

## TOWARDS A JRC EARTH OBSERVATION DATA AND PROCESSING PLATFORM

*P. Soille<sup>1</sup>, A. Burger<sup>2</sup>, D. Rodriguez<sup>1</sup>, V. Syrris<sup>1</sup>, and V. Vasilev<sup>2</sup>*

European Commission, Joint Research Centre (JRC)

<sup>1</sup>Institute for the Protection and Security of the Citizens, Global Security and Crisis Management Unit

<sup>2</sup>Institute for Environment and Sustainability, Digital Earth and Reference Data Unit

### ABSTRACT

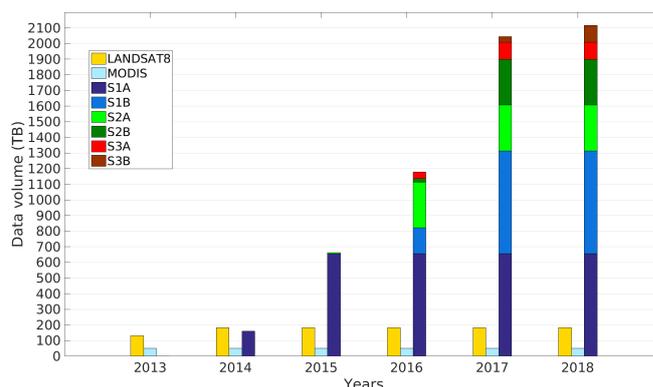
The Copernicus programme of the European Union with its fleet of Sentinel satellites operated by the European Space Agency are effectively making Earth Observation (EO) entering the big data era. Consequently, most application projects at continental or global scale cannot be addressed with conventional techniques. That is, the EO data revolution brought in by Copernicus needs to be matched by a processing revolution. Existing approaches such as those based on the processing of massive archives of Landsat data are reviewed and the concept of the Joint Research Centre Earth Observation Data and Processing platform is briefly presented.

**Index Terms**— Earth Observation, Sentinel, Copernicus, Infrastructure

### 1. INTRODUCTION

To date, the United States (U.S.) Government is the largest provider of environmental and Earth system data in the world<sup>1</sup>. A first data revolution happened in 2008 when the U.S. Geological Survey decided to release for free to the public its Landsat archive which is the worlds largest collection of Earth imagery [11]. Still, the European Commission, with its ambitious Copernicus programme and associated Sentinel missions (S1 to S6 satellite series) operated by the European Space Agency and complemented by a range of contributing missions, is on the way to become the main provider of global EO data with a free, full, and open access data policy. With expected data *volumes* of 10 TB per day (when *all* Sentinel series will reach full operational capacity), data *velocity* highlighted by the production of global coverage with repeat time as short as 2 days for Sentinel-3, and data *variety* resulting from sensors in the optical and radar ranges at various spatial, spectral, and temporal resolutions, the Copernicus programme is a game changer making EO data effectively entering the big data era [10]. Figure 1 shows the overall estimated data throughput for the Sentinel 1–3 missions compared to those delivered by the Landsat 8/MODIS satellites.

<sup>1</sup><http://dels.nas.edu/resources/static-assets/besr/miscellaneous/Stryker.pdf>



**Fig. 1.** Yearly data flow estimates from Sentinel 1–3 (assuming full operational capacity) compared to MODIS and Landsat 8 data flows.

Whilst the European Union (EU) is making EO entering the big data era in terms of data production, innovative developments need to be pursued to fully exploit the potential of the generated data whether for academic, institutional, or commercial applications. This also applies to the Joint Research Centre where the current fragmented approach of EO data storage and processing is no longer sustainable.

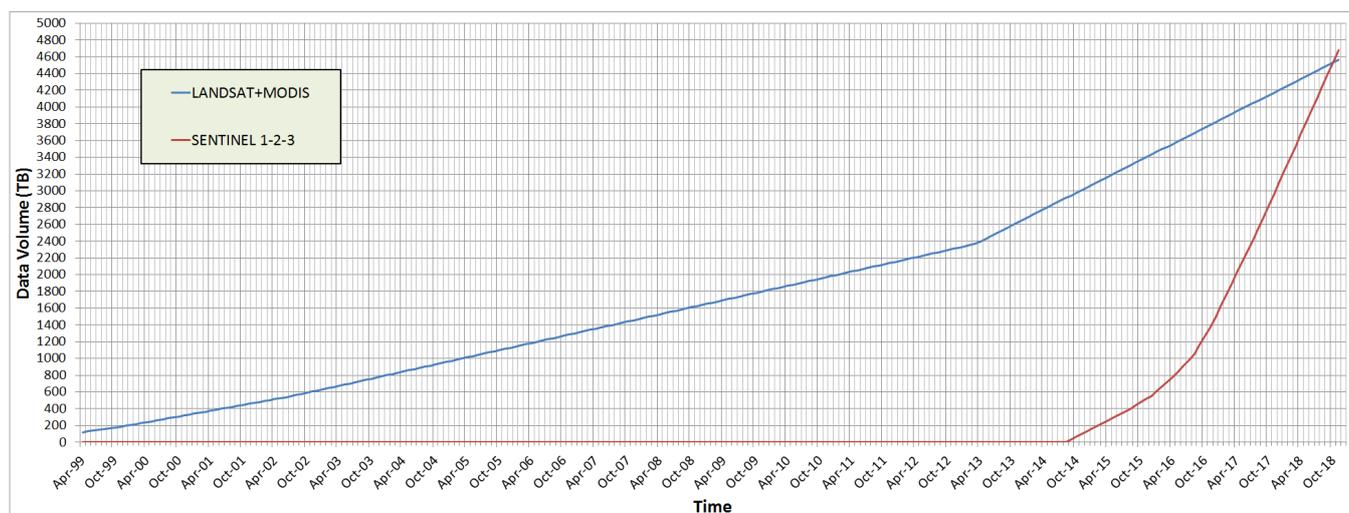
### 2. THE SENTINELS AND THE BIG EO DATA ERA

The evolution of the cumulative data produced by the Landsat missions and Sentinel 1-2-3 with estimations until summer 2018 is shown in Fig. 2. The underlying calculations are based on the following assumptions and considerations:

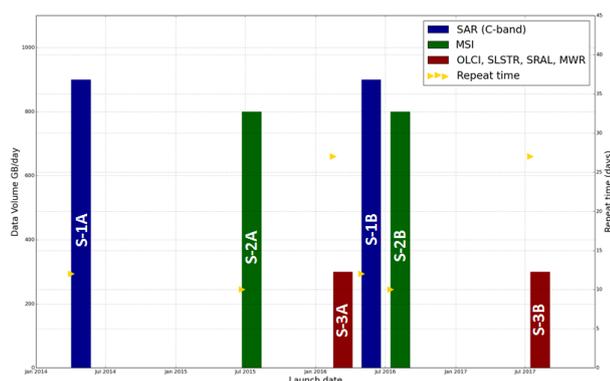
1. The data volume of Landsat 1-6 missions is  $\sim 120 \text{ TB}^2$ ;
2. The data volume of Landsat 7 and 8 is estimated from the relating metadata files provided by USGS<sup>3</sup>;
3. MODIS Terra and Aqua generate 70GB/day each [7];

<sup>2</sup><http://academic.emporia.edu/aberjame/remote/landsat/landsat.htm>

<sup>3</sup><http://landsat.usgs.gov/metadatalist.php>



**Fig. 2.** Data volume: evolution of the data produced by the MODIS/Landsat missions and Sentinel 1-2-3 (estimations based on data throughput in full operational capacity after six months of the scheduled launch dates).



**Fig. 3.** Data velocity of Sentinel 1–3: daily data throughput in full operational capacity and orbit repeat time (in days).

4. The data volume of Sentinel 1, 2 and 3 in their twin constellations is approximately 3.6 TB/day, 1.6 TB/day and 0.6 TB/day respectively<sup>4</sup>; all starting dates are considered to be around six months after the respective mission launch (assuming that data are generated in full operational capacity by that date).

These estimations indicate that the cumulated data generated by the Sentinel 1–3 missions will exceed the data generated by all Landsat missions during 2018.

The data velocity daily of S1–3/A-B is highlighted by their data throughput and orbit repeat times, see Fig. 3. Finally, data variety of the Sentinel 1 to 3 satellite series is summarised in Table 1. Revisit times are not indicated since they increase with latitude.

<sup>4</sup>[https://www.ffg.at/sites/default/files/esa\\_sentinel\\_core\\_products\\_overview.pdf](https://www.ffg.at/sites/default/files/esa_sentinel_core_products_overview.pdf)

### 3. PLATFORMS FOR BIG EO DATA

This section presents a brief survey of some existing platforms and other initiatives to address the needs of big EO data. Given the available space for this short paper, it is by no means comprehensive. In particular, platforms mostly devoted to data dissemination are not considered here.

#### 3.1. Examples from public sector

- NASA Earth Exchange (NEX)<sup>5</sup> is a platform for scientific collaboration, knowledge sharing and research in the Earth science community;
- ESA Exploitation Platforms (EP): Thematic Exploitation Platforms (TEPs), Earth Exploitation Platforms (EOPs) and Sentinel Application Platform (SNAP);
- DLR GeoFarm hardware organisation follows the cloud-like vitalisation of processing hardware, see also [4].
- The Theia Land Data Centre is a French national inter-agency organisation designed to foster the use of images issued from the space observation of land surfaces [6];
- Earth Observation Data Center (EODC)<sup>6</sup> public/private initiative [2] that combines HPC with data provision and collaborative development;
- Australian Geoscience Data Cube (AGDC) at National Computational Infrastructure in Canberra, Australia [5].

<sup>5</sup><https://nex.nasa.gov>

<sup>6</sup><https://www.eodc.eu>

**Table 1.** Sentinel 1 to 3 data variety: spectral, temporal, and spatial (repeat cycle/global coverage for 1 satellite).

Mission	Sensors	Applications	Repeat cycle/ Global coverage	Resolution	Formats
Sentinel-1	C-Band SAR	Monitoring: sea ice, oil spills, marine winds and waves, land-use change, respond emergencies such as floods and earthquakes	12 days/12 days	Strip map mode 80km swath 5x5m; Interferometric wide swath 240km 5x20m; Extra wide swath 400km 25x100m; Wave mode 20x20km at 5x20m	.SAFE with GEO-TIFF, XML, PNG, XSD, HTML and netCDF files
Sentinel-2	MSI (13 bands from 443 nm to 2,190 nm)	Monitoring agriculture, forests, land-use change, land-cover change; mapping biophysical variables; monitoring coastal and inland waters; risk mapping and disaster mapping	10 days/10 days	10m, 20m, and 60m spatial resolution to identify spatial details consistent with 1ha minimum mapping unit)	.SAFE with JPG2000, XML, GML and HTML files
Sentinel-3	OLCI (21 bands from 400 nm to 1,020 nm), SLSTR (9 bands from 555 nm to 10,850 nm), SRAL, MWR	Monitoring sea (ocean circulation, tides), coastal zone, inland waters and land, ice and sea-ice, climate geodesy and geophysics, and land topography	27 days/3 days	OLCI 300m, SLSTR (500m for solar reflectance, 1km for thermal infrared bands)	.SAFE with GEO-TIFF, XML, PNG, XSD, HTML and netCDF files

### 3.2. Examples from private sector

Several companies are hosting large amounts of EO data in combination with processing capabilities, e.g., Google Earth Engine (GEE)<sup>7</sup> for a web-based platform with dedicated web API, Amazon Web Services (AWS) with availability of Sentinel-2<sup>8</sup> and Landsat-8<sup>9</sup>, and CloudEO<sup>10</sup> that proposes a geo-infrastructure as a service.

### 3.3. Big EO data initiatives

In parallel to the development of public and private platforms, a number of governmental and research initiatives aiming at addressing the needs of big geospatial data are flourishing; see for example:

- BEDI is a U.S.A. government big data initiative on civil Earth observation<sup>11</sup>;
- The Big SkyEarth EU COST action<sup>12</sup>;
- The EarthServer: <http://www.earthserver.eu/> see also [1];
- *EarthCube* is a joint initiative between the Division of Advanced Cyberinfrastructure (ACI) and the Geosciences Directorate (GEO) of the US National Science Foundation (NSF).

<sup>7</sup><https://earthengine.google.com>

<sup>8</sup><http://sentinel-pds.s3-website.eu-central-1.amazonaws.com>

<sup>9</sup><http://aws.amazon.com/public-data-sets/landsat>

<sup>10</sup><http://www.cloudeo-ag.com>

<sup>11</sup>[http://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/national\\_plan\\_for\\_civil\\_earth\\_observations\\_-\\_july\\_2014.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/national_plan_for_civil_earth_observations_-_july_2014.pdf)

<sup>12</sup>[http://www.cost.eu/COST\\_Actions/TDP/Actions/TD1403](http://www.cost.eu/COST_Actions/TDP/Actions/TD1403)

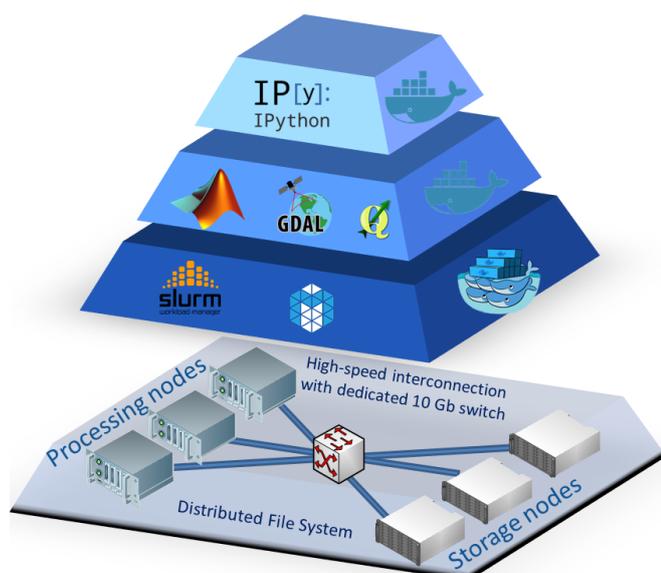
- The Earth System Grid Federation: An Open Infrastructure for Access to Distributed Geospatial Data;
- NOAA's Big Data Project within which a set of Data Alliances are being formed with providers of IaaS.

## 4. JRC EO DATA AND PROCESSING PLATFORM

A number of JRC projects are exploiting Earth Observation data to achieve their goals. In the Sentinel era, this can only be addressed by an integrated approach combining data storage and data processing. This leads to the concept of JRC EO Data and Processing Platform (JEODPP) developed in the framework of the JRC EO&Social Sensing Big Data (EO&SS@BD) pilot project.

The envisaged architecture consists of processing servers accessing the data provided by a series of storage servers and their directly attached storage (Just a Bunch of Disks or JBODs) in a distributed file system environment. The I/O bottleneck typically observed with network attached storage is avoided by considering appropriate high speed server intercommunication topology (switched fabric in fibre channel). This topology has the best scalability compared to arbitrated loop and point-to-point alternatives. Storage servers are automatically populated with the data requested by the applications. For example, the automatic download of Sentinel-2A data is achieved by using a time-based job scheduler launching OpenSearch and OpenData (ODat) scripts taking into account user requirements (geographical areas, cloud coverage, seasonality, etc.).

Processing can be performed at various levels through a sandbox environment. In its simplest utilisation, users have direct access to a batch job scheduler allowing the automatic and distributed processing of EO data at continental or global scale. For example, the detection of clouds on the full S2A



**Fig. 4.** Components of the planned JRC EO data and processing platform (JEODPP).

expert user data (about 30 TB of JPEG2000 compressed imagery) in only 5 days on a computing cluster with 10 processing nodes. Similarly, the automatic mosaicing of 2.5m SPOT (Copernicus CORE\_3 data set) covering the whole territory of the EU plus 11 additional states (European Environment Agency member and associate members) [9] as well as the computation of Global Human Settlement Layers [8] from 4 multitemporal global Landsat data sets are achieved within one week.

The sandbox will offer different levels of user interaction with the platform:

- direct access to the batch job scheduler (slurm or HTCondor);
- virtualisation/operating-system-level virtualisation of the desired environment for prototyping (based on Docker Linux containers);
- interactive EO visualisation/processing capabilities;
- interactive data science and scientific computing through Jupyter web application (IPython notebooks), see also [3].

The various components of the planned EO data and processing platform are sketched in Figure 4.

## 5. CONCLUDING REMARKS

The EU Copernicus programme with its series of Sentinel missions acts as a game changer by bringing EO into the big data era. The value of the produced data depends on our capacity to extract information from it. The velocity, variety,

and volume of the generated data combined with the need for using other non EO data sources calls for innovative approaches in data storage and processing. The scope of the JRC EO&Social Sensing Big Data (EO&SS@BD) pilot project is to propose innovative solutions addressing the needs of JRC projects. We are currently testing and optimising the various components of the JEODPP to ensure its scalability.

## 6. REFERENCES

- [1] P. Baumann et al. Big data analytics for earth sciences: the earthserver approach. *International Journal of Digital Earth*, 1–27, 2015. doi: 10.1080/17538947.2014.1003106.
- [2] C. Briese, et al. Challenges in the exploitation of big earth observation data. In *Proc. of BiDS'14*, 49–52. 2014. doi: 10.2788/1823.
- [3] A. Burger et al. Towards an infrastructure for interactive Earth Observation data analysis and processing. In B. Chan, editor, *Cloud Services for Synchronisation and Sharing*, p. 27. ETHZ, 2016. doi: 10.5281/zenodo.44783.
- [4] S. Kiemle, et al. Big data management in earth observation - the German satellite data archive at DLR. In *Proc. of BiDS'14*, pages 45–49. doi: 10.2788/1823.
- [5] A. Lewis, et al. Iterating Petabyte-Scale Earth Observation Processes in The Australian Geoscience Data Cube. In *Proc. of BiDS'14*. doi: 10.2788/1823.
- [6] C. L'Helguen et al. Muscate: Multi-satellites, multi-sensors and multi-temporal theia data centre. In *Proc. of BiDS'14*, 114–118. doi: 10.2788/1823.
- [7] E.J. Masuoka, et al. Key characteristics of MODIS data products. *IEEE Transactions on Geoscience and Remote Sensing*, 36(4):1313–1323, 1998. doi: 10.1109/36.701081.
- [8] M. Pesaresi, et al. Operating procedure for the production of the global human settlement layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. Technical Report, JRC, 2015. doi: 10.2788/253582.
- [9] P. Soille. Seamless mosaicing of very high resolution satellite data at continental scale. In P. Soille et al., editors, *Proc. of BiDS'14*, 222–223. doi: 10.2788/1823.
- [10] P. Soille et al. Preface. In *Proc. of the BiDS'14* doi: 10.2788/1823.
- [11] The Landsat Science Team. Free access to Landsat imagery. *Science*, 320(5879):1011, 2008. doi: 10.1126/science.320.5879.1011a.