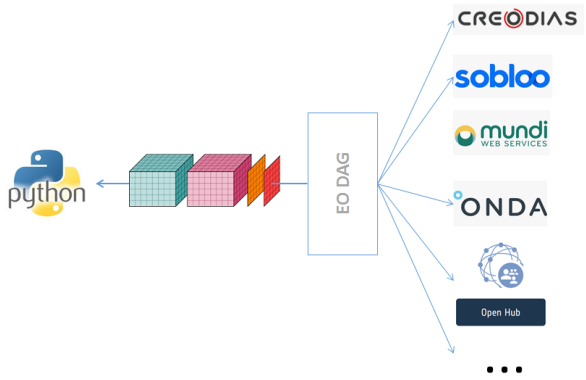


FIGURE 4: EODAG CONCEPT

An experimental product cropping functionality, where a user application can access a subset of an EO product without completely downloading is also available.

EODAG uses a two-level plugin system. Plugin topics are abstract interfaces for a specific functionality of EODAG like Search or Download. EODAG providers are implementations of at least one plugin topic. The more plugin topics are implemented by a provider, the more functionality of EODAG are available for this provider.

EODAG plugin topics are standalone Python packages or modules that comply with EODAG plugin APIs, the Plugin manager load them at runtime.

The strength of the EO DAG plugin system is that the EODAG plugin APIs are basics and simple to implement facilitating external contributions and a growing adoption of the system.

Complete providers (List/Search/Load) are available for all operational DIAS platforms and the Scihub. This means that by using EODAG whatever the DIAS platform or platforms on which you decide to deploy your services, the way you access the data remain the same.

4. CONCLUSION

We have shown in this article that the combined use of SafeScale and EODAG open-source tools makes it easy to create and deploy powerful and secured Earth Observation applications portable on any clouds

5. REFERENCES

[1] Ali Ghodsi, Matei Zaharia, Benjamin Hindman, Andy Konwinski, Scott Shenker, Ion Stoica, “Dominant resource fairness: Fair allocation of multiple resource types, University of California, Berkeley ,Jan. 2011.

[2] Wang, Yaoshu & Qin, Jianbin & Wang, Wei, “Efficient Approximate Entity Matching Using Jaro-Winkler Distance”, International Conference on Web Information Systems Engineering, Oct.2017.

[3] Hofheinz, Dennis; Kiltz, Eike, "Secure Hybrid Encryption from Weakened Key Encapsulation" (PDF). Advances in Cryptology, CRYPTO 2007. Springer. pp. 553–571, 2007

[4] Rodrigo Rodrigues and Barbara Liskov, “High Availability in DHTs: Erasure Coding vs. Replication”, Peer-to-Peer Systems IV 4th International Workshop IPTPS 2005, (Ithaca, New York), Feb. 2005.

BIG DATA GNSS FOR INTERMEDIATE FREQUENCY RECORDING STATIONS

Vicente Navarro[1], Rok Dittrich [2], Konstantin Skaburskas [3], Yeqiu Ying[4], Marc-Elian Bégin[5], Fernando Perez[6]

[1]ESA-ESAC, Madrid, Spain

[2]ESA-ESTEC, Noordwijk, The Netherlands

[3,5] SixSq, Geneva, Switzerland

[4]NSL, Nottingham, UK

[6]RHEA for ESA-ESAC, Madrid, Spain

ABSTRACT

Digitized intermediate frequency (IF) data is the first and most fundamental measurement available following antenna signal receipt. Due to its data rate, digital IF data cannot be stored consistently and is converted to lower density measurements such as pseudoranges, code and carrier phase which generate much lower data rate. The conversion step from IF to observables therefore leads to an unrecoverable loss of information. At present, initiatives to collect IF data focus on applications where the recording requirements are limited to short periods (hours / days). Systematic recording of digital IF data would allow offline re-processing applying any signal processing technique. This scenario opens the door to new science use cases applying innovative processing techniques.

This paper describes a set of use cases that would profit from the systematic, long term storage of digitized IF data.

Moreover, the paper presents a hybrid architecture that seamlessly integrates remote, edge-based GNSS Intermediate Frequency Recording Stations (GIFRES), in charge of collecting the IF Data, with centralized, cloud-based infrastructure in charge of implementing task orchestration and advanced data analysis techniques.

***Index Terms*—**Galileo, GNSS, Big Data, Edge Computing, Cloud Computing, Data Science.

1. INTRODUCTION

The GNSS Big Data activity attempts to develop an initial pilot to demonstrate the possibility of recording and storing IF data in the long term [1].

The main objectives of the activity are:

- demonstrate the potential of continuous or adaptive recording of digitized intermediate frequency data at reference and potentially mobile stations using a distributed or centralised IT facility approach.
- identify relevant application/study cases during the pilot operation period which demonstrate and promote the benefits of a systematic long term digitized intermediate frequency data storage.
- achieve a cost-effective station system trade-off between RF front-end, pre-processing equipment, recording configuration (signal frequencies, bandwidth, number of bits and sampling, effective methods for storing/reducing/processing the large amount of data), storage infrastructure (central versus distributed, HDD and/or tape libraries….), scalability and operation cost.
- analyse the technical feasibility of Big Data solutions [2] to process remotely collected IF Samplers of GNSS signals. Identify the necessary remote and centralised infrastructure including the communication.
- demonstrate through the testbed/station, its use in its use in relevant applications.

2. RELEVANT USE CASES

This section presents relevant uses cases identified as part of the project.

2.1. Geo-hazard / structure health monitoring (High Frequency Vibration)

Throughout the world, Continuously Operating Reference Stations (CORS) provide GNSS raw data such as carrier phase and code range measurements. These high-quality measurements are vital to the GNSS industry and science community. They are essential in providing the service for centimetre-level user positioning, which is highly demanded in markets such as construction, surveying and agriculture. CORS stations can also be used in meteorology, ionosphere study, crustal movement monitoring, geoid determination, geophysical applications, orbit and clock determination etc.

doi:10.2760/848593

Potential GIFRES network shall provide comparable functions/performance to these CORS stations used in the international/regional network.

CORS networks provide a great opportunity to capture the displacements of the Earth's surface or structures, in a level of detail and accuracy unmatchable before the GNSS era. It has become an important resource to monitor geo-hazards such as earthquakes or landslides, and also an important tool to monitor structural heath.

Traditional COTS receivers generally provide reliable positioning output at a rate up to 20Hz. This means that finer details regarding the seismic response and the structure response frequency will be lost due to the technology limitation. Recording IF data will allow higher rate positioning output, either through a post-processing approach or a smarter SDR real time processing. This will be especially valuable for the vibration studies interested in frequencies higher than 20Hz.

2.2. Ionosphere indices generation

Ionospheric perturbations is another important type of vulnerability threat to the GNSS system. Ionospheric perturbations could be caused by several phenomena of which scintillation and space weather events are the two most influential. To tackle this problem, different ionosphere monitoring networks have been established around the world. These monitoring networks consist of specialised CORS stations that can output a set of ionospheric indices, which statistically describe the signal amplitude / phase / spectrum fluctuation. These indices are important to ionosphere monitoring and mitigation. Typically they can only be produced in the specialised ionosphere monitoring type of receiver. Recorded IF data could be utilized in SDR to output these parameters in high quality and in a more standard approach. Unlike the ionospheric monitoring type of receiver, the raw IF data could be used to generate the ionosphere monitoring indices without full-tracking of the GNSS signal. This is particularly useful when the signal suffers severe loss of lock and cycle slips, and could maximize the availability of these indices

2.3. Configurable tracking scheme

There are different scenarios that may require the configurable receiver processing.

One scenario is the adaptive receiver processing according to the environment. A GNSS receiver may lose lock when using traditional tracking configurations under harsh environments, such as severe ionosphere disturbances. Conventionally, some key receiver parameters (e.g. tracking loop bandwidth) are fixed values optimized for nominal ionospheric conditions, therefore they may have difficulty in keeping track on the signal when the ionosphere disturbance is severe. If these parameters could be optimized according to the actual ionospheric conditions, the receiver could be re-configured and provide a more robust and sustainable performance under the unusual environment. In particular, this reconfiguration would be useful if it originates from the IF data, rather than the fully decoded baseband signal, as it could adjust the receiver before the signal becomes un-trackable. Although this adaptive-configuration will be at the expense of the GNSS measurement quality, the loss of tracking problem could be much improved and the GNSS observation availability could be enhanced.

A second scenario is receiver technology development. The GIFRES recorded IF data could allow the re-analysis of the same data in post-processing, with different tracking strategy and receiver settings, which is not feasible with the traditional approaches. This will be particularly useful in incubating and facilitating the development of future advanced receiver technologies.

2.4. Satellite anomaly monitoring

The GNSS signals may suffer from different types of anomalies. These anomalous signals will jeopardize the system integrity and degrade the positioning performance. Some well-known signal anomalies include (but are not limited to):

- Waveform anomaly: as a satellite ages, it may start to transmit distorted signals, possibly caused by a hardware component failure (GPS SVN 19).
- Clock anomaly: space vehicles typically have multiple atomic clocks on board. If one clock unit begins to malfunction it will greatly distort the transmitted signal and degrade performance.
- RF antenna: if improperly designed, the satellite on board antenna may bring a harmful impact to the transmitting signal in a manner of satellite vehicle multipath (GPS SVN 49).
- Launch failure: launch problems may put the satellite in the wrong orbit (Galileo FOC FM1 and FM2), which meant that some receivers are unable to track satellites due to their Doppler shift being outside the nominal range.

These above anomalies have been observed and well-studied in recent years. The analysis has some limitations. With the GIFRES recorded raw IF data, if a feared event is discovered, it could be analysed not only on the impact on the positioning performance, but also in the signal spectrum and waveform level.

2.5. Man-made vulnerability - Radio Freq Interference

The GNSS bands are protected, and as such, the frequency bands should only contain white noise, and the GNSS signals themselves, all of which appear under the noise floor due to

their spread-spectrum characteristics. Since GNSS signals are very weak when they arrive at the receiver antenna, any excessive Radio Frequency Interference (RFI) that falls into the dedicated GNSS radio frequency bands can cause a significant negative impact to GNSS receivers. In the recent years, the appearance of large amount of inexpensive radio equipment whose intention is to disturb or deceive the GNSS receiver functionality has made the situation worse.

There are different scenarios where GIFRES supports Man-made RFI monitoring.

One scenario is testing power levels. Under the normal conditions of an appropriately sited station, the power levels at the receiver should not vary much. If interference is present, it will manifest itself as extra power in the band(s). Interference that is destructive enough to cause problems to GNSS systems will have enough power that it should be easily visible in the FE samples, provided there is no AGC turned on to compensate. The power of the FE samples (or the value of the AGC) can therefore be monitored for signs of RFI. It is possible for anyone with access to the GIFRES sampled data to carry out this testing for power disturbances.

Another scenario is testing the spectrum. By using the FE samples to compute the spectrum, it is possible to look for interference at much smaller bandwidths than simply looking at the power over the whole GNSS band(s). The spectrum-test is usually much more sensitive; however the computational burden is much higher. It could be difficult to run this in real-time over the entire data. The spectrum testing can be done even in the presence of an AGC, though the results are poorer.

3. ARCHITECTURE TRADE-OFF

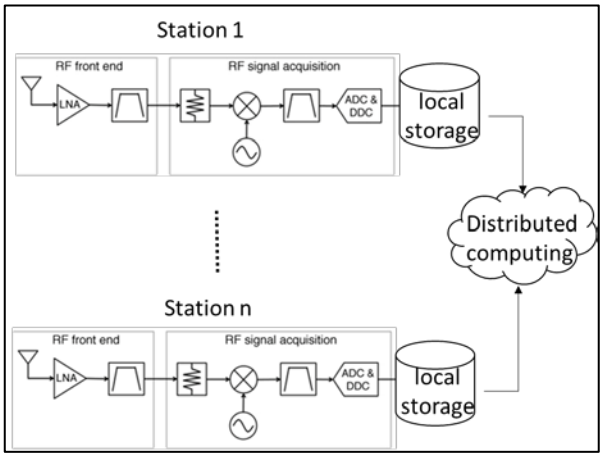
COTS equipment which is potentially suitable to perform IF data recordings are available on the market [4], [5]. Those solutions are mostly used in record and replay applications where the recording requirements are limited to short periods (hours / days). Such solutions may therefore not be optimized for the cost effective and scalable continuous recording solution. Other COTS equipment are of lower quality and cost aiming at (software defined) radio-amateurs.

No attempt to optimize the terminal quality, adoption of standard storage, processing, access to data, and continued permanent recording for multiple applications exists. This case requires hundreds of Terabytes of data per terminal (if all bands were to be recorded continuously). One possible hardware architecture that could be used to capture all the signals in the L band in a storage optimized way is proposed in [3]. Signals whose signal bandwidths overlap or are close to each other would be grouped together as a single signal to reduce the overall needed bandwidth.

During the execution of the work an architecture trade-off analysis has been carried out considering three different deployment options:

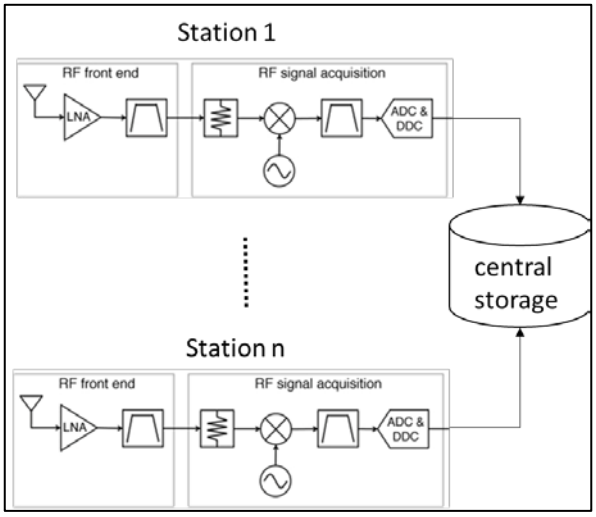
3.1. Distributed storage approach

This architecture highly delegates the storage and processing capabilities to the edge nodes which are coordinated by an orchestration layer in charge of routing data requests to the relevant nodes. This approach minimizes network requirements for data transfer while increases coordination complexity and cost of edge nodes.



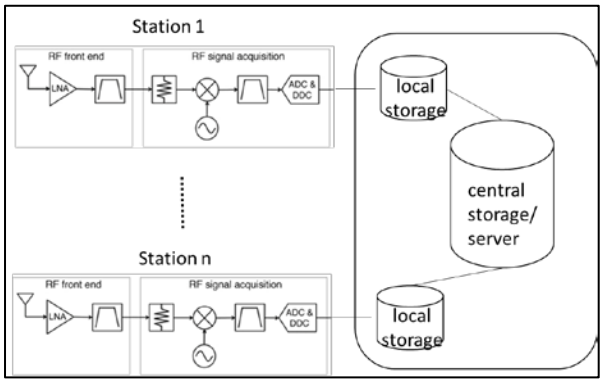
3.2. Centralised storage approach

Contrary to the previous case this architecture highly delegates the storage and processing capabilities to a cloud centric storage and computing resource. This approach maximizes network requirements for data transfer while simplifies coordination and reduces the cost of edge nodes.



3.3. Hybrid storage approach

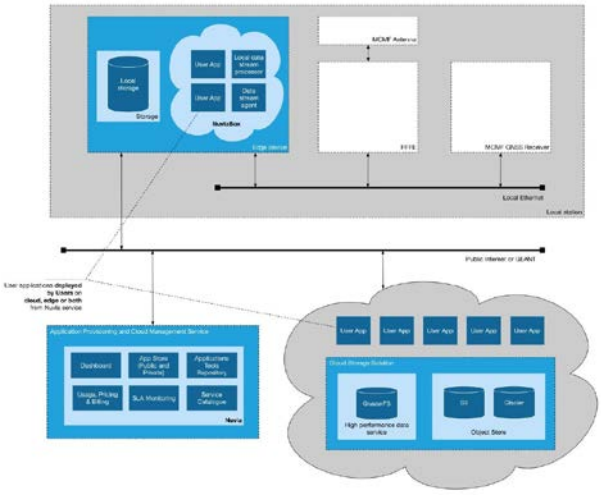
Last alternative consists of a hybrid solution with edge nodes responsible for recording and storage of high fidelity data and cloud computational resources in charge of orchestration and long term preservation of all relevant data.



This deployment approach has been considered to offer the best balance.

4. IT SOLUTION

The IT Solution implemented for the project is highly modular, spanning Edge and Cloud domains. The IT solution as a whole brings new capabilities to the GNSS IF data processing domain via adaptive IF raw data collection, immediate and long-term availability of the IF raw data to support real-time and on-demand processing



At the core of this solution, there is an Adaptive Data Upload Agent responsible to adjust data rates to the science use case to be supported and available bandwidth. Using the wide area network, the station agent will be transferring to its peer agents in the cloud. At both ends, the agents are designed to work in cluster mode, in order to parallelise the reads and writes, to work around the limited single thread I/O performance of object stores.

5. CONCLUSIONS AND FUTURE WORK

At present the activity is about to start the deployment of GIFRES stations at three different locations. This will trigger the beginning of eight months of operations where data will be recorded to validate the processing techniques described in this paper. In parallel, the architecture of the system is being aligned with ESA's on-going initiative "Science Exploitation and Preservation Platform" (SEPP) [6]. Among other things, this alignment will improve the flexibility of GIFRES architecture introducing support for Docker Containers technology. This improvement, combined with the highly modular and flexible architecture already in place is expected to boost computing capabilities across edge and cloud domains. The resulting system and outputs of this activity are to be integrated and deployed on the GNSS Science Support Centre[7] at ESAC.

6. REFERENCES

[1] Recording and Replay for Multiple Constellations and Frequency Bands, Steve Hickling and Tony Haddrell, March 2014, gpsworld.come.

[2] Big Data Challenges @ CERN, Dirk Duellmann, Data & Storage Services, CERN-IT, Australian Bureau of Meteorology CIO Visit, 18. Sep 2015, Geneva.

[3] Report on the Recording of GNSS Digital Data for Scientific Applications, 15 May 2017, GSAC

[4] RF Record and Playback Test System, National Instruments, http://sine.ni.com/nips/cds/view/p/lang/en/nid/206806

[5] Multi-Channel RF Record & Playback, Averna, RP-6100 Series datasheet

[6] Use Case for the ESAC Science Exploitation And Preservation Platform, Christophe Arviset, Vicente Navarro, Ruben Alvarez, et al.

[7] GNSS Science Support Centre – gssc.esa.int

doi:10.2760/848593

GEOINFORMATION SERVICE OF THE RUSSIAN EO-SPACE SYSTEMS INFORMATION PRODUCTS

Markov A.N., Vasilyev A.I., Olshevskiy N.A., Krylov A.V., Salimonov B.B., Stremov A.S.

JSC "Russian Space Systems", Research Center for Earth Operative Monitoring

ABSTRACT

The article discussed processing and distribution technologies of the data based on the Russian Earth observation (EO) space satellites such as Meteor-M, Kanopus-V, Resurs-P. These technologies are realized in the frame of "Basic products bank" geoinformation service designed for high-level Earth remote sensing (ERS) information products creation that are the base for thematic tasks solution. At first, the data and information products verifications in conjunction with the regular ongoing sensors calibrations ensure the quality of provided information. At second, the parallel realization of the ERS data processing algorithms based on the OpenCL/CUDA technologies provides the possibility of the linear technological chains building from the receiving stations to the storage systems. At third, the information products distribution and publication are realized based on the asynchronous model ensuring the balance between the orders quantity and accessible computing resources.

Index Terms **—** Basic products bank, geoinformation service, Meteor-M, Kanopus-V, Resurs-P

1. INTRODUCTION

The United territory distributed information system of the ERS data (ETRIS DZZ) was created in the Russian Federation in 2006-2015 years. Its developing is provided in 2016-2025. ETRIS DZZ is designed for integration to the single geoinformation space of the information resources providing the organization of the target use of the Russian orbital constellation (see table 1), the functioning coordination of the Russian data receiving and processing centers which obtained from the Russian and foreign space Earth observation satellites, ERS data distribution and providing to the users and consumers. The main purpose of ETRIS creation is the full and timely supporting the consumers by the ERS data and relevant products through the special information portals/geo-portals and web-services [1, 2]. The one of such information services is a "Basic products bank" (http://bbp.ntsomz.ru).

The main mission of the "Basic products bank" geoinformation service (BBP GS) is the high-level information products providing to the Russian and foreign consumers and obtained from the Russian satellites.

The article discussed the main types of the information products generated based on the Russian ERS data and supplied by the service as well as the technologies used for its creation and distribution.

Table 1: Specifications of the Russian EO-space systems optical sensors

Satellite, sensor	GSD	Swath	Bands (number)
Meteor-M,MSU-MR	1km	2800km	RED-NIR-MIR-TIR (6)
Meteor -M, KMSS	60/120m,	450/900km	Vis-NIR (6)
Kanopus-V, PSS/MSS	2.5/11m,	23km	PAN(1)/Vis-NIR (4)
Resurs-P, KShMSA	12/24m,	98km	PAN-Vis-NIR (6)
Resurs-P, GSA (hyperspectral)	30m	25km	Vis-NIR (130)
Resurs-P, Geoton-L1	0.7/3m,	38km	PAN-Vis-NIR (8)

2. ERS PRODUCT TYPES

ERS data products formed by BBP GS and provided to the consumers include the following:

processing level 1 products (as CEOS classification [3]) automatically formed based on the primary space data obtained on the receiving stations;

processing level 2 products as CEOS classification (see fig. 1), automatically formed based on the processing level 1 products (CEOS classification). Various vegetation indexes are related to these products (for example, NDVI, SR, RGR, ARVI, EVI, BAI) including ones based on the hyperspectral data obtained from Resurs-P space satellite (for example, NDVI705, mSR705, mNDVI705, VOG1, VOG2, VOG3, PRI, SIPI, PSRI, WBI);

processing level 3 products CEOS (see fig. 2) are also formed based on the processing level 1 and 2 products (CEOS classification).

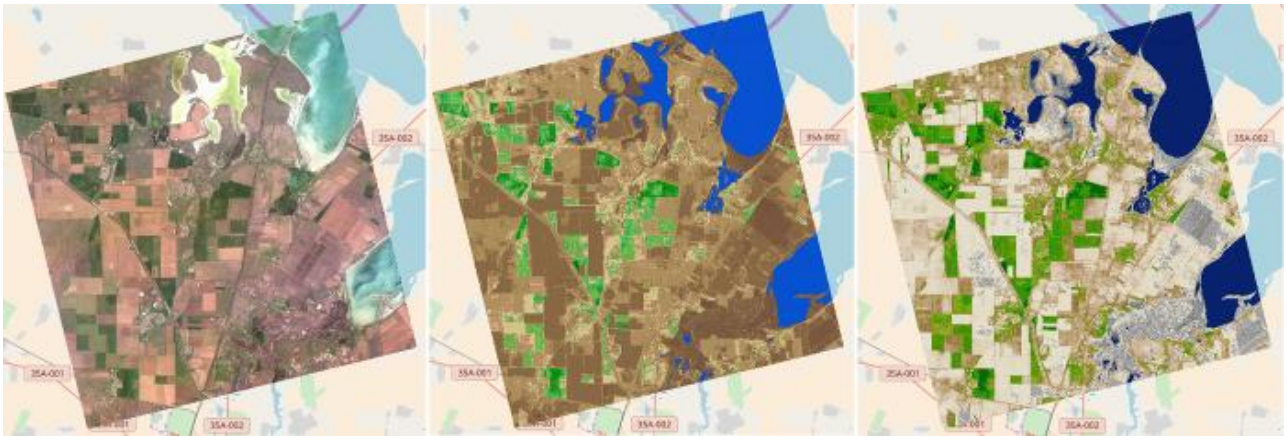


Fig.1: ERS information products formed on Kanopus-V MSS sensor data: on the left – 1D CEOS level product, in center and right – 2 CEOS level products (SR and NDVI respectively) presented in pallets

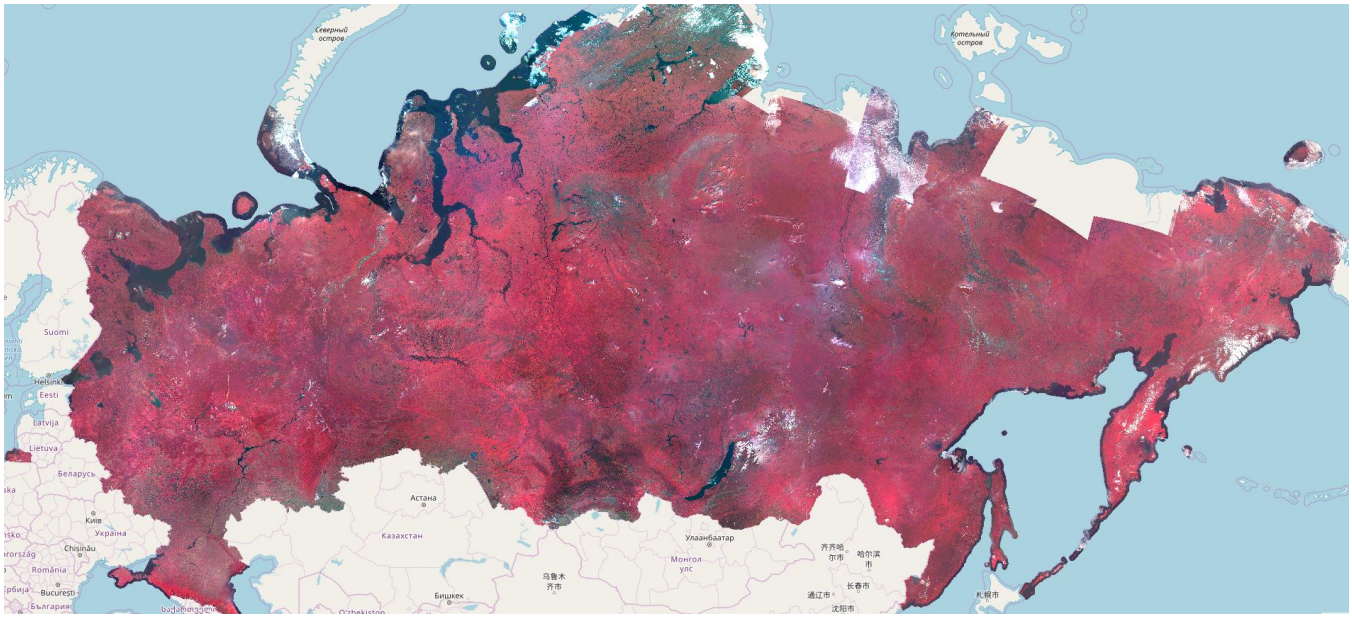


Fig.2: Mosaic coverage in false colors (NIR-Red-Green) of the Russian Federation, formed on Meteor-M KMSS sensor data

3. ERS PRODUCTS FORMATION AND DISTRIBUTION TECHNOLOGIES

Technological models of "Basic products bank" service ERS information products formation are presented in article [4,5]. In the frames of this article we'll discuss the key technological solutions used in related processing level information products forming.

Owing to the significant ERS data volumes the processing (as the re-projecting task) may take the substantial amount of CPU time. At the same time the ERS data structure corresponds to the matrix structure. Therefore it is advisable for high-productive processing to optimize the algorithms taking into account the characteristics of data structures and hardware. The use of the general purpose graphics processing units (GP GPU) as the hardware allows alignment of the data structures and the hardware multiprocessor architecture. OpenCL/CUDA technologies are widely disseminated for GPU programming.

The technologies of 1D CEOS level products formation on the primary space data from the Russian Meteor-M, Kanopus-V, Resurs-P EO-space satellites are based on parallel realization of ERS data processing algorithms (for example, [6,7,8]) using OpenCL/CUDA technologies. In table 2 the results of very high resolution Resurs-P Geoton-L1 sensor data orthotransformation based on SRTM model are presented using different hardware. The results of GPU application demonstrate the possibility of data processing in space information receiving mode (more than 1200 Mbit/s).

Table 2: The test results of parallel processing (orthotransformation of the Resurs-P Geoton-L1 data) using different hardware

Hardware	Speed, MB/s
Intel i7-4820K CPUx4 HTT (x2)	30
Xeon E5-2670 v2 CPUx10 (x2)	96
NVidia GeForce GTX 760	140
NVidia Tesla K40c	349

The formed CEOS level 1 products are fragmented on the scenes (on width) and archived in storage system. Such approach ensures the determined processing speed of any scene.

CEOS level 2 products formation is based on the regular control of cameras radiometric characteristics. In articles [8,9,10] Meteor-M, Kanopus-V, Resurs-P EO-satellite data with foreign Landsat and Terra/Aqua satellite data cross-calibration and comparison results are presented. It demonstrates the stability of cameras and possibility of data application for spectrometric tasks solution.

The formation and distribution of CEOS level 2 products (as well as level 1 products distribution) are realized based on the asynchronous model. That is, the consumers address to the geoinformation service resource, for example, using the graphic web-interface (fig.3), executing the data search in the areas of interest, browsing the quicklook imageries and forming the order on different processing level products. Then BBP GS computing resources provided the formation of information products order pack are used. The order pack downloading is realized by means of HTTP. The formation average time for one order is no more than 10 minutes.

Fig.3: Web-interface screenshot of "Basic product bank" geoinformation service

The ERS products distribution approach allows providing the balance between the appeals quantity/orders of the customers and accessible computing resources.

Mosaic formation technology (CEOS level 3 products) includes the following stages: 1) single cartographic coordinate system data providing; 2) local and global alignment parameters estimation (taking into account the cloud cover masking and cut lines generation); 3) single imagery generation. On the picture 2 the mosaic over the territory of Russia is presented (in summer time of 2017), formed with Meteor-M KMSS data in automatic mode.

Additionally for CEOS level 1 and 2 products formed in order scenes and also for CEOS level 3 mosaics the publication of tiles is executed (fig. 4).

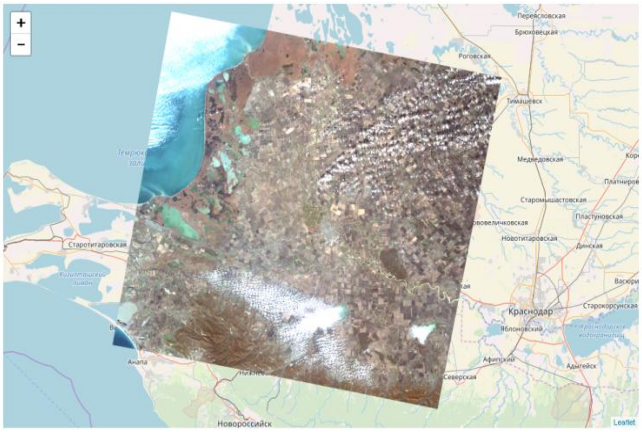


Fig.4: Tile presentation of 1D CEOS level information product formed on "Resurs-P" satellite KSHMSA data

The geoinformation service and external information systems and services of ETRIS DZZ integration is realized through program interface (web-API). Taking into account the open standards and interoperability, the web-API structure corresponds to the requirements for integration to the world exchange catalogues such as CWIC [12].

4. CONCLUSION

On the current stage it may be highlighted the main directions of geoinformation technologies and information systems development, which is oriented to the science and applied tasks solution in the sphere of ERS. It means the cloud technologies and service model of services rendering. The developers of satellite data processing technologies provide to the consumers the services oriented to processing efficiency and high-level information products providing convenience including global and regional monitoring. In this, the work with the services is realized through the web-applications, working in browser and do not need the special software installation on the user's work station. In view of the foregoing, in the article it appears that the "Basic products bank" geoinformation service technological solutions comply with marked world trends.

Further development of "Basic products bank" geoinformation service is performing in the frame of ETRIS DZZ development in 2016-2025 in the directions of information products range increase based on the data of perspective Russian optical and radar EO-space systems, customer rendering service system creation (land and water surface monitoring, atmosphere monitoring, ecology and emergency situations), as well as the automatic mosaic formation technologies based on the Russian and foreign ERS space systems data.

5. REFERENCES

[1] Yu.I. Nosenko, P.A. Loshkarev, "ETRIS DZZ– problems, solutions, perspectives (part 1)", *Geomatics*, pp. 28-32, No 3 (8), 2010 (In Russian).

[2] P.A. Loshkarev, O.O. Tokhiyan, A.M. Kurlykov, A.P. Gladkov, "Progressing of ETRIS DZZ using the cloud computing", *Geomatics*, pp. 22-26, No 4. 2013 (In Russian).

[3] Interoperable Catalogue System, CEOS/WGISS/ICS/Valids, April 2005, Issue 1.2, http://wgiss.ceos.org/ics/documents/ics/Valids-1_2_5.pdf

[4] A.N. Markov, A.I. Vasilyev, N.A. Olshevsky, A.P. Korshunov, R.A. Mikhalenkov, B.B. Salimonov, A.S. Stremov, "Architecture of the Basic Product Bank geoinformation service", *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2016, Vol. 13, No. 5, pp. 39–51.

[5] A.N. Markov, A.I. Vasilyev, D.V. Stepanova, M.A. Evlashkin, A.V. Krylov, B.B. Salimonov, "Technological and Program Models of Remote Sensing Basic Products Formation", *Raketno-kosmicheskoe priborostroenie i informacionnye sistemy*, 2018, Vol. 5, No. 3, pp. 29–38.

[6] A.I. Vasilyev, "Calibration of Kanopus-V satellite sensor during its operation", *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2015, Vol. 12, No. 1, pp. 203–214.

[7] A.I. Vasilyev, A.A. Boguslavskiy, S.M. Sokolov, "Parallel SIFT-detector implementation for images matching", *Proc. of the 21st Conference on Computer Graphics and Vision*, GraphiCon'2011, September 26-30, 2011, Moscow, pp. 173-176

[8] A.I. Vasilyev, A.P. Karpenko, E.L. Shtanov, "Informativity analysis of remote sensing data using NVidia graphic processor units", *Proc. of the International supercomputer conf. "Scientific service in the Internet"*, September 22-27, 2014, Novorossiysk, pp. 45-48 (In Russian).

[9] T.V. Kondrat'eva, B.S. Zhukov, I.V. Poljanskij, A.A. Forsh "Comparison of reflectances of natural objects from Meteor-M No.1 Multispectral Satellite Imaging System and Terra MODIS spectroradiometer", *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2015, Vol. 12, No. 1, pp. 215–224

[10] A.I. Vasiliev, A.S. Stremov, V.P. Kovalenko, A.A. Mikheev, "Methodology of Kanopus-V MSS and Landsat ETM+ basic product comparison", *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2018, Vol. 15, No. 4, pp. 36-48

[11] A.I. Vasiliev, A.S. Stremov, V.P. Kovalenko, "Study of Resurs-P wide-swath multispectral equipment data applicability to spectrometric tasks", *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2017, Vol. 14, No. 4, pp. 36-51

[12] CWIC Client Partner Guide (OpenSearch), CWIC-DOC-14-001r010, May 2017, http://ceos.org/document_management/Working_Groups/WGISS/Projects/CWIC/OpenSearch/CWIC_OpenSearch_Client-Guide.pdf

COPERNICUS AUSTRALASIA – TYRANNY OF DISTANCE

Simon Oliver[1], Alla Metlenko[1], Joshua Sixsmith[1], Edward King[2], Dan Tindall[3], Matthew Adams[4], Tony Gill[5], Rafael Kargren[6], Ben Evans[7]

[1]Geoscience Australia, Corner of Jerrabomberra Avenue and Hindmarsh Drive, Canberra, Australia - simon.oliver@ga.gov.au

[2]Commonwealth Scientific and Industrial Research Organisation, CSIRO Marine and Atmospheric Research, Castray Esplanade, Hobart, TAS, 7001, Australia

[3]Queensland Department of Environment and Science, Ecosciences Precinct, 41 Boggo Road, Dutton Park, QLD 4102

4Landgate, Midland Square, Midland, WA 6056

[5]New South Wales Government Office of Environment and Heritage, Level 1, 48-52 Wingewarra St, Dubbo

[6]New Zealand Centre for Space Science Technology, Centre for Space Science Technology, Level 1, 50 Centennial Ave, Alexandra 9340

[7]National Computational Infrastructure Australia, The Australian National University, 143 Ward Road Acton, ACT, 2601

ABSTRACT

Copernicus Australasia [2] is a regional Hub supporting Europe's Sentinel Online Copernicus programme within the South-East Asia and South Pacific region. The Hub supports government and commercial information requirements and enhances access to Earth observation (EO) data by research, industry and civil society, and facilitates collaboration between Australia, New Zealand, Europe, South-East Asia and the nations of the South Pacific. The Hub is hosted on the National Computational Infrastructure (NCI Australia) in Canberra, Australia, and provides free and open access to the hosted datasets. The NCI is Australia's primary eResearch infrastructure provider, delivering the underlying data services, storage and compute services for the Hub. The Hub was established to ensure a single point of truth for critical EO data assets in the region as well as providing a coordination function for partnerships with major international providers of EO data. Partner organisations replicate subsets of the data from the Hub to store and maintain duplicate copies and enable business operations as well as contingent services if and when they may be required.

The Hub enables research and development, and the delivery of operational Government programs. It is supported by the partner agencies contributions, including data archiving, scientific computing and storage resources, improved imagery corrections, calibration and validation activities and algorithm development. It also provides a mechanism for information access requirements to be shared within Australia, Europe and amongst other international partners as part of the Australian Government's commitment to provide greater tangible contributions to the international EO community and industry as per the strategy outlined in the Earth Observation Australia (EOA) Community Plan [10].

The Hub utilizes government research and digital infrastructure to move large volumes of satellite data across the globe in a timely and cost effective manner. Sponsored by both Australian and New Zealand government organisations, the Hub supports programs which depend on a consistent and high volume data stream from supported missions including: Geoscience Australia's Digital Earth Australia program (DEA); Australian state government land cover and land management and monitoring programs; and several Commonwealth Scientific and Industrial Research Organisation (CSIRO) research programs.

The regional Hub is established under an agreement between Australia and the European Union. Implementation of the Hub is facilitated by arrangements between the European Space Agency (ESA), the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) and Geoscience Australia (GA). The project is operated collaboratively by Geoscience Australia, Queensland Department of Environment and Science, New South Wales Office of Environment and Heritage, Western Australian Land Information Authority (Landgate), CSIRO and New Zealand's Centre for Space Science Technology (NZCSST) who joined the consortium in 2018.

Index Terms— remote sensing, Copernicus, data hub, Sentinel

1. INTRODUCTION

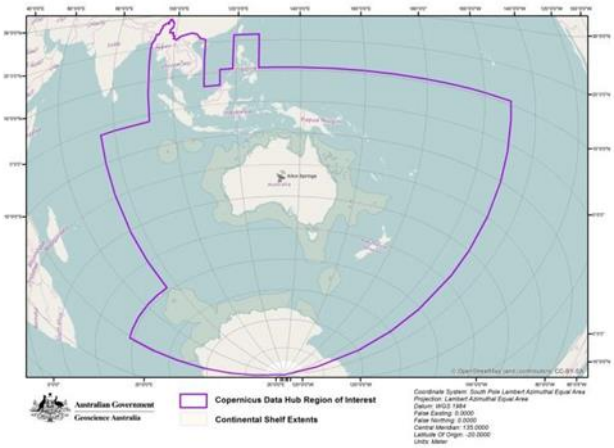


Fig.1 Copernicus Australasia Data Hub Region of Interest

Through the Copernicus Australasia Regional Data Hub project, Geoscience Australia, on behalf of its partner entities, manages a contract with the National Computational Infrastructure (NCI) to provide the data syncing, storage and access services which underpin the Hub's operation. Copernicus Australasia provides user access to satellite Earth observation (EO) data from Europe's Sentinel Online Copernicus programme in order to facilitate uptake and access to the data within the Asia Pacific region. As of August 2018, Copernicus Australasia stored 1.5 PB of data with an average user download volume of ~200 TB / month across all products.

2. BACKGROUND

Copernicus Australasia was primarily established to provide a unified Australasian approach to ensure that Sentinel data was regularly captured over Australia and to coordinate and consolidate the retrieval of Sentinel mission data from Europe to the Asia Pacific region. The Copernicus Australasia Hub provides a single point of truth in Australia for data from the Sentinel missions and exploits synergies between overlapping interests of states and national agencies.

A secondary reason to establish the Hub was to get Copernicus datasets closer to the computing resource from which products are produced. The Copernicus Australasia Hub is housed within the NCI and made available through data services as well as the computational systems. For example, the Raijin supercomputer provides over 80,000 CPU cores and the ability to process and derive national-scale Level-2 and Level-3 products [12]. Partners also copy subsets of data via the data services to their own infrastructure to meet their individual program objectives.

The Hub is an exemplar for multi-agency collaboration: Geoscience Australia manages the project and administers the contracts with partners and the facility manager; partners contribute financially to a share of the Hub; partnership provides for input to the strategic decision making as well to development of technical solutions and user engagement strategies, and; the NCI provides the data syncing and access capability and connectivity via Australia's Academic and Research Network (operated by AARNet [11]). Copernicus Australasia provides data syncing and access functions which enable free and open data online. Public data access is facilitated via a number of services: the Sentinel Australasia Regional Access (SARA [1]) portal which provides a map-based search and download function, as well as an Application Programming Interface (API) which enables programmatic search and retrieval of data - SARA is powered by RESTo (REstful Semantic search Tool for geospatial), a solution to searching and retrieving EO images from very large databases; The auscophub code repository [7] for users wanting to make use of the API for batch processing [3]; NCI's general purpose THREDDS data server [13] (http://dap.nci.org.au/thredds/catalog.html), and; direct file system access to files on NCI's Raijin supercomputer for registered users.

3. CHALLENGES

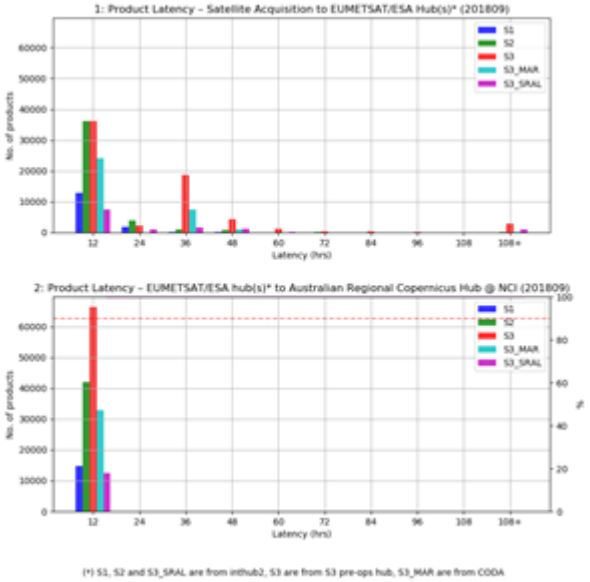


Fig. 2 Product latency comparing satellite acquisitions to Europe hubs(at top), and time to download to Copernicus Australasia (at bottom)

The Hub commenced routine operations in 2018 and by October 2018 had delivered nearly 3 petabytes of data to users external to the NCI. In the nascent stages of establishing operations, a great deal of work was undertaken to improve data transfer, stability and quality of service. The Hub technical team worked with our counterparts at ESA to improve network performance and increase bandwidth

between ESA's data centre in Frankfurt and NCI. ESA, NCI and other international partners have tuned the network for long distance transfer. ESA also changed the maximum transmission unit (MTU) size configuration on the International DataHub network to improve data transfer. Further improvement will involve tuning of not just the network, but also middleware and the DataHub software. Successful operation of the Regional Data Hub depends on being able to serve a complete and valid copy of regional Sentinel data with acceptable latency, equal to ESA's Open Access Hub. Additionally, any user should be able to easily access update-to-date information about completeness, integrity and latency. Under nominal operations, the Hub aims to make 90% of the data available within 24 hours of arriving on the European hubs (Fig. 2 provides an example of the Hub operating within these requirements.).

MD5 checksums are used to confirm successful data replication and routine data checks are undertaken on the data to ensure consistency with European data hubs.

Table 1. Copernicus Australasia supported products (2018)

Mission	Products	Area of Interest
Sentinel-1	level-0, level-1 and level-2	Regional
Sentinel-2	level-1C	Regional
Sentinel-3	level-1, level-2 land, level-2 marine	Global

The products outlined in Table 1 are queried every hour for new data in the region of interest (Fig.1). The range of products on offer is constantly reviewed and updated as new sources become available.

4. APPLICATIONS OF SENTINEL DATA IN THE REGION

Data from Copernicus Australasia is being used within the local region to support programs of work delivered by the partners, including:

- near real-time and long-term monitoring of native vegetation changes and woody vegetation extent in support of legislative and regulatory requirements;
- improving the quantification of ground cover information for monitoring wind and water erosion and agricultural land management and productivity
- modelling water quality and land condition of Great Barrier Reef catchments and receiving waters
- mapping of coal seam gas physical infrastructures to support regulatory and activity monitoring
- development of methodologies to monitor ground deformation and subsidence
- measurement of natural water flows and monitor water quality both coastal and inland
- development of "near-real-time" burnt area mapping and feed into fire-simulation tools to model fire behaviour and improve greenhouse emissions programs
- mapping fire scars and fuel loading
- development of methodologies for monitoring broad crop types
- development of vegetation biomass and net primary productivity products to inform emerging market-based instruments for carbon and other greenhouse gas trading schemes
- responding to, and recover from disasters
- monitoring, detecting and characterising land, water and infrastructure changes across Australia
- climate impact research and studies

This work is being undertaken through programs including: Queensland Government's Statewide Landcover and Trees Study (SLATS [5]), Land Use Mapping Program (QLUMP); New South Wales Statewide Landcover and Trees Study, Tree Clearing Early Detection, Landuse mapping program, and Ground cover change mapping; Western Australia's Land Monitor [6], FireWatch-Pro and FireWatch Aurora, Floodmap, DataWA, and Digital Earth Australia [4] (DEA) - an infrastructure and service delivery program that provides government and industry access to standardised Earth Observation data across Australia, powered by OpenDataCube's [8] API, for developing Earth monitoring and mapping applications. The DEA program also leverages NCI's GSKY [9] which provides a high performance online data service that provides Open Geospatial Consortium (OGC) APIs to satellite data collections.

5. CONCLUSION

Copernicus Australasia was established in order to coordinate engagement with and access to data from the European Union's Sentinel satellites. After overcoming significant technical challenges in its formative phase, the Hub is achieving its primary aim of delivering reliable and dependable replication of data from Europe and has become a point of truth for users within the Asia Pacific Region. The Hub has proved effective in reducing the load on European data hubs and has made it easier for users to access large volumes of data quickly. The Hub partnership has now grown to include membership outside Australia and has proven to be an exemplar model for collaboration across the region. Copernicus Australia is returning real benefits to its constituents and underpins a growing number of important national and state-wide activities which are realising the value of free and open access to Copernicus data.

6. ACKNOWLEDGEMENTS

The authors would like to thank the Copernicus Australasia Regional Data Hub partners for their contribution to this work including: Geoscience Australia, New Zealand Centre

for Space Science Technology, Queensland Department of Environment and Science, Western Australian Government Land Information Authority (Landgate) , Commonwealth Scientific Research and Industrial Organisation (CSIRO), New South Wales Office of Environment and Heritage and the National Computational Infrastructure at the Australian National University.

7. REFERENCES

[1] Sentinel Australasia Regional Access portal. [Online]. Available: https://copernicus.nci.org.au/sara.client/#/home. [Accessed: 10- Oct- 2018].

[2] "Copernicus Australasia", Copernicus.gov.au, 2018. [Online]. Available: http://www.copernicus.gov.au/. [Accessed: 10- Oct- 2018].

[3] "Auscophub bitbucket repository", 2018. [Online]. Available: https://bitbucket.org/chchrsc/auscophub/src/default/. [Accessed: 10- Oct- 2018].

[4] "Digital Earth Australia - Geoscience Australia", ga.gov.au, 2018. [Online]. Available: http://www.ga.gov.au/about/projects/geographic/digital-earth-australia. [Accessed: 10- Oct- 2018].

[5] "Statewide Landcover and Trees Study (SLATS) | Environment, land and water | Queensland Government", Qld.gov.au, 2018. [Online]. Available: https://www.qld.gov.au/environment/land/vegetation/mapping/slats . [Accessed: 10- Oct- 2018].

[6] "Land Monitor", Landmonitor.wa.gov.au, 2018. [Online]. Available: http://www.landmonitor.wa.gov.au/ . [Accessed: 10- Oct- 2018].

[7] "RESTo code repository", GitHub, 2018. [Online]. Available: https://github.com/jjrom/resto . [Accessed: 10- Oct- 2018].

[8] "OpenDataCube", opendatacube, 2018. [Online]. Available: https://www.opendatacube.org/ . [Accessed: 10- Oct- 2018].

[9] "GSKY", http://gsky.nci.org.au, 2018. [Online]. Available: http://gsky.nci.org.au/. [Accessed: 10- Oct- 2018].

[10] "The Plan", Earth Observation Australia, 2018. [Online]. Available: https://www.eoa.org.au/aeocp-the-plan/. [Accessed: 10- Oct- 2018].

[11] "AARNet", Aarnet.edu.au, 2018. [Online]. Available: https://www.aarnet.edu.au. [Accessed: 10- Oct- 2018].

[12] "Data Processing Levels | Science Mission Directorate", Science.nasa.gov, 2018. [Online]. Available: https://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products. [Accessed: 11- Oct- 2018].

[13] "Unidata | THREDDS Data Server (TDS)", Unidata.ucar.edu, 2018. [Online]. Available: https://www.unidata.ucar.edu/software/thredds/current/tds/. [Accessed: 11- Oct- 2018].

NEW INFRASTRUCTURE & AUTOMATIC INFORMATION EXTRACTION FOR DISRUPTIVE SERVICES BASED ON EO PRODUCTS

F. Tromeur
J. Helbert

Telespazio France

ABSTRACT

Today space based services are facing a new challenge to cope with a ***variety of sensors***, and a ***huge quantity of data***. To respond to these challenges, Telespazio France is putting in place new disruptive architectures coming from GAFAMs [1] and algorithms based on artificial intelligence. The objective is mainly to ***bring new added value*** in short time at large scale by supporting different kinds of EO data (SAR and optical data) and also information coming from other sources such as AIS for maritime surveillance.

1. INTRODUCTION

Telespazio France (TPZF) is building a portfolio of geo-information services to its clients from the civil and military domains among which the French Navy, Maritime Affairs from various countries, intelligence services as well as actors from agriculture, forestry and environment monitoring. These services rely on space imagery to extract valuable insights and to provide end users with high-level processed information.

In the spirit of continuous improvement, one has to propose better added value and to address a larger diversity of data together with larger data volumes to cope with larger areas.

Today Cloud & Big Data technologies, Artificial Intelligence are able to disrupt the geo-information services as they did in social media, and retail.

This paper presents the solutions that TPZF has put in place to deal with these challenges; and concludes on the envisaged improvements in the future.

2. STATE OF THE ART

Formerly, TPZF was using a classic SDI (Spatial Data infrastructure) to host its geo-information services. This kind of system is able to ingest, index, store and visualize georeferenced data with a classic service oriented architecture. The platform was based upon virtualization to optimize the servers' power and to reduce costs.

The fact is that this sort of architecture is not really suited for cloud, due to the monolithic aspect of the system and the difficulty to deploy it (manually following a manual installation). Today new **devops** technologies enabling to code the infrastructure are becoming very simple to deploy the system. Furthermore this technology allows to hybrid our infrastructure with private and public clouds in order to generate private data on premise and then burst them on a cloud; and to process huge volume of data (e.g. large volume of Sentinel imagery on DIAS).

In addition, the former system was not able to process imagery data and had difficulties to deal with huge data volume due to classic SQL storage (e.g. worldwide AIS data). Today **NoSQL** databases [1] are able to treat easily such data in an efficient manner.

Last, new patterns of architecture known as **micro services** bring the capacity to make the system very scalable by isolating simple business functions, and to horizontally scale the system in order to simplify the maintenance and the cost of the infrastructure.

All these paradigms allow to imagine a new platform to be able to cope with huge data volumes, large variety of data and to be versatile to respond to new innovative services in the geospatial field.

Besides, Artificial Intelligence (AI) is bringing a new paradigm in image processing and is fostering applications and services derived from Earth Observation (EO) products. Since early studies on neural networks showing the potential of learning techniques for image recognition (LeCun et al. [2], Figure 1), progress in deep learning algorithms based on convolutional neural networks (CNN) – with important actors from the IT such as Google, Facebook or Microsoft, offering algorithm frameworks and participating actively in the research on AI (Tensorflow, PyTorch, Microsoft Cognitive Toolkit…) –, hardware – with the emergence of GPU computing, large storage and processing capabilities (Google Cloud, Amazon Web Services) – as well as availability of large EO image catalogs (ESA SciHub, USGS EarthExplorer) has enabled the development of services based on massive image processing for environment and urban monitoring, security and surveillance, intelligence...

[1] GAFAM: Google, Apple, Facebook, Amazon, Microsoft

doi:10.2760/848593

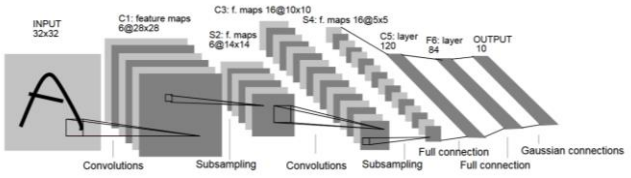


Figure 1. Example of LeNet-5 CNN used for handwritten digit recognition, from LeCun et al [3].

2. SERVICE REQUIREMENTS

To be efficient, monitoring and surveillance services require to download and to process numerous images on a regular basis and potentially over large areas so as to provide useful added value information to the end users. In the case of maritime security Sentinel-1 & 2 images are regularly made available through several dissemination sites (e.g. SciHub), which represent thousands of images available each year. Furthermore AIS data are recovered in NRT and require high level performance tools to decode the AIS frames efficiently without any bottleneck. This leads to the necessity of developing dedicated infrastructures and software to be able to process the data as soon as a new image is available, in a near real time approach so as to provide the users with new products without delays. Consequently infrastructures enabling large storage capabilities, fast and automatic processing are required.

4. INFRASTRUCTURE DEVELOPMENT

Objectives

In order to deal with these challenges TPZF have put in place a new geo-platform called STORM. It is a homemade product built for its own geospatial services; in particular for its maritime surveillance services. It has the capacity to ingest, index, store, and display geo data (satellite imagery, IoT data, positioning data), through space and time. Moreover the system is able to host processing, to create added-value information from raw data with dedicated algorithms (e.g. AI algorithms for object detection in remote sensing images). The platform is entirely designed taking into account new cloud and big data technologies.
Five challenges have been addressed:

1. Solve the problem of velocity for the AIS ingestion;
2. Have a more versatile system to follow incremental business development, reducing the time to market;
3. Have several clients/projects on a unique platform;
4. Be able to process huge volume of EO data;
5. Be able to deploy a processing chain on a public cloud to minimize costs.

Methodology

To support all the developments an **agile** methodology following **SAFE**[2] framework has been put in place. All the R&D and evolutions of the platform are pushed by the business with the product manager and the business owners. Then the architect together with the scrum master and team design develop the system according to an incremental approach (Sprint, Release). This allows to fit strictly to the needs and to avoid techno-push effect.

Architecture

The chosen architecture relies on **micro services** [6] being able to isolate each business function and to simplify the evolving maintenance.
We have also chosen the **Docker** technology which is a container solution to isolate functional service and algorithms from the platform (OS, libraries). This disruptive new standard is a cornerstone to bring a cost effective solution for processing deployment and to ease the deployment on whatever cloud solution.
A message bus based upon **Rabbit MQ** technology orchestrates all the processes from the image ingestion to the dissemination of a result or report.

To be able to address multiple clients of the same platform and to guarantee the strict separation of data, we have put in place a SSO (Single Sign On) solution. This multi-tenant architecture is based upon the **keycloak** solution.

To deal with huge data volume (e.g. AIS data, object detections on satellite imagery) we set up a noSQL database, **Elastic Search,** able to index geospatial data in a simple manner offering the capacity to retrieve data in an efficient way. Moreover this solution is scalable on multiple servers.

[2] Safe : https://www.scaledagileframework.com/

doi:10.2760/848593

Figure 2 - Overview of the STORM Platform Architecture

The infrastructure is composed of three subsystems:

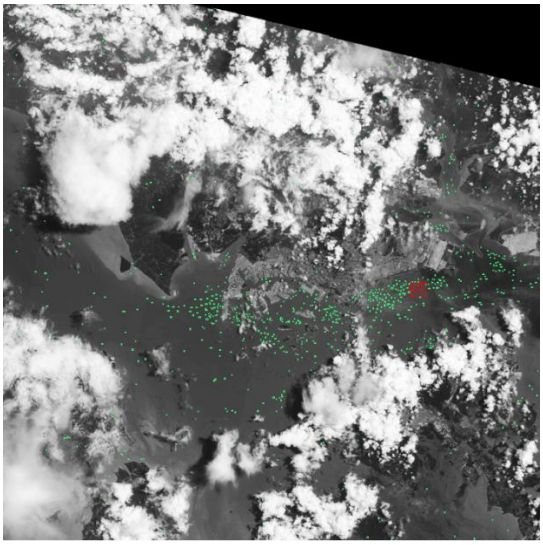
1. STORM Portal to visualize 2D data and to communicate with the end user (requests/results). It is based upon Angular 5, Openlayers 3 for the GIS component;

2. STORM Engine with:
   - Storm Feeder to ingest all sorts of data: EO data (Sentinel, Pleiades, COSMO-SkyMed, TerraSAR-X, RADARSAT-2, SPOT), AIS (NMEA/VDM), Vector Data (KML, shapefile), Raster Data (RNC S61, GeoTIFF, JPEG 2000),
   - Storm Catalog to index & store data, with respect to opensearch standard,
   - Storm Map to disseminate georeferenced data with respect to OGC standards (WCS/ WMS) and TMS,
   - Storm Processing to launch and to monitor the processing and to expose them with OGC WPS standard (*coming soon*);

3. STORM PaaS to host the whole software in a cloud agnostic architecture. It is based upon cutting edge open source technologies. Its architecture is cloud ready with **Marathon/Mesos** and Docker technologies. These frameworks allow to have a scalable & cloud agnostic platform: unlike HPC architectures which are static and need very performant nodes and networks to perform, cloud architectures are horizontally scalable, dynamic (it is easy to add nodes when needed), and only need basic server to perform. Last, STORM PaaS also benefits from the "Infrastructure As Code" technologies (Cobbler, Terraform, Ansible) to do the provisioning and the bootstrapping of the platform in premises or on a public cloud.

5. PROCESSING CHAIN

The processing chains developed for maritime security services leverage the availability of large and continuously growing amounts of images from satellite constellations such as Sentinel. Every image available over the area and period of interest of the user is downloaded and processed automatically in order to generate the value added products. In the case of object detection – in optical or radar images – the processing algorithms are based on recent machine and deep learning techniques developed for computer vision, precisely these algorithms make use of CNNs in order to classify and to localize the objects in the scene. The techniques and networks used in our services have been developed in-house or benefit from published algorithms such as YOLO, faster-RCNN ([4], [5])... The networks have been trained with a ground truth dataset composed of thousands of ship thumbnails stored in a homemade database populated from visual inspection of satellite images by our remote sensing analysts. The quality of the ground truth as well as the diversity of the objects present in the database allow to detect a large variety of ships, from fishing ships to large tankers, under various sea state conditions, from calm seas to rough conditions.



doi:10.2760/848593

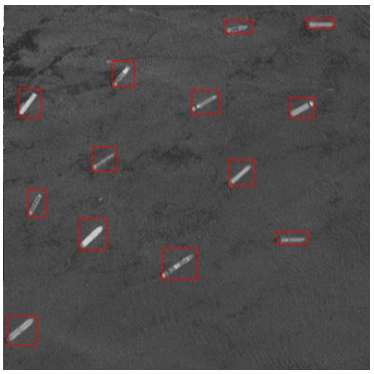


Figure 3. Automatic ship detection in a Sentinel-2 image acquired over the Singapore straight (top: entire tile showing detected ships in green; bottom: zoom on red square visible on top).

6. RESULTS

6.1. Infrastructure

This new architecture allows to deal with high volume of imagery, and process in real time worldwide AIS data thanks to the Elastic technology. The docker and cloud techniques have really disrupted the way algorithms can be developed, tested, deployed and scaled on the cluster; as a consequence the "commoditization" of the hardware reduces the global owning costs.
Moreover the infrastructure as code technologies has reduced with a scale factor the provisioning and bootstrapping of machines. It takes only 30 minutes to deploy the whole TPZF cluster in a standard configuration.

6.2. Algorithms

In the context of maritime security and surveillance, the object detection algorithms implemented in the TPZF services provide end users with reliable information on exclusive economic zone visits for fighting against illegal fishing. These algorithms have proven their efficiency in terms of accuracy and speed. Performances obtained on ship detection are better than 80% in Sentinel-2 optical images and about 95% in Sentinel-1 radar images. These performances – recall values – are really good since they have been compared with ground truth datasets including very small ships hardly visible in the images (especially the optical images) and under very different sea conditions. Processing times range between 45 seconds to 20 minutes depending on the implementation used. Detection of all the objects of interest, which are small objects compared to spatial resolution, in very large images and very high variety is still a challenge. AI is progressing continuously and better performances can be expected from improvement of algorithms, learning techniques and hardware.

7. CONCLUSION

TPZF has put in place an innovative solution to serve its own business and its customers. The methodology and the architecture allow to shorten the time to market. It also opens new fields and possibilities with smart processing on huge areas. Moreover, customers can benefit from near real time services thanks to the automation of image processing with AI algorithms. All this at lower cost due to OPEX reduction. Next steps are the enhancement of AI algorithms in terms of accuracy and speed, as well as the improvement of the architecture including the optimization of data storage and the addition of image streaming. Automation of the cloud bursting depending on customers' requests is also envisaged.

10. REFERENCES

[1] Gupta A., Tyagi S., Panwar N., Sachdeva S., Saxena U. 2018. NoSQL databases: Critical analysis and comparison. 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN).

[2] LeCun Y., Jackel L. D., Boser B., Denker J. S., Graf H. P., Guyon I., Henderson D., Howard R. E., Hubbard W. 1989. Handwritten digit recognition: Applications of neural net chips and automatic learning. IEEE Communication, pages 41-46. Invited paper.

[3] LeCun Y., Bottou L., Bengio Y., Haffner P. 1998. Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE, 86(11): 2278-2324.

[4] Shaoqing R., He K., Girshick R., Sun J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497.

[5] Redmon J. and Farhadi A. 2016. YOLO9000: Better, Faster, Stronger. arXiv:1612.08242.

[6] Di Francesco P., Malavolta I., Lago P. 2017. Research on Architecting Microservices: Trends, Focus, and Potential for Industrial Adoption. International Conference on Software Architecture (ICSA).

TOWARDS A HERITAGE MISSION VALORISATION ENVIRONMENT

Paulo Sacramento[1], *Giancarlo Rivolta*[2], *Joost van Bemmelen*[3]

[1] Solenix c/o European Space Agency (ESA-ESRIN), Frascati, Italy

[2] Progressive Systems c/o European Space Agency (ESA-ESRIN), Frascati, Italy

[3] European Space Agency (ESA-ESRIN), Frascati, Italy

ABSTRACT

In order to support users in extracting value from data from Earth Observation satellites and performing research involving trend-analysis over long periods of time, the European Space Agency has specified use cases and requirements and developed an architecture for a Heritage Mission Valorisation Environment (HM-VE). Heritage data are extremely valuable for long time-series studies such as those concerning climate change, due to their uniqueness, temporal coverage and the impossibility of collecting further observations back in time. The paper highlights how Heritage Mission data is relevant to the Big Data theme – particularly its Value, Variety and Volume - and then presents the main requirements for the environment, as well as the proposed architecture, emphasizing how ESA's long-term investment in the Research and Service Support service plays a role, particularly for the use cases related to hosted processing.

Index Terms **—** Heritage Missions, Time-series analysis, Valorisation, Research

1. INTRODUCTION

As part of its Heritage Data Programme (LTDP+), the European Space Agency has a mandate to promote the usage and exploitation to the maximum extent and in the easiest way possible of its Heritage Mission data holdings. Such mandate stems from the value perceived in said holdings by its Member States. Indeed, given that observations are, by definition, unique and that it is impossible to go back in time to take a snapshot of an area of particular interest, data from Heritage Missions such as ERS and ENVISAT are the only resource available to climate researchers and scientists wanting to understand changes and trends spanning several decades, starting from the 1990s and even before – the LTDP+ programme manages Third-Party mission data starting from the late 1970s. And this is in spite of limitations such as limited geographical coverage or revisit times and very low-resolution for today's standards – having such data enables scientists to perform analyses and compare their results over decades, not possible otherwise. Putting this in relation to the five Vs of Big Data – Velocity, Volume, Value, Variety and Veracity – it is evident that data from Heritage Missions has high Value, but it is also characterized by high Variety (multiple, heterogeneous missions and sensor types are in scope) and even Volume. In effect, although the total data volumes pertaining to Heritage Missions cannot be considered large for the current era – and this problem will be perennially smaller -, they are so in relative terms (full mission lifespan).

Without proper tools and services that can help mine and extract the value within the data, as well as putting it in context to the most recent one, there is a clear and large risk that this value is not adequately exploited. Several initiatives, ESA and otherwise, such as the Thematic Exploitation Platforms (https://tep.eo.esa.int/) or projects developed in the EU Framework Programme and H2020 context (https://ever-est.eu/, http://www.geoportal.org/) have addressed such issues targeting the missions and thematic areas most relevant to their project scopes, and have played an important role in advancing the technology required for this. Yet, relatively little focus has been dedicated specifically to the Heritage Missions.

In order to tackle this gap and pursue its mandate, ESA has gathered a set of use cases, specified a set of requirements and established an architecture for a so-called Heritage Mission Valorisation Environment (HM-VE). Valorisation is one of the five pillars of the LTDP+ Programme, which are introduced in the next section.

2. THE 5 PILLARS OF THE LTDP+ PROGRAMME

The Heritage Data Programme encompasses five strategic pillars that include all the functionality that needs to be covered by the LTDP+ supporting infrastructure. The figure below shows these five pillars: Preservation, Discovery, Access, Valorisation and Exploitation. Although the specified environment focuses on the Valorisation pillar, and some pillars like Preservation are only marginally or indirectly addressed, the set of use cases and requirements also covers aspects of Discovery, Access and Exploitation which are relevant for Valorisation.

Valorisation is explicitly distinguished from Exploitation on one hand to clearly focus the nature of the environment on the aspect of preparing data and extracting value from it, without actually exploiting it, on the other hand to highlight that the actual exploitation of the data is a higher-level concern, to be done downstream.

doi:10.2760/848593

FIGURE 1: THE 5 PILLARS OF THE LTDP+ PROGRAMME

3. SCOPE OF VALORISATION

The term "valorisation" may be subject to several interpretations. Therefore, it is important to define the meaning with which it is used, particularly when applied to Heritage Mission data and an environment for its valorisation.

First of all, it is considered that such a valorisation environment needs to encompass end-to-end functionality. Therefore, basic data and associated knowledge search and access use cases are included, even if those alone are not at the core of valorisation (they are founding pillars). However, it is possible that innovative ways of implementing search and access may add value in their own right, therefore contributing to the overall valorisation goal.

Secondly, valorisation is deemed to entail, at its simplest but most important level, some kind of data processing. In fact, it is assumed that most valorisation comes as a result of processing, which may be extremely simple - just deriving a single parameter from a product or generating a higher-level product using standard inputs - or extremely complex, consisting of the chaining of dozens of different algorithms, which more often than not will be in a state that renders integration difficult and requires substantial engineering work (it can be argued that in this case the valorisation will come not only from the end result – the processing chain – but also from the engineering work). Even a concept such as data fusion, which may not be understood as processing, involves and is made possible through processing (either the one which adapts the data - or extracts the part of interest - to enable fusion, or the one which takes different inputs to create a uniform output).

Finally, valorisation is also considered to include use cases such as quality-checking and quality-flagging (making such information available to the whole community), simple extraction, as well as new product and visualization formats generation. Such use cases, which can also be seen as forms of processing, are also contemplated.

4. MAIN USE CASES AND REQUIREMENTS

The figure below depicts how use cases have been organized. The logic is hierarchical, i.e. at the bottom level are the basic capabilities – search and access -, in the middle are use cases concerned with data preparation and processing, at the top are use cases concerned with visualization and valorisation, and at the highest level are the policy & decision making use cases, which represent the ultimate goal of the Heritage Mission data valorisation functionality, that is to enable policy and decision makers to act based on relevant, complete, accurate and large-scale information.

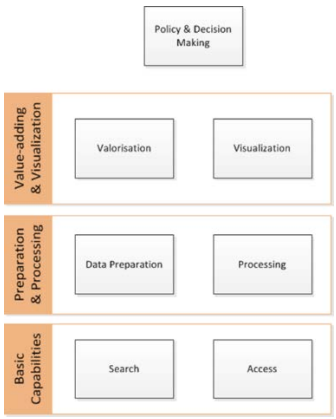


FIGURE 2: VALORISATION USE CASE ORGANIZATION

For reasons of brevity, the use cases and requirements related to search and access are not discussed in this paper and only brief information is provided about the other ones, by means of the table below. The table is not exhaustive but attempts to capture the most important use cases and requirements.

Use Case	Requirements
Evolution trends	• Integration of visualization tools with other functionality; • Visualization of evolution of parameters of interest over time, regardless of mission, on local and global scale.
Tool availability	• Visualization collocated with access/processing capabilities, to avoid need for separate tools.
Basic visualization	• Zoom in/out, sub-setting, clipping; • Access collocated with processing, to avoid download
Parameters of interest	• Adapt visualization mechanism to parameter being visualized (type, dimension, geographic and temporal granularity)
Primitive operations	• Apply interpolation and extrapolation across time (trend) and space (image); • Apply mean, difference, addition, superposition, correlation to relatable results
On-demand processing	• Browse available algorithms to run on-demand over data subset • Provide own algorithm to run on-demand over data subset
Large-scale processing	• Provide own algorithm to run systematically over data subset; • Make algorithm available to others

doi:10.2760/848593

Global processing	• Run algorithm on full geographical extent of the world
Quality checking	• Access data quality indicators; • Flag and validate data quality issues; • Add flagged and validated issues to data as metadata; • Track validation of flagged data quality issues
Simple valorisation	• Annotate data by adding/editing metadata
Valorisation by processing	• Add information produced or deduced by processing to data as metadata
New product generation	• Generate new data products (private or public); • Define new data products specifying starting inputs, aux data and processing algorithm
Data preparation	• Have tools available and use them to prepare and reorganize data for subsequent processing or advanced access
Climate analysis, Soil subsidence (as example)	• Search for parameters of interest regardless of missions and product types involved (local and global scale); • Extend support to new missions by configuration, taking advantage of similarities (different generations of same optical/radar sensors, common spectral bands);
Long time-series	• Pixel or n-dimension subset access (not only full products); • Explore pre-defined time-series of parameters of interest (last 1/3/6 months, last 1/5 years, last decade, maximum); • Configure new or edit existing time-series of parameters of interest;

TABLE 1: HM-VE MAIN USE CASES AND REQUIREMENTS

5. ARCHITECTURE

The figure below shows the conceptual architecture of the HM-VE, which organizes the functionality specified in the requirements into five functional blocks:

- DISC – Discovery: grouping all functionality relating to the discovery of data collections, data products, knowledge items (documents, videos, etc.) and services;
- ACC – Access: grouping all functionality that allows users to gain access to data collections (in the form of metadata), data products, knowledge items and services;
- VIZ – Visualization: grouping all functionality pertaining to the enhanced visualization of data products, long multi-mission time-series, climate phenomena, etc.;
- PREP – Data Preparation/Valorisation: grouping all functionality concerned with data preparation and value extraction, which is tightly related to processing;
- PRO – Data Processing: grouping all functionality that enables data to be processed. In general, this means hosted, not local processing, given the large timeframes, data volumes and geographical areas (e.g. global coverages) involved.

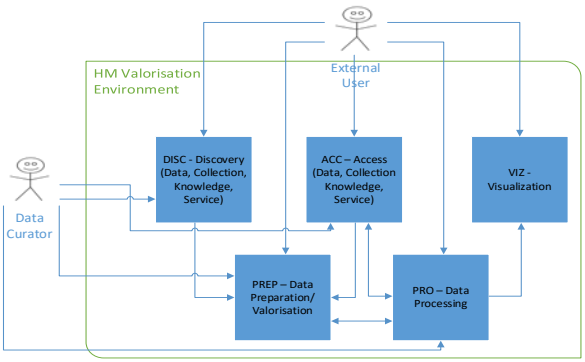


FIGURE 3: PRELIMINARY ARCHITECTURE OF THE HM-VE

Two different user types are considered – External Users and the Data Curator. The needs of the latter have higher priority.

6. RESPONSE TO BIG DATA CHALLENGES

As mentioned in the introduction, Heritage Mission data is characterized by high Value and Variety, two of the main challenges of Big Data.

The high Value stems from its uniqueness, temporal coverage and the impossibility of going back in time to gather further observations. This challenge can be addressed by providing users with simple and adequate search and access mechanisms, ways of extracting value from the data (preparation, valorisation and processing) and visualizing it. All this functionality is contemplated by the architecture introduced in the previous section.

The high Variety is a natural consequence of the global temporal span of the missions in scope – 40 to 50 years - and their heterogeneity - different sensor and product types, different instrument generations, different data rates, resolutions, swath sizes and geographical coverages. This is a difficult challenge to tackle but the proposed architecture takes this consideration on board by grouping the data preparation and valorisation functionality, making it a pre-requisite for further processing and advanced access mechanisms (e.g. datacubes). To be able to deal with such large Variety, the environment needs to ensure that data is previously examined and prepared, pre-processed, re-organized and re-formatted as necessary, both through manual steps and through dedicated software functionalities (e.g. a harmonization layer).

Very importantly, there is a specific requirement for the HM-VE that it shall be possible to extend support to new missions by configuration, taking advantage of similarities (different generations of same optical/radar sensors, common

doi:10.2760/848593

spectral bands). The ability to bring in new missions is clearly necessary because the set of missions managed by the LTDP+ programme is dynamic (they come into scope 5 years after conclusion of in-orbit operations). However, this means additional heterogeneity and Variety, which cannot be easily predicted beforehand, and thus requires the HM-VE implementation to be very flexible. It has to be able to register/import the new data and then allow users to include it as part of analyses across different missions, i.e. ensure it is seamlessly integrated into the existing user flows and interfaces.

7. IMPLEMENTATION STRATEGY

Having established an architecture, it should be mentioned that ESA has important starting constraints for the implementation of such an environment. First and foremost, the implementation strategy shall be based on heavy re-use of all existing ESA – and in particular the Earth Observation Programme directorate - infrastructure and services, which represent decades of ESA Member State investment. Apart from the obvious motivation of building on previous investment, there are two other reasons for this strategy: the first one is the desire not to "re-invent the wheel", as there is a wealth of know-how and technology within ESA that needs to be taken advantage of; the second one are the LTDP+ programme budget constraints, which prevent the undertaking of large implementation projects. It is, in fact, key to the success of the HM-VE that a proper gap-analysis is performed and an appropriate re-use strategy is employed.

The operations of the Heritage Missions managed by the LTDP+ programme currently rely on several elements (hardware/software) of the Earth Observation Multi-Mission Payload Data Ground Segment (EO PDGS). Whilst several elements of this PDGS can be no doubt employed as part of the HM-VE, there are also services funded by ESA that play a current role in these operations and should play a role in the HM-VE. One of those services is the ESA Research and Service Support (RSS) service ([1]), which since years supports EO researchers in the data processing algorithm development and integration phases as well as in running (hosted) processing campaigns using those algorithms. This process is actually key in the valorisation of Heritage Mission data ([2], [3]).

The ESA Research and Service Support (RSS) service has the mission to support the exploitation of Earth Observation data, by providing:

- flexible and customizable solutions useful during the algorithm development phase; and
- scalable cloud-based processing solutions that can be tailored and configured in accordance with the requirements defined by the user/stakeholder/project, for on-demand processing on limited datasets as well as for massive processing campaigns (e.g. decades of EO data).

Given its nature, RSS is particularly suited for the 'on-demand processing', 'large-scale processing', 'global processing', 'valorisation by processing' and 'new product generation' use cases and respective requirements from Table 1, among others. Currently, RSS services already respond to part of these requirements.

In particular, on-demand, large-scale and global processing related requirements are supported, as well as those concerning the generation of new products. As a matter of fact, existing RSS scalable solutions enable the EO user community on the one hand to run on-demand their own algorithms and on the other hand to run the same algorithms systematically over selected areas of interest (data subset). Further scaling-up to the global level is supported as well, thanks to the infrastructure independent RSS flexible service model virtually capable to resort to any capacity provider. Besides the direct use of own algorithms either for on-demand or for large/global processing, RSS users can make their algorithms available to others, and of course browse available (i.e. made available by others) algorithms and run them over selected data subsets.

Regarding the generation and the definition of new products, RSS users have the possibility to use available algorithms specifically designed for such aim. Alternatively, users can provide new algorithms, specifying as well inputs, auxiliary data, orchestration rules and other relevant information defining the desired output, and run them as needed to generate new products.

8. CONCLUSIONS

This paper has described on-going work at ESA to specify and implement a Heritage Missions Valorisation Environment, which will be realised by re-using existing ESA EOP infrastructure and services. Requirements and use cases have been gathered and an architecture proposed, placing high priority on the needs of the Data Curator user type. Implementation is planned to take place during 2019.

9. REFERENCES

[1] P.G. Marchetti, G. Rivolta, S. D'Elia, J. Farres, G. Mason and N. Gobron: "A Model for the Scientific Exploitation of Earth Observation Missions: The ESA Research and Service Support", IEEE Geoscience and Remote Sensing(162): 10-18, 2012

[2] M. Albani, J. van Bemmelen, G. Rivolta: "On the shoulders of giants: prototyping The HERO virtual research environment for data valorisation of heritage missions", Proceedings of the 2017 conference on Big Data from Space (BiDS'17), doi: 10.2760/383579: 214-216, 2017

[3] P. Sacramento, G. Rivolta, J. van Bemmelen: "ESA's Research and Service Support as a Virtual Research Environment for Heritage Mission data valorisation", Proceedings of the 2018 conference on adding value and preserving data (PV2018), RAL-CONF-2018-001, ISSN 1362-0231, 147-152, 2018

REGARDS – A GENERIC CATALOG ACCESS SYSTEM AND DATA VALORIZATION TOOL

Claire Caillet, Benoît Chausserie-Laprée, Dominique Heulet

CNES, 18 av E. Belin, 31401 Toulouse Cedex 9, France

ABSTRACT

For CNES archives, SITools2 and SIPAD-NG are the current two main systems used to manage space mission data. However, the architecture of these tools are now becoming obsolete due to the new needs of long term data preservation: cloud-type architecture to handle the amount of data and performances of archives functions, open data policy to implement some interoperability standards.

The new system REGARDS (REnewal of Generic tools to Access and aRchive Space Data) has been developed to address these new needs to merge the functions of SITools2 and SIPAD-NG.

In this paper we will present successively: the context of the CNES archives, the REGARDS characteristics (functions, architecture, framework), the main OAIS (Open Archive Information System) functions as implemented by REGARDS (ingest and storage, catalog, access), the deployment aspects and the short term planning.

Index Terms— Preservation, Long-term archive, OAIS, Web services, Web architecture, Interoperability, Scalability

1. CONTEXT

CNES manages a large variety of space missions, addressing various topics from Earth Observation to Astronomy, as well as physical sciences and technology. Data produced during these missions can be processed, archived and distributed either by CNES or by scientific laboratories.

CNES or partners Data Centers are responsible for the long-term preservation of all the data produced by these missions, see Ref [1] for more details.

These Data Centers (CDPP : French Data Center for space Plasma Physics), MEDOC (Data Center for Solar Physics Data), CADMOS (Data Center in the microgravity domain), AVISO (Data Center for altimetry missions), SERAD (Data Centre for other space missions) are based on two main tools (SITools2 (http://adsabs.harvard.edu/abs/2012ASPC..461..821M) and SIPAD-NG (http://vds.cnes.fr/VDS-Sipad.html)) which are becoming obsolete.

REGARDS is the next generation of Catalog Access Systems. It merges the functions of the SITools-2 and SIPAD-NG tools (see Ref [2]). The targeted main objectives of REGARDS are the following:

- Have a unique tool to optimize development and maintenance costs,
- Be able to cope with huge data volumes expected from space missions in the 2020 and beyond (see Figure 1),
- Address the interoperability needs,
- Meet the need to bring the processing as close as possible to the data,
- Make it an open source software (under GPLv3 license),
- Get a true ground segment product with high capabilities of configuration and adaptability aiming to be implemented in Mission or Data Centers located at CNES or in partner laboratories
- Be compatible with a cloud-type architecture

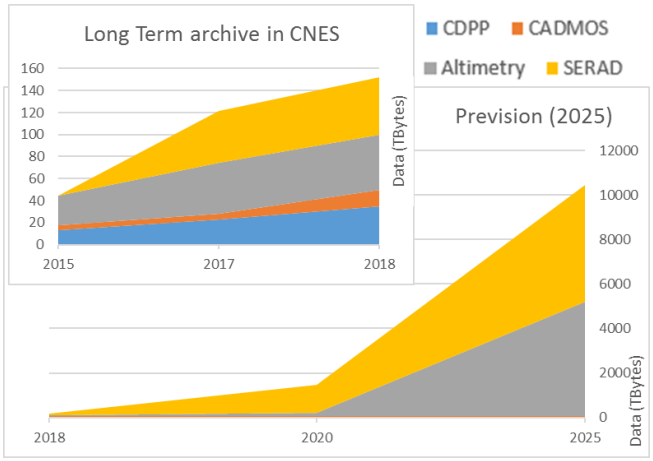


Figure 1: Archive evolution in CNES

REGARDS will also contribute to better implement the FAIR principles for CNES Archives:

- Findable: better referencing of CNES archives
- Accessible and Interoperable: pertinence of search and selection tools, compatibility with standard protocols of interoperability
- Reusable: standard metadata descriptions, documentation, services (visualization, web processing services)

2. REGARDS FUNCTIONS

REGARDS main functionalities rely on the OAIS model for long-term preservation and access to digital data, as shown in Figure 2.

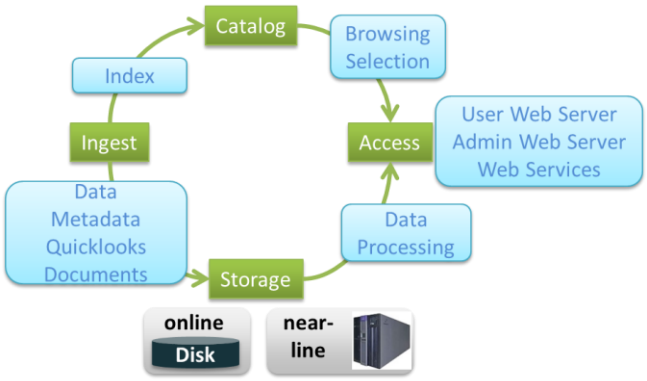


Figure 2: REGARDS functions

In addition to the long term storage capability of data, REGARDS also offers functionalities:

- To facilitate the system administration in a simple and ergonomic way.
- To launch back-end processing through standard protocols (WPS: https://www.opengeospatial.org/standards/wps, UWS: http://www.ivoa.net/documents/UWS/20161024/index.html).
- To be supervised.

3. REGARDS ARCHITECTURE

REGARDS relies on a web-oriented architecture. The 'frontend' (client application) has access to the 'backend' (services) through a gateway. The 'backend' is composed of several Java micro-services. Each micro-service is a web server providing REST endpoints (Figure 3).

Each micro-service matches an elementary REGARDS function (single accountability) and has its own context of execution and its own configuration (modularity). It is built, tested and deployed separately from other services (modularity, serviceability, scalability) and provides a REST API which relies on a service contract.

Such an architecture enables REGARDS to allocate at best the needs in terms of horizontal scalability: several sessions of the same service can be simultaneously deployed on the System, allowing it to bear load spin-up thanks to a "load balancing" mechanism. Such a functioning allows the System to be highly fault-tolerant and enables REGARDS to absorb the different mission loads. It allows high-performance ingestion of data and metadata available from providers, either in standard or non-standard formats. In this way, it allows to valorize metadata through upgrade and to provide them to the community thanks to standard protocols of interoperability and advanced search interface (Open search with the Geo and Time extensions and the parameter extension, Faceted search, multiple geospatial projection support).

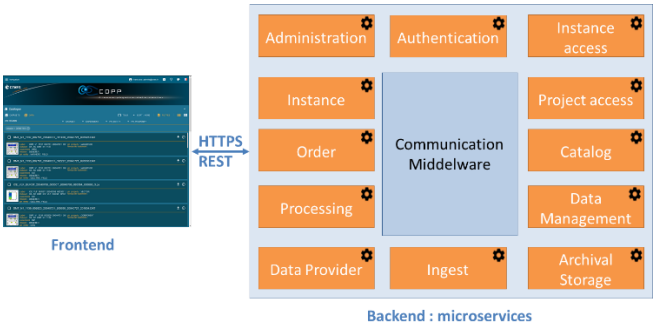


Figure 3: REGARDS architecture

4. REGARDS FRAMEWORK

REGARDS is a highly customizable and adaptable framework:

- The administration web GUI (Graphical User Interface) provides high customizable capabilities (data model definition as in Figure 4, ingest and storage configuration, user interface configuration, …)

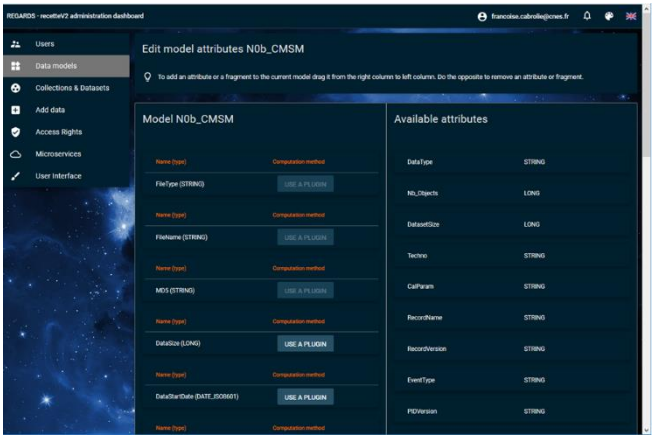


Figure 4: REGARDS administration GUI

- Plugins are used to adapt each component (micro-services or frontend) to the characteristics of the mission or data center (see Figure 5). For instance, a plugin can be developed to transform data and provide it in the expected format of the Ingest Micro-service. Another example, is to build a plugin to expose the catalog interface in ISO19115 to allow the harvesting from other earth observation data catalogs.

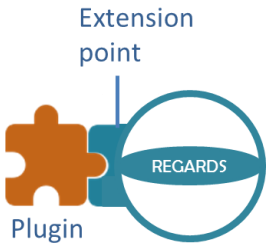


Figure 5: REGARDS plugins

- Extensions can be created to provide a new function to REGARDS (see Figure 6). An extension will expose new REST web services which would be aggregated to the public API.

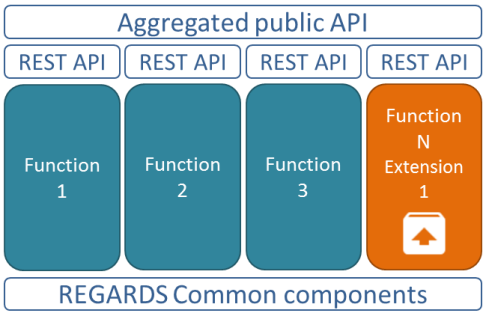


Figure 6: REGARDS extensions

5. INGEST AND STORAGE

The backend micro-services are constituted by chaining steps, where each step is implemented by a plugin. The Figure 7 shows this principle.

The "Data provider" micro-service transforms data provided by an external service into a Submission Information Package (SIP).

The "Ingest" micro-service transforms a SIP into an Archive Information Packages (AIP).

The "Storage" micro-service stores AIP.

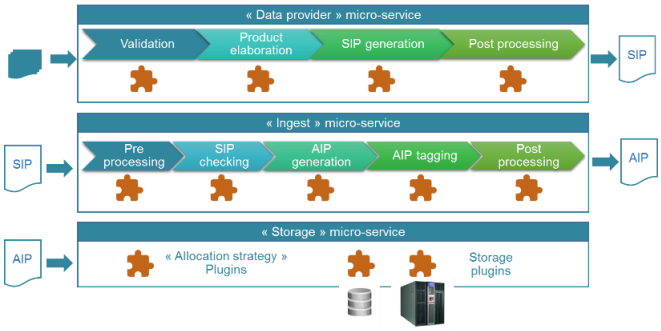


Figure 7: Ingest and Storage functions

6. CATALOG

The Catalog is based on an Elastic Search database which can be fed by two kinds of data sources:

- Internal "data sources"
- External "data sources".

Each data source is indexed and made searchable through dedicated plugins via "Data management" and "Access" components (see Figure 8).

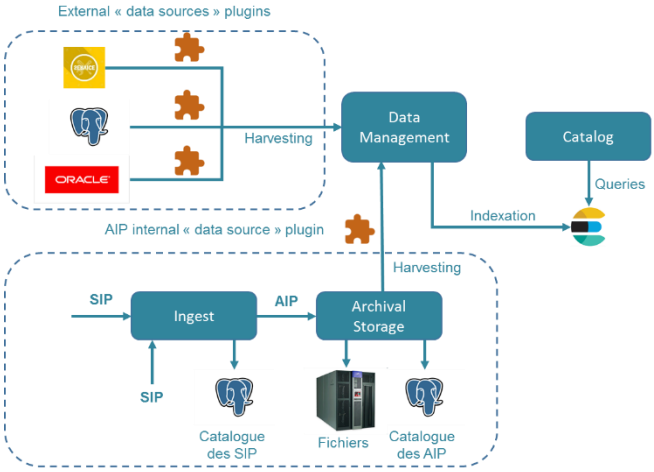


Figure 8: REGARDS Catalog

7. ACCESS

The access function includes three parts: user interface configuration, interoperability access and data ordering.

User Interface configuration:
For each mission to be archived with REGARDS, the data (products) to be archived and distributed, criteria to search available data for users and web GUI look and feel need to be defined. This process is iterative with the project team and the REGARDS team specialized in long term archiving.

This process is facilitated by the structuration in modules. Each module is a plugin: the construction of the user interface is done without any development, directly by assembling modules.

Data ordering:
The ordering (order micro-service) provides access to the data either through the local archive or through the external "data sources" (URL or web services processing). The data download can be synchronous or asynchronous (ie. with the possibility to process the data (processing micro-service) or in case of high volume to download it as a background task).

Dedicated plugins can be developed depending on the exposed services of the "data source".

Interoperability access (example of MIZAR cartographic component):
REGARDS cartographic component for visualization is the OpenSource product MIZAR (https://github.com/MizarWeb).

MIZAR can be embedded in the REGARDS GUI and be connected to REGARDS catalog using the exposed web services (Opensearch web services). Doing so, the users can research data from REGARDS using MIZAR GUI and its cartographic search capabilities.

Selected products can then be ordered (to be downloaded) by users through MIZAR, using the order capabilities of REGARDS.

MIZAR is able to interface several catalogs and could be used to provide search capabilities into various catalogs including different REGARDS projects or external catalogs.

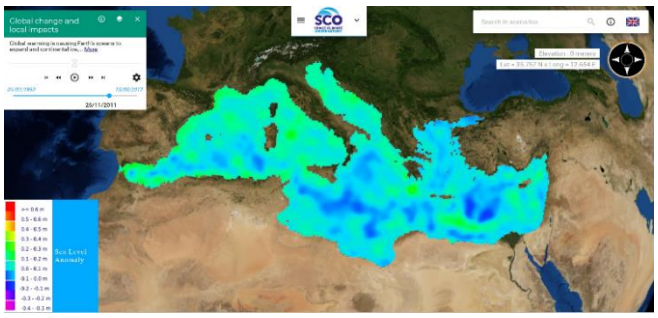


Figure 9: MIZAR Interface (SCO - Space Climate Observatory demonstrator)

8. REGARDS DEPLOYMENT

REGARDS can be deployed on one single server or it is possible to separate the deployment of its components on several servers. It is also possible to deploy only a part of the micro services.

REGARDS is compatible with a cloud-type architecture, based on virtual machine servers: each micro service can be deployed in one or more instances on one or more servers, for instance:

- Ingestion, ordering and restitution services on one server
- Data access service on another server
- COTS (Rabbit MQ, Elastic Serarch) on a last server

REGARDS is deployed using the IZPack tool and the installation can be done with an MMI or in full command line mode. This mode will allow the deployment of REGARDS to be encapsulated and to be fully automatic. With this objective in mind, the deployment using Ansible or Docker, will be available in the future versions of REGARDS.

Sources and documentation are available here https://github.com/RegardsOss. All APIs provided by REGARDS are available, and the documentation will provide tutorials for developers who would like to develop plugins and integrate them

9. DEPLOYMENT AND MIGRATION PLANNING

All missions currently using the SIPAD-NG will be migrated to REGARDS. In addition, new CNES projects will use REGARDS to archive and distribute their data. Hereafter are the milestones of REGARDS development, migrations and new deployments:

- Nov 2015: development start
- Oct 2018: REGARDS V3 including all main functionalities (Ingest, Catalog, Storage, Access)
- 2019: deployment at ONERA for MICROSCOPE mission data
- Mid-2019: REGARDS V4
- 2020: all migrations finished (CDPP, SERAD, CADMOS, SMOS Ifremer, Altimetry (SSALTO, CFOSAT), IAS Orsay),
- 2020: Deployment for SPOT (SPOT World Heritage)
- 2021: Deployment of MICROCARB and SWOT catalogs
- 2022: Deployment for PLEIADES World Heritage

10. REFERENCES

[1] Evolution of CNES Tools and Processes for Long Term Preservation of Space Science Data, PV 2018, RAL-CONF-2018-001

[2] REGARDS: the new CNES generic system to access and archive space data, PV 2015 Conference

[3] CCSDS The Consultative Committee for Space Data Systems, “Reference Model for an Open Archival Information System (OAIS)”, *CCSDS publication* [online database]

[4] OGC OpenSearch Extension for Earth Observation, docs.opengeospatial.org/is/13-026r8/13-026r8.html

[5] OpenSearch Parameter Extension, http://www.opensearch.org/Specifications/OpenSearch/Extensions/Parameter/1.0

DISTRIBUTING BIG ASTRONOMICAL CATALOGUES WITH GREENPLUM

Pilar de Teodoro, Juan González, Sara Nieto, Jesús Salgado

European Space Astronomy Centre, Madrid, Spain

ABSTRACT

When there is no option to continue scaling up resources, there is a need for scaling out. At the ESAC Science Data Centre (ESDC) we envisage a growth of the archive data stored in operational databases from 25TB up to 50TB in 3 years. The current technology used, which is an open source relational database system, vanilla PostgreSQL will not be enough to keep the data and maintain good performances. In order to fulfill the user requirements for the different missions with such big amounts of data, and demanding heavy queries, distributed databases will be necessary. After testing other flavours of distributed PostgreSQL such as Citusdata and Postgres-XL, we investigated Greenplum in its commercial and open source flavours. This paper completes a number of tests performed with Gaia DR1, DR2 catalogues and a Euclid simulated catalogue with the aim to check the feasibility of the solution.

***Index Terms*—**ESDC, database, PostgreSQL, Greenplum, distributed databases, scale-out solution

1. INTRODUCTION

At ESDC there is a need to evaluate if all our databases could run on a single platform and allow joins between different missions catalogues. It is not only the size of the data on the database but also how it is queried. Partitioning data on a single machine may be not enough and running queries on multiple nodes multiplying the number of cores will be necessary. The investigation of distributing databases with Greenplum at ESDC started in 2012 in Amazon EC2 for the Gaia mission archive. A scale-up solution was chosen at that moment due to the size of the first data releases. We have retaken the investigation with the Greenplum improved version in 2018. Previously and as reported for the BIDS'17, we completed the study for two other flavours of PostgreSQL distributed databases: Postgres-XL [1] and Citus data [2]. The three solutions are grounded on a common base, one master node, which orchestrates and several nodes where the data resides in PostgreSQL instances and which provides the possibility to scale the size of the database by adding more nodes. All of them use massively parallel processing (MPP) techniques. Tests on Postgres-XL and Citus were reported in [3] nevertheless we introduce the basics for understanding the three solutions explored in the following section. Also at the end of this paper we will make a comparison summary. In the other sections an overview of the tests run on Greenplum are explained with their conclusions.

2. SOLUTIONS EXPLORED

2.1. Greenplum

The open source and commercial Greenplum Database (GPDB) cluster consists of a master node and segment nodes. All of the data resides on the segment nodes and the system catalogue information is stored in the master nodes. Segment nodes run one or more segments, which are modified PostgreSQL database instances and are assigned a content identifier. For each table the data is divided among the segment nodes based on the distribution column keys specified by the user in the data definition language. When a query enters the master node, it is parsed, planned and dispatched to all of the segments to execute the query plan and either return the requested data or insert the result of the query into a database table. The Structured Query Language, version SQL:2003, is used to present queries to the system. Transaction semantics comply with constraints known as ACID (atomicity, consistency, isolation, and durability).

2.2. Postgres-XL

This open source solution is currently supported by the company 2ndQuadrant. The architecture counts with one or several coordinators (entry points) and one or more datanodes where the data is distributed. Database tables can be created in the datanodes specified, not necessarily in all of them, which allows distributing the data load. On top of that, having several coordinators increases the number of connections to the cluster or the possibility to connect to the local database. Tables can be replicated in all the datanodes with the purpose of making joins more optimal as the join is done locally. When you issue queries, Postgres-XL determines where the target data is stored and dispatches corresponding plans to the servers containing the target data. To keep track of the transactions a global transaction manager (GTM) provides unique and ordered transaction id to each transaction running on Postgres-XL servers. GTM is a single point of failure and can cause bottlenecks due to the serialization of the transactions.

doi:10.2760/848593

2.3. Citus

Citus is a PostgreSQL extension that allows commodity database servers (called *nodes*) to coordinate with one another in a "shared nothing" architecture. The nodes form a cluster that allows PostgreSQL to hold more data and use more CPU cores than would be possible on a single computer. This architecture also allows the database to scale by simply adding more nodes to the cluster. Citus architecture is based on a coordinator (in old versions called master) server and one or more worker nodes. Applications send their queries to the coordinator node, which relays it to the relevant workers and accumulates the results. Every query is either run on a single node, or in several, in a parallel way.

3. TESTING WITH GREENPLUM

We have fulfilled different tests to check the GPDB solution: ingestion, compression, high availability, performance and scalability.

3.1. Testing Architectures

The recommended architecture by Greenplum is bare metal servers with a shared-nothing architecture and local storage in each node with the possibility to use SSD or NVMe storage. It was not possible to use this architecture for the proof of concept as SSDs were not available, but we could compare two different architectures.

3.1.1. Private Cloud

The test performed ran on a private cloud with 1 master node and 6 8-core datanodes (4 segments on each) with 32GB RAM each, the storage used in NetApp [4] shared by all nodes for the binary files and different mounted volumes for supporting data files. The IO in this volume for writing was about 180MB/s and for reading 84MB/s.

3.1.2. Bare Metal

The cluster is composed of 1 master node and 5 48-core data nodes with 256GB RAM. GPDB binaries were installed locally in each node. 1TB SAS drive was allocated as local volume but the performance was poor, the IO was about 80MB/s for writing, it is a single drive so it could not benefit from the writing and reading parallelism as several segments are running on the same volume and the spindle needs to access different sectors on the disk, blocking the readings and writings. Then we installed the cluster using NetApp as storage, using different NetApp volumes for each node going to 475MB/s for writing.

3.2. Ingestion on different table formats

Greenplum supports different methods for ingesting data into the cluster database: from file, gpload and gpfdist, which is a parallel file distribution program. We tested the ingestion with this tool, reading from csv files of the flagship_mock_1_5_2_s astronomical simulated catalogue from the Euclid mission with a total of 2739541922 astronomical sources (rows) and describing the data in 118 columns, which occupies in csv format 3.1TB and 1/3 in the database. An external table was created for reading the data from the gpfdist servers and then ingested into the corresponding table. A summary of the tests run on the private cloud architecture with different numbers of segments to check scalability is presented below. The goal was to choose the more suitable format for our use case as we make large ingestions only when new catalogues are released.

Format	Segments	gpfdist servers	Time (h)	MB/s
Heap table	2	1	8	104
Heap table	12	1	5.5	156
Heap table	24	1	7	123
Columnar	24	6	5.5	156
Append only	24	6	3	287

Append only format is the fastest way to ingest data, in this case this ingestion eliminates the storage overhead of the per-row update visibility information, saving about 20 bytes per row. This allows for a leaner and easier-to-optimize page structure. The storage model of append-optimized tables is optimized for bulk data loading such as gpfidst with external tables but not for single inserts.

3.3. Compression

Greenplum offers different compression types with different compression levels. ESDC data in databases will grow, therefore it is necessary to check if it is worthy to use compressed data as CPU time and resources are used for the compressed and uncompressed work. For the proof of concept, the compression of zlib and quicklz [5] were tested, creating a compressed table and populating it from the heap format one. Compression rates and time to compress for the Euclid simulated catalogue is shown below.

Compression lib	Time to compress 1.3TB	Compression Ratio
Zlib-level5	1h 28 min	1.27
Quicklz	1h 58min	1.07

Performance was not better than on the uncompressed tables except for the count with quicklz, which was faster.

In our case compressing the data, which is "hot", will not be worthy but for keeping "old" data can be an advantage.

Some query result time values are shown in the performance and scalability section.

doi:10.2760/848593

3.4. High Availability (HA)

We have evaluated the option of creating standby instances for the master and every segment in our cluster. Greenplum Database high availability works using standby replication for each node. For the tests performed, the standby master was created after the primary cluster was created and running. The steps were done following the Pivotal documentation to initiate the standby master and later adding the standby segments. When a primary segment goes down, the mirror activates and becomes the primary server. After recovering the failed node, the mirror remains the primary and the failed segment becomes the mirror. This can be later reversed and returned to first preferred roles. On the other hand, for the master recovery of a failure it is not possible to recover to the preferred role. In that case it is necessary to create a new standby and activate again in the old hostname+port. Once this is done, a new standby for the recovered master must be created. All these steps have been tested in our setup. It is important to remember that double the space is necessary when mirroring is configured.

3.5. Performance and Scalability

Tests were run on the two different architectures:
-Private cloud cluster: Testing from 4 datanodes with 4 segments each to 6 datanodes with 4 segments, making a total of 24 segments. Expanding a 4.5TB database cluster from 3 to 6 datanodes took 15 hours.
-Bare metal cluster: we could scale from 3 datanodes with 4 segments each to 5 datanodes with 4 segments on each (6 hours to expand) and then later to 8 segments on each for a total of 40 segments (1 hour to expand).

For expanding, some downtime has to be expected, as it will affect the whole cluster. This point may be a caveat. We could never reshuffle data with this volume using Postgres-XL. A query profiling exercise was performed for Gaia catalogue data [6] and Euclid catalogue data [7] to compare different options: different table formats and different number of segments. The tables were distributed on their primary key making sure that the skew was minimum so every segment contained approximately the same number of rows. The distribution of the Gaia and Euclid catalogues is done by the primary key column. For understanding purposes here is the number of rows and size of the tables presented in this paper:

Table	Rows number	Size in DB
Euclid.flagship_mock	2739541922	1,3TB
Gaiadr2.gaia_source	1692919135	862GB
Gaiadr1.gaia_source	1142679769	301GB

Q3C[8] and pg_Sphere[9] PostgreSQL extensions were installed on the database to allow geometrical queries. Some representative examples of the different queries executed are presented in the following sections.

3.5.1. Count

Counting the number of rows in the flagship simulated catalogue for Euclid:

```
select count(*) from euclid.flagship_mock;
```

This full table scan shows that the time depends on the format of the table and also on the number of segments advancing that columnar format is the fastest in our configuration and that increasing the number of segments also scales.

Bare Metal	Nodes/ segments	Format	Compression/ DB	Time (s)
N	2/8	heap	No/GPDB	8055
N	6/24	heap	No/GPDB	1808
N	6/24	columnar	No/GPDB	33
N	6/24	AO	No/GPDB	1693
N	6/24	columnar	zlib-5/GPDB	378
N	6/24	columnar	quicklz/GPDB	22
N	10	heap	No/Postgres-XL	9324
Y	3/12	AO	No/GPDB	2567
Y	5/20	AO	No/GPDB	1308
Y	5/40	AO	No/GPDB	1077
Y	5/40	columnar	quicklz/GPDB	14

3.5.2. Query on Gaia DR2

The following query will retrieve all sources lying within a 3 degree radius from the given coordinates (ra=34.7, dec=57), having parallax and proper motion in both ra, dec directions larger than 0 and limiting the number of rows to retrieve to 500000.

```
select source_id, ra, dec , parallax, parallax_error , phot_g_mean_mag, phot_bp_mean_mag , phot_rp_mean_mag
 from gaiadr2.gaia_source
where (q3c_join(34.7,57,"ra","dec",3))= '1' and ("parallax" >= 0.0 and "pmra" >= 0.0 AND "pmdec" >= 0.0) Limit 500000;
```

The query execution plan shows that using the legacy Greenplum optimizer, a bitmap index scan is done on q3c_ang2ipix(ra, "dec"), with the bare metal configuration giving better results.

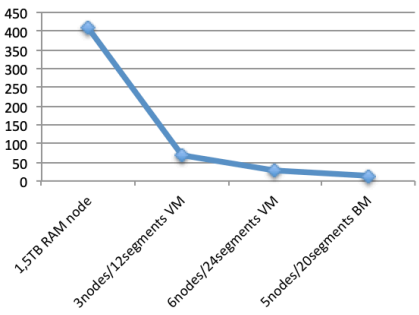


FIGURE 1: SETUP VS TIME (S)

3.5.3. Crossmatch Query on Gaia DR2 with Gaia DR1

One of the challenges proposed was the possibility to perform crossmatches between large catalogues, which is a desired capability by the scientists, but difficult to achieve with the current hardware limits. Crossmatches will allow identifying when a source in one catalogue is the same in another, and that as a first approach is used to identify close-neighbours between catalogues as possible candidates, who will be refined later with, e.g. astronomical emission models. In this case the crossmatch is done on gaia_source from data release 1 and data release 2.

```
SELECT gdr1.source_id AS iddr1, gdr2.source_id AS iddr2
FROM gaiadr1.gaia_source AS gdr1, gaiadr2.gaia_source AS gdr2
WHERE q3c_join(gdr2.ra, gdr2.dec, gdr1.ra, gdr1.dec, 0.5/3600)
```

This query takes about 5.5 hours to finish in a 1.5TB RAM machine using SSDs for storing the catalogues. Distributing this data in our POC infrastructure for bare metal took about 6 hours in the attached storage. This result proves that we can scale with smaller nodes but that we need a master node with enough memory, about 56GB at least.

4. CONCLUSION

As the technology evolves, the satellites are able to send more data and more data need to be selected to get maximum benefits for understanding. SQL databases give this chance to get the information required directly from the data but hardware and databases needs to evolve hand in hand. Tests performed inferred that Greenplum is more mature as a distributed database but still some important issues are not covered as the integration with latest versions of open source PostgreSQL, which limit us in using some already implemented functionality for vanilla PostgreSQL and the possibility of using the commercial Greenplum, which is currently under evaluation. The main difference with the open source version is the dedicated support and the quicklz compression license among others. With the tests performed we have proven that is possible to scale-out and keep or improve performance with the current bare metal architecture tested. Finally, the following table shows a comparison among the three distributed solutions studied.

Solution	**Greenplum**	**Postgres-XL**	**Citus**
Support Company	Pivotal	2ndQuadrant	Citusdata
Last version released*	5.6.15	9.5	8_11
Tested version	5.6.14	10 alpha	7_10
PostgreSQL version	8.3	9.5	10
Allow multiple coordinators?	No	Yes	Yes
Data nodes Mirroring?	Yes	Yes	Yes
Possibility to work on individual nodes?	No	Yes	No
Foreign Data Wrappers allowed?	PXF tools	No (in development)	No
Is it a PostgreSQL fork?	Yes	Yes	No
Commercial support offered	Yes	Yes	Yes
SQL, ACID	Yes	Yes	Yes
Open source community	Yes	Yes	Yes
Support answering speed	Ongoing	Slow	Very slow
Performance on ESDC testing	Good	Good but bug found	Poor
Expansion-reshuffling (3TB)	Good	Bad – crashed	-
Ingestion	Parallelize with GPFDIST	Use multiple coordinators	-
Monitoring tool	Yes (commercial)	No	No

5. REFERENCES

[1] P. Teodoro. Distributing postgres with Postgres-XL for very large astronomical databases. Pgconf.us 2018: https://postgresconf.org/conferences/2018/program/proposals/distributing-postgres-with-postgres-xl-for-very-large-astronomical-databases

[2] C. Kerstiens. How to horizontally scale your Postgres Database using Citus: https://dzone.com/articles/how-to-horizontally-scale-your-postgres-database-using-citus

[3] P. Teodoro, S. Nieto, J. Salgado, C. Arviset. Considering scale out alternatives for big data volume databases with PostgreSQL. Big data from space conference. 2017 p.193: https://publications.europa.eu/s/gPYM

[4] NetApp: https://www.netapp.com/

[5] Zlib and Quicklz compression libraries: https://www.quicklz.com/bench.html

[6] T.Prusti et al. 2016. The Gaia Mission: http://dx.doi.org/10.1051/0004-6361/201629272

[7] M. Poncet, C. Dabin, J.-J. Metge, K. Noddle, M. Hollimann, M. Melchior, A. Belikov, and J. Koppenhoeffer, "Euclid: "Big data from dark space" – Science ground segment challenges for next decade," in BiDS, 2014.

[8] Koposov, S., Bartunov, 2006. Q3C, Quad Tree Cube – (Cone Search and Xmatch) in Open Source Database PostgreSQL, in: ADASS XV, p. 735.

[9] I. Chikingarian, O.Bartunov. PostgreSQL: the suitable DBMS solution for astronomical databases. ADASS XIII, p.227.

Author Index

Abstract

Big Data from Space refers to the massive spatio-temporal Earth and Space observation data collected by a variety of sensors - ranging from ground based to space-borne - and the synergy with data coming from other sources and communities. This domain is currently facing sharp development with numerous new initiatives and breakthroughs from intelligent sensors' networks to data science application. These developments are empowering new approaches and applications in various and diverse domains influencing life on earth and societal aspects, from sensing cities, monitoring human settlements and urban areas to climate change and security.

The goal of the Big Data from Space conference is to bring together researchers, engineers, developers, and users in the area of Big Data from Space. It is co-organised by ESA, the Joint Research Centre (JRC) of the European Commission, and the European Union Satellite Centre (SatCen). The 2019 edition of the conference was hosted by the German Aerospace Center (DLR) and held in the Alte Kongresshalle of Munich (Germany) from the 19th to the 21st of February 2019.

These proceedings consist of a collection of 75 short papers accepted for oral or poster presentation at the conference as a result of the peer-review process by the conference programme committee. The papers are lined up around the topics matching the oral sessions as well as the poster session, also organised by topics. These contributions provide a snapshot of the current research activities, developments, and initiatives in Big Data from Space.

This 4th edition of the Big Data from Space conference is directed towards 'Turning Data into Insights'. Indeed, while the first editions of the conference concentrated on technologies and platforms capable of sustaining the sharp increase of data streams originating from space sensors, the development of efficient and effective methodologies and algorithms capable of extracting insights from these data is gradually becoming the main challenge. In this context, artificial intelligence and machine learning techniques have started to play a key role as illustrated by numerous papers of this conference edition. Methodological developments are motivated by the pressing need to extract information on large areas and/or over long time series to better understand the dynamics of the processes that are shaping our planet and indeed our universe in the case of data collected by telescopes. The topic of analysis ready data has also emerged since the last edition and is closely linked with the development of new data cube representations. Big data from space is also introducing some new legal challenges and the need for further developments of standards and interoperable interfaces between the growing number of platforms hosting multi-petabyte scale data co-located with processing capabilities. All these topics as well as other generic key aspects of big data are mirrored in dedicated sections of these proceedings.

KJ-NA-29660-EN-N

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

Joint Research Centre

EU Science Hub

Publications Office

doi: 10.2760/848593

ISBN 978-92-76-00034-1