

European  
Commission

# JRC SCIENTIFIC AND POLICY REPORTS

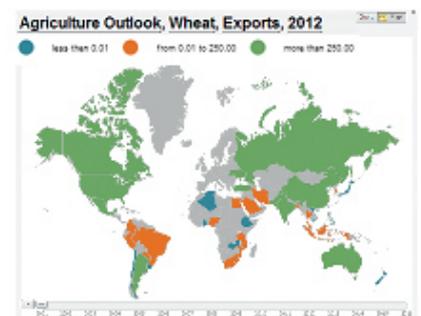
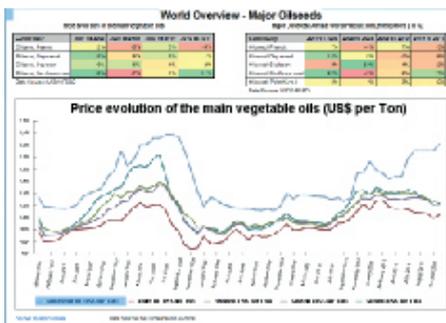
## DataM

# Data on Agriculture, Trade and Models

A tool for flexible management,  
extension and integration of  
(model) databases

Sophie H elaine  
Mihaly Himics  
Robert M'barek  
Arnaldo Caivano

2013



Report EUR 25976 EN

European Commission  
Joint Research Centre  
Institute for Prospective Technological Studies

Contact information

Address: Edificio Expo. c/ Inca Garcilaso, 3. E-41092 Seville (Spain)  
E-mail: [jrc-ipts-secretariat@ec.europa.eu](mailto:jrc-ipts-secretariat@ec.europa.eu)  
Tel.: +34 954488318  
Fax: +34 954488300

<http://ipts.jrc.ec.europa.eu>  
<http://www.jrc.ec.europa.eu>

Legal Notice

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

Europe Direct is a service to help you find answers to your questions about the European Union  
Freephone number (\*): 00 800 6 7 8 9 10 11

(\* Certain mobile telephone operators do not allow access to 00 800 numbers or these calls may be billed.

A great deal of additional information on the European Union is available on the Internet.  
It can be accessed through the Europa server <http://europa.eu/>.

JRC 81668

EUR 25976 EN

ISBN 978-92-79-29915-5 (pdf)

ISSN 1831-9424 (online)

doi:10.2791/17240

Luxembourg: Publications Office of the European Union, 2013

© European Union, 2013

Reproduction is authorised provided the source is acknowledged.

Printed in Spain



# JRC SCIENTIFIC AND POLICY REPORTS

## DataM Data on Agriculture, Trade and Models

A tool for flexible management, extension  
and integration of (model) databases

Sophie H elaine, Mihaly Himics, Robert M'barek, Arnaldo  
Caivano

2013

Disclaimer:

The views expressed in this document are purely those of the authors and may not in any  
circumstances be regarded as stating an official position of the European Commission.

Joint Research Centre



# Acknowledgements

This report has been prepared by JRC IPTS staff from the AGRILIFE unit.

We would like to thank in particular Ana Nicalau, Sergey Svitkov and Vladimir Eskin and all the team at PROGNOZ for the excellent cooperation in developing DataM.

We are grateful to Anna Atkinson from the JRC IPTS for the language editing and formatting of the report and to Fabien Santini for his invaluable advice.

Finally, we would also like to thank Jacques Delincé, Head of the AGRILIFE unit, for his comments and support throughout the project.



# Contents

<b>1 Introduction</b>	<b>7</b>
<b>2 Benefits of the tool</b>	<b>9</b>
2.1 Easy access to the data	9
2.2 Up-to-date data	10
2.3 Data harmonisation	10
2.4 Source documentation and meta-information	10
2.5 Data analysis and reporting	10
<b>3 Databases included in DataM</b>	<b>11</b>
<b>4 The main features of the tool</b>	<b>13</b>
4.1 The source documentation and the metadata definition tools	14
4.2 The search functionality	15
4.3 The data visualisation in the 'Workbook'	16
4.4 The reporting	18
4.5 The export and import of data	20
4.6 Transformations	20
4.7 The 'Management console'	21
<b>5 Databases structure</b>	<b>23</b>
5.1 A tool designed for time series	23
5.2 Nomenclature and classification schemes	23
5.3 Specific databases organisation	23
5.4 Groups and hierarchies	23
<b>6 Data harmonisation</b>	<b>25</b>
6.1 A pragmatic approach	25
6.2 The common dictionaries	25
6.3 Linking databases to the common dictionaries	27
6.4 Examples of benefits for using the common dictionaries	27
<b>7 Technical infrastructure</b>	<b>29</b>
7.1 Back office application	30
7.2 Web application	30
<b>8 Data flow in DataM</b>	<b>31</b>
8.1 A data warehouse solution	31
8.2 Hyper-Cubes	31
8.3 Data marts	32
8.4 ETL, Extracting - Transforming - Loading	33
8.5 Synchronisation	33
<b>9 DataM – summary and next steps</b>	<b>35</b>
<b>Annex 1: List of the balance sheet common attributes</b>	<b>39</b>
<b>Annex 2: List of the world reference prices</b>	<b>41</b>

## Acronyms

---

<b>AMECO</b>	Annual macroeconomic database of the European Commission's Directorate General for Economic and Financial Affairs
<b>CAP</b>	Common Agricultural Policy
<b>CAPRI</b>	Common Agricultural Policy Regional Impact Analysis
<b>CGE</b>	Computable General Equilibrium
<b>CN</b>	Combined Nomenclature
<b>COMEXT</b>	Statistical database on trade of goods managed by Eurostat
<b>COMTRADE</b>	Commodity Trade database (United Nations)
<b>CPI</b>	Consumer Price Index
<b>DBMS</b>	Database Management System
<b>DG AGRI</b>	Directorate General for Agriculture and Rural Development
<b>DG ECFIN</b>	Directorate General for Economic and Financial Affairs
<b>EC</b>	European Commission
<b>ESIM</b>	European Simulation Model
<b>ETL</b>	Extracting - Transforming - Loading
<b>EU</b>	European Union
<b>EU-27</b>	27 EU Member States after 2007 enlargement
<b>Eurostat</b>	Statistical Office of the European Commission
<b>FAO</b>	Food and Agriculture Organization of the United Nations
<b>FAPRI</b>	Food and Agricultural Policy Research Institute, US
<b>GAMS</b>	General Algebraic Modeling System
<b>GUI</b>	Graphical User Interface
<b>HS</b>	Harmonized Commodity Description and Coding System
<b>IMF</b>	International Monetary Fund
<b>IPTS</b>	Institute for Prospective Technological Studies
<b>JRC</b>	Joint Research Centre
<b>OECD</b>	Organisation for Economic Co-operation and Development
<b>OLAP</b>	Off-line Analytical Processing
<b>OLTP</b>	On-line Transactional Processing
<b>PE</b>	Partial Equilibrium
<b>PSE</b>	Producer Support Estimate
<b>RDP</b>	Remote Desktop Protocol
<b>US</b>	United States of America
<b>USDA</b>	US Department of Agriculture
<b>w.s.e.</b>	White sugar equivalent

# 1 Introduction

Agriculture and trade analysis have regained importance during recent years due to the increase in food prices and the implications for global food security. Also, agriculture is increasingly becoming more integrated into the macroeconomy and linked to other sectors such as the energy sector. Consequently the tools for quantitative analysis, as well as the management of data and models, are confronted with new and more complex challenges.

The JRC IPTS responded to these challenges with the creation of the 'integrated Modelling Platform for Agro-economic Commodity and Policy Analysis (iMAP)' in 2005. iMAP aims to facilitate the use of selected Partial Equilibrium (PE) and Computable General Equilibrium (CGE) models, either in stand-alone mode or in combination.<sup>1</sup> Important efforts are also focused on achieving the availability and accessibility of high-quality and reliable data, updated on a regular basis, not least to feed economic models which provide science-based policy support.

Even if these models differ with respect to regional and commodity coverage, they nevertheless share the need for a large amount of data (e.g. time series on wheat production in EU Member States). A systematic approach, i.e. a harmonised database used as input for all models, would be ideal and would ensure that differences in model results were only due to differences in model structures. Reaching this target is very ambitious and will require time. A first

step in that direction is to ease the access to the main data required for modelling exercises by linking them by common nomenclatures.

In 2007, the JRC IPTS started to develop concepts for data consolidation, harmonisation and management together with external partners.<sup>2</sup> Building on the experience of this project, the JRC IPTS began a second initiative in 2009, together with the PROGNOZ software company, to design and implement a software platform for data harmonisation and management. The resulting product, DataM, is a database management tool intended to simplify the daily data work of analysts and modellers. DataM facilitates the input of data into economic models, verification of statistical information and/or analysis of results.

Over the course of its development it turned out that including a multitude of important statistical databases makes DataM a valuable product on its own right and not only as an input for economic models. The rich data content and the powerful IT framework offer a unique opportunity to develop data dissemination tools on top of the DataM platform. Thus the current developments also focus on improving the accessibility for users; both in terms of access to the software (a web-based solution that is accessible with a simple web browser) and in terms of access to the relevant information (the introduction of 'dashboards').

<sup>1</sup> JRC Scientific and Technical Report (2012): An integrated Modelling Platform for Agro-economic Commodity and Policy Analysis (iMAP). European Commission, JRC 69667. <http://ftp.jrc.es/EURdoc/JRC69667.pdf>

<sup>2</sup> JRC Scientific and Technical Report (2008). Potentials of a Harmonised Database for Agricultural Market Modelling. EUR 23417. <http://ftp.jrc.es/EURdoc/JRC43298.pdf>



## 2 Benefits of the tool

DataM is a database management tool focusing on data on agriculture, trade and models from the main data providers. The main strength of DataM is its ability to link these different databases by their nomenclatures. Users can quickly retrieve information from different sources without specific knowledge of the nomenclature used by the data providers. Using one interface only, users can rapidly access and compare the main agricultural, macroeconomic and trade databases as well as the model databases. The tool addresses different needs, ranging from data collection and data checks to advanced reporting, with the possibility to export data.

### 2.1 Easy access to the data

With DataM, users can access most of the databases they need through a unique interface. They can retrieve information from several databases with only one account and using a single query tool. The databases are in a unique format regardless of the data origin. It accelerates the data work as users no longer need to handle different file formats or perform certain transformations, and they can directly compare the data through the graphical interface. In addition, DataM offers a user-friendly interface, similar to those in Office® applications<sup>3</sup>.

A powerful search engine enables the relevant information to be found in the available databases. This search functionality is one of the main advantages of the tool because it allows many different databases to be searched simultaneously and the results can be displayed in a table with just one click. Moreover, the linkage between the nomenclatures ensures that when searching using the common dictionary (see chapter 6.2) none of the data series related to the required topic are missed.

Within the iMAP modelling initiative, different agro-economic models (in particular AGLINK-COSIMO<sup>4</sup>, CAPRI<sup>5</sup> and ESIM<sup>6</sup>) are used to build yearly agricultural baselines<sup>7</sup>, analyse trade agreements or the Common Agricultural Policy (CAP). Some of these models do not have a user-friendly Graphic User Interface and DataM can be used to analyse the results of these models. In addition, DataM allows users to analyse model results without knowledge of the specific programming languages (e.g. GAMS, TROLL). In order to rapidly import model results (and databases) integrated 'data loaders' have been added to DataM. Moreover the three core models can then be easily compared in DataM using either the input (historic) data or the simulated results.

In DataM the organisation of the databases is optimized to ease the data access. The databases are organised by topics (outlooks, prices, trade, models, macroeconomic data, farm structure and income). In addition, while keeping the original data classifications, as far as possible the databases were organised following a similar structure: Country, Commodity and Attributes. In the specific case of Eurostat, the multiple databases available on the web are grouped together by subject to speed up the data search and limit the list of data series. For the main agricultural outlooks, the different releases are compiled in one dataset, so that users can easily compare the changes in the projected values.

The data selected can be easily exported in different formats, e.g. in Excel for further analysis or in GDX format to feed a model database.

<sup>4</sup> OECD (2006). Documentation of the AGLINK-COSIMO model, Working Party on Agricultural Policies and Markets, Organisation for Economic Cooperation and Development (OECD), Paris.

<sup>5</sup> CAPRI (Common Agricultural Policy Regional Impact): highly disaggregated (regions NUTS 2, products) partial equilibrium (PE) model. A model documentation is provided by Britz, W. and H.-P. Witzke (eds.) (2008): CAPRI Model Documentation 2008, Version 2. Institute for Food and Resource Economics, University of Bonn, Germany. [http://www.capri-model.org/docs/capri\\_documentation.pdf](http://www.capri-model.org/docs/capri_documentation.pdf).

<sup>6</sup> Banse, M., H. Grethe (2008): 'European Simulation Model (ESIM) in GAMS: Model Documentation (Version 2.0).' Model documentation prepared for DG AGRI, European Commission, The Hague and Berlin.

<sup>7</sup> A baseline is a consistent set of market and sector income prospects elaborated on the basis of specific policy and macroeconomic assumptions. It is also known as an 'Outlook'.

<sup>3</sup> The aim is to reduce learning costs, as users are already familiar with the interface.

## 2.2 Up-to-date data

In DataM, the data is updated regularly: the daily and weekly databases are updated twice a month; the monthly databases are updated every month and the databases released once a year are updated within 5 or 10 working days depending on the urgency of the work. The information on the data version (i.e. last update of the data by the source provider) available in DataM is displayed in the main form of the tool.

Updates are scheduled every week following a calendar agreed with PROGNOZ who are in charge of loading the data from the websites of the data providers. If there is a licence issue, the data is loaded by the JRC IPTS and sent to PROGNOZ for integration in the tool. More details on the data flow are available in chapter 8.

As part of the update process PROGNOZ also ensures the consistency of the variable codes. For example, the nomenclature of trade data in the COMEXT database changes every year so PROGNOZ establishes the correspondence between the codes to keep data series as long as possible.

As explained in chapter 2.2, the model data (baseline, scenarios) can rapidly be uploaded or updated using 'data loaders'. Users are fully autonomous with regards to the management of the model data and do not need PROGNOZ to update these data.

## 2.3 Data harmonisation

To ease the data search and comparison, the nomenclature of the databases are linked via common dictionaries. These enable the user to work with data from different sources under one unique nomenclature (the common one). They include a list of countries, commodities and attributes (like feed use or production). For example 'maize' is part of the common dictionary of commodities and is linked to 'maize' or 'corn' according to the nomenclature of the different databases. A search on 'maize' will therefore return all data series on 'maize' and 'corn'. As a result, as long as users are familiar with the common dictionaries, they do not need to know all the specific databases' nomenclature to look for data. This is a significant benefit because very often junior economic modellers are not familiar with agricultural databases. In addition, in services with a high staff turnover this harmonisation decreases the entry barrier to work on agricultural data.

Furthermore, for the common dictionaries, data series are converted into common units of measurement, allowing a direct data comparison. Time series from a single data source are also harmonised in case there is a change in definitions over years. This is typically the case for COMEXT but also for the agricultural Outlook databases. All details on the data harmonisation are provided in chapter 6.

## 2.4 Source documentation and meta-information

The source documentation (methodology, variable definitions) is stored in the tool and is directly accessible from the main form of DataM. In addition, a summary report informs the user of the data source provider, the potential need for a licence, the data frequency, the update frequency in DataM and the geographical coverage. The meta-information can also be consulted directly when visualising the data. The availability of the main information on the databases directly in the tool is particularly useful in the context of a high staff turnover.

## 2.5 Data analysis and reporting

DataM provides visualisation tools (tables and graphic presentation) for data analysis and for the comparison of different data sources:

- historic data versus model input data,
- different scenario results of a single model,
- projections of different institutions and in-house models.

The data can be analysed using time series' statistical/econometric functionalities: time series' calculations (percentage change, summations, moving averages...), collapse (from Monthly to Yearly data) or basic econometric analysis. For data validation, rules can be defined to look for missing data or detect outliers, e.g. 'abnormal' growth rates and levels.

DataM is also equipped with a reporting tool. Reports for repeated data analysis, extraction tasks or for dissemination purposes can be pre-defined. Ready-to-use and customisable report templates allow an analysis to be updated rapidly, for example in one click a report can display new scenario results. The reporting tool can also be used for rapid collection of model inputs, e.g. AGLINK-COSIMO data transformed into ESIM format for the calibration of the ESIM model or macroeconomic data to feed the ESIM model. These reports will be automatically updated whenever there is a new baseline version or when there is a new release of macroeconomic forecasts.

The data selected or the reports can be directly exported into MS Office applications for further formatting or for publishing. Reports and data selections can be shared with all data users or with specific ones if the data is under licence or if the work is in progress. For example, reports can be built to inform other users of the latest AGLINK-COSIMO baseline or to compare the baselines of different models or different published outlooks.

Users with specific rights can modify the data, e.g. if a mistake is noticed, these changes are recorded (author and motivation). Any data point can be traced back to its origin, i.e. data versioning is available. Given these functionalities, DataM is also ideal for the storage of the scenario results of a policy report.

## 3 Databases included in DataM

The choice of databases is oriented to the needs of the users. One of the main tasks of the AGRILIFE Unit is to provide policy support in the agricultural field using different agro-economic models. As a consequence, besides model databases, the main agricultural databases with historic data on balances, the major agricultural outlooks, trade and prices databases are included in the tool. The number of databases in DataM can evolve over time according to the

needs. The databases currently available in DataM are listed in Table 3.1.

Certain databases are under licence (e.g. COMTRADE) and DataM allows user rights to be defined at the level of each database. Specifically, users or a group of users may have no rights to access a database, have read-only rights or editing rights depending on the databases (see chapter 4.7).

**Table 3.1 Main data sources available in DataM**

### Commodity balances and outlooks

DG AGRI market outlook

OECD-FAO outlook

FAPRI US and world outlook

USDA projections & USDA PS&D

Eurostat Balance sheets and Production

FAO Balance sheets and Production

Oil World database

### Prices

DG AGRI prices

Eurostat Prices

Oil World Prices

FAO Prices

World Bank Commodity prices

Futures (Wall Street Journal, MATIF, LIFFE)

### Macroeconomic data

AMECO – DG ECFIN database

Eurostat

IMF

Global Insight (selection of macroeconomic data)

UN Population Demographic Statistics & UN World Population

### Trade

UN COMTRADE (HS1992, HS2002, HS2007, SITC2, SITC4)

Eurostat – COMEXT

FAO trade



**Model databases**

AGLINK-COSIMO

ESIM

CAPRI

GTAP

AGROSAM

**Structure&Income**

Eurostat (EAA, FSS)

FADN data

**Other**

OECD PSE

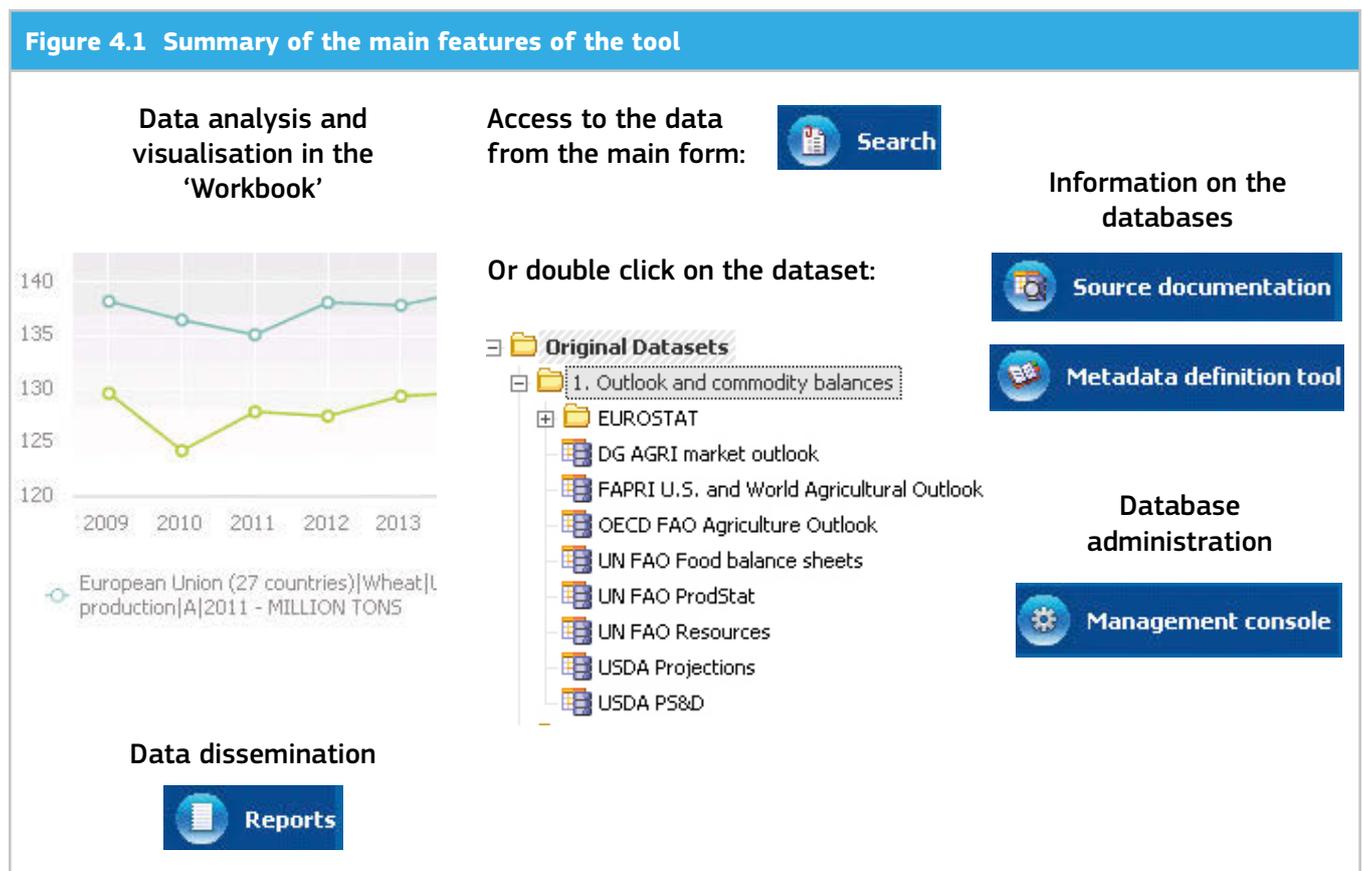
DG AGRI Budget expenditure

World Development Indicators (WDI)

---

# 4 The main features of the tool

This chapter provides an overview of the main features available in DataM. These features are summarised in Figure 4.1.



## 4.1 The source documentation and the metadata definition tools

Access to the source documentation is available on the main form of DataM together with information on the data version (i.e. the date of the last update of the data by the source provider) available in DataM. In one click users can open a summary report with summarised information on the database. In addition, when meta-information is available on the source provider websites it is displayed in the source documentation report.

The source provider's meta-information is also attached to the element and added as metadata to the original database nomenclatures. These definitions, the original databases' nomenclatures and the data harmonisation are managed

via the metadata definition tool. To enhance knowledge on the data, additional comments on the variables can be added by the users (with administrative rights for the related database). For example, in the DG AGRI Market Outlook database the crop year definition was introduced in the commodity classification as a comment. This 'free' field is meant to be complemented over time.

The meta-information can also be directly consulted when visualising the data in the 'Workbook' via a viewer (see Figure 4.2). It is also in this window that new groups of elements and different elements' classifications can be defined (see chapter 5.4). The linkage to the common dictionary is made in this interface (see chapter 6).

In sum, the metadata and source documentation information can be accessed through three different functionalities (source documentation, metadata definition, and visualisation) of DataM.

**Figure 4.2 Source documentation and metadata for production in the OECD-FAO outlook**

The figure displays two side-by-side screenshots from the DataM application. The left screenshot, titled 'Report in the source documentation', shows a web browser window with the URL 'OECD FAO Agruculture Outlook' and a document titled 'USDA Projec'. The browser's address bar shows '110%' zoom. The page content includes a 'Summary' tab and a 'Definitions of Variables' tab. A table is visible with the following content:

A	
<b>Production:</b>	
<b>Crop Production</b>	Crop product harvesting ar includes the i consumed or production p; production tc production fa obtained as a statistics is er area, which v

The right screenshot, titled 'Metadata viewer in the Workbook', shows a 'Metadata viewer' window. It has tabs for 'Contextual', 'Production events', and 'Objects'. The 'Contextual' tab is active, showing a list of properties for 'Production - kt'. The list includes:

Property	Value
Attributes	Production - kt
Name	Production - kt
Code	QP
Common Code	Prod
Common Attribute	Production
ISAGR	
Comments	
Metadata	Crop production data refer actual harvested productic the field or orchard and ga excluding harvesting and threshing losses and that p crop not harvested for any reason. Production therefc includes the quantities of tl

## 4.2 The search functionality

The search functionality is available on the main form of DataM and permits searches of all or part of the databases using the name or the code of the elements. The search

returns the list and number of data series containing the chain of characters searched for, organised by database and providing the years available (see Figure 4.3). For models, results are additionally classified by scenario. The results of the search can be selected and opened in a table in the 'Workbook' (see chapter 4.3).

Figure 4.3 Results for a search on 'EU-27 maize production'

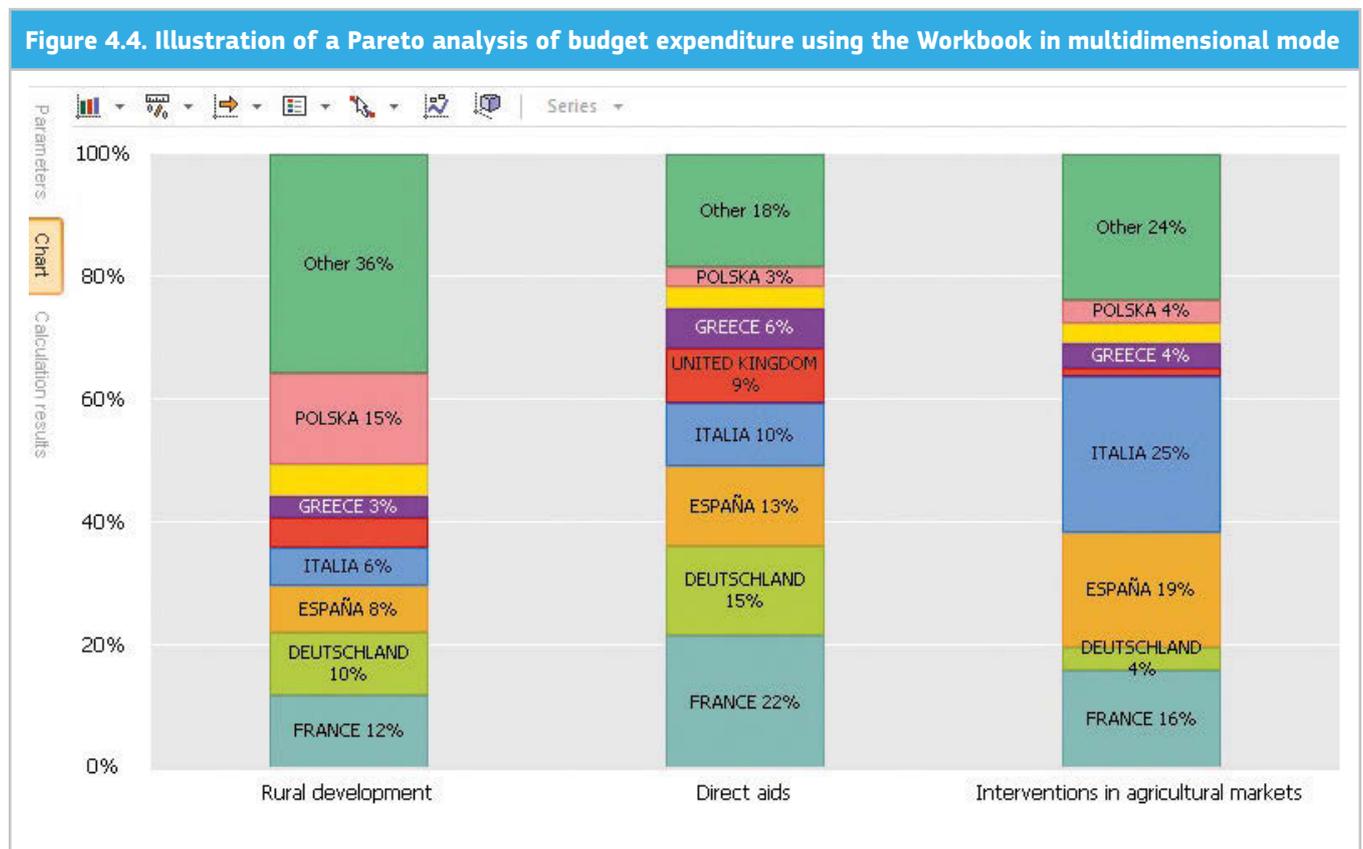
The screenshot displays the search results interface for 'EU-27 maize production'. At the top, there are navigation buttons for Search, Source documentation, News, and Help. Below these, a search bar contains the query 'EU-27 maize production'. The search results are organized into a list of categories, each with a plus icon and a document icon. The first category is '1. Outlook and commodity balances (Total 14)'. Under this category, there are several search options: 'Search in source classifications' (unselected), 'Search in common classifications' (selected), 'Whole word' (unselected), 'Match case' (unselected), and 'To display results with common names' (unselected). Below these options is an 'Open' button. The search results list includes:

- EUROSTAT Crop production&Land use (1, 2002-2009)
- DG AGRI market outlook (3, 2006-2020)
  - European Union (27 countries)|Maize|Production (common)|Annual|2009
  - European Union (27 countries)|Maize|Production (common)|Annual|2010
  - European Union (27 countries)|Maize|Production (common)|Annual|2011
- FAPRI U.S. and World Agricultural Outlook (3, 1999-2025)
- OECD FAO Agriculture Outlook (1, 1995-2019)
- UN FAO Food balance sheets (1, 1961-2009)
- UN FAO ProdStat (4, 1961-2010)
- USDA Projections (2, 2009-2021)
- USDA PS&D (1, 1999-2012)

### 4.3 The data visualisation in the ‘Workbook’

By clicking on the databases listed on the main form or opening the results of the search, the user will be redirected to the ‘Workbook’. This feature (‘Workbook’) is the principal visualising tool of raw data in DataM. Within the Workbook, users can select the data series they want to display in the data area and they can choose to work either in an OLAP mode (multidimensional mode) or in a time series mode. The data is by default automatically filtered, meaning only the non-empty data series are displayed. The data selected is directly graphed for a rapid visualisation.

A user-friendly multidimensional mode is available (see Figure 4.4) allowing the selected data to be organised in rows and columns or permitting the preparation of graphs. Find functions are available to accelerate the process of selecting elements in different dimensions. The list of elements can be displayed by name or by code. For trade or budget data with complex nomenclatures or geographical information at different NUTS<sup>8</sup> levels both the code and the name of the elements can be presented. Users can perform different calculations by rows and columns such as sums, averages, percentage changes and many more. Data can be sorted and a Pareto analysis allows users to select the most significant information, for example the EU Member States receiving 80% of the direct aids in a given year (see Figure 4.4). The rest of the countries are aggregated in an ‘other’ group.



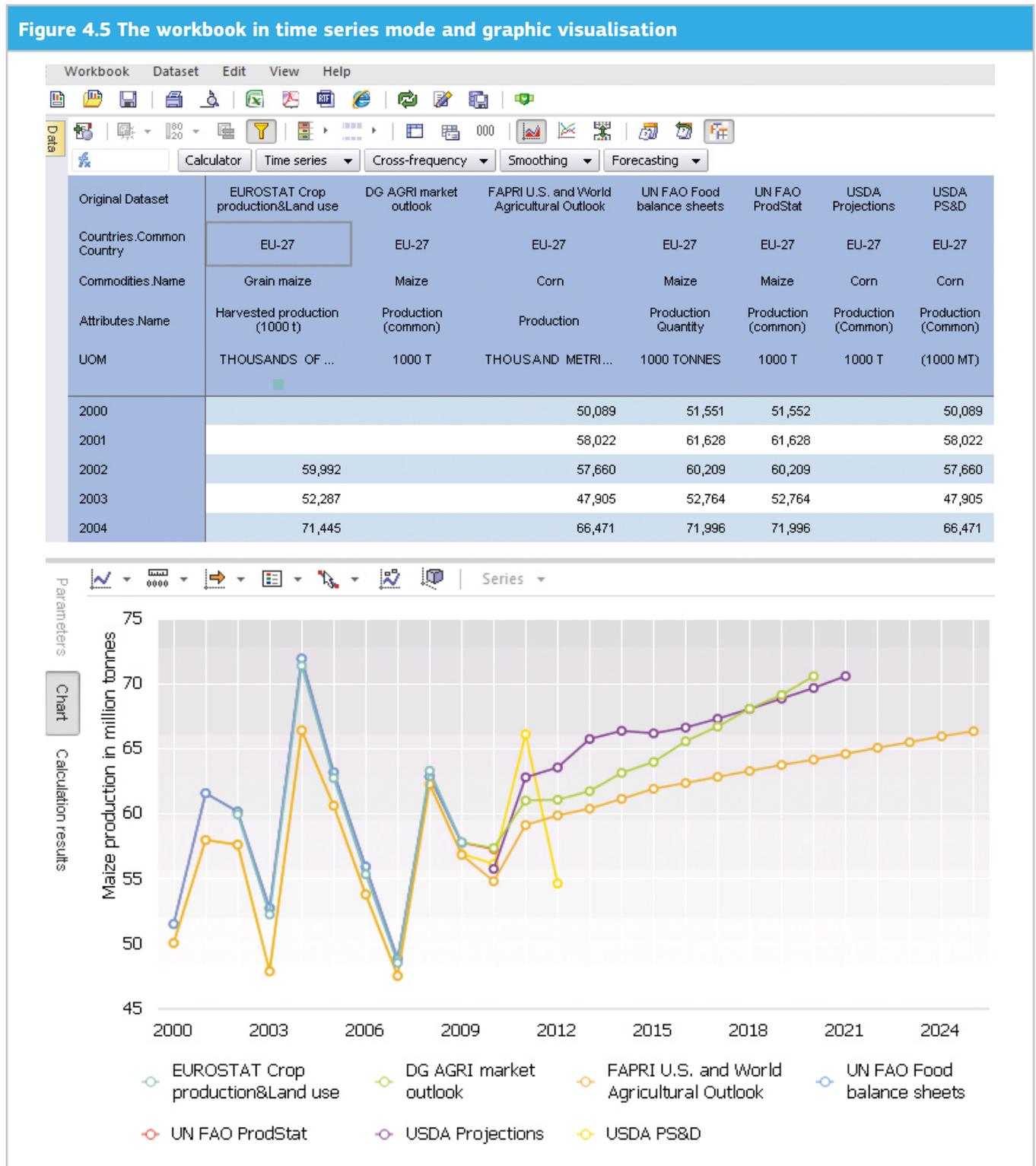
8 The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU.

The other way to select and visualise the data is the ‘time series mode’ (see Figure 4.5). In this mode the time is always indicated and users can choose the range of years they want to display. Users can order the data by dimensions and have the same possibilities to select the elements in the lists as in multidimensional mode, and data is automatically filtered too. Unlike the multidimensional mode, from the time series’ one, users can open as many different databases as they want for a rapid comparison. After a search on the main

form, the data series are displayed in the workbook in this mode.

Simple calculations can be performed like summations, moving averages or yearly growth. Collapse functions are available to transform Monthly data into Yearly data for example. Another very useful function to compare data series is the rebase function: in two clicks it allows data series to be indexed in a year of the user’s choice.

Figure 4.5 The workbook in time series mode and graphic visualisation



Users can save their selection, so that the next time they will open their workbook it will be automatically updated (provided that the underlying data was updated in the meantime). The workbook as a whole or only a graph or a worksheet can be exported to Excel for further processing. Users may prefer to use DataM to search for the data and Excel for the data analysis and visualisation according to their preferred way of working.

When users select a data series they can visualise the metadata associated with the selection in a viewer (see Figure 4.2). The data can also be edited and saved by data administrators. In this case the administrator is asked to register the motivation for the change. The new version of the data is directly available to all users, in addition to the previous versions.

Users can implement series of tests to validate the data and register these validation rules to apply them on a

regular basis. The following tests can be done: missing data, level analysis, growth rate, ratio, trend, statistical outliers, correlation, or even customisation of a validation rule based on the specificities of certain databases.

## 4.4 The reporting

---

DataM contains a reporting feature which can be used for data dissemination but also to build routines of structured data extraction or to analyse scenario results. Reports correspond to an enhanced visualisation of data in support of its analysis. These reports can consist in a combination of tables, graphs and maps. They are updated automatically if the underlying datasets have been updated. In addition these reports can be designed to be dynamic so that users can change the data selection via a control box. In the example presented below users can select the year and commodity of their choice (see Figure 4.6).

Figure 4.6 Report on the 10 biggest producers, importers and exporters by commodity and year in USDA PS&amp;D

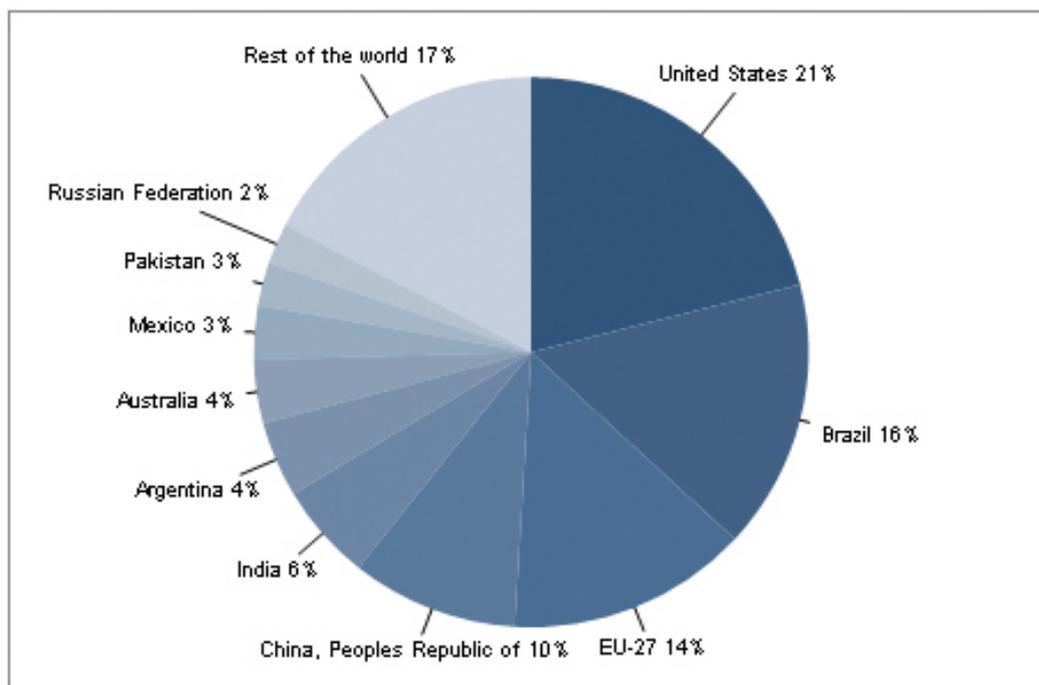
1) Top 10 Producers, Importers, Exporters

100%

**Commodities:** Meat, Beef and **Year:** 2011

10 Biggest Producers, Importers, Exporters

Producers		Production (1000T)
		2011
1.	<b>United States</b>	<b>11,988</b>
2.	<b>Brazil</b>	<b>9,030</b>
3.	<b>EU-27</b>	<b>8,023</b>
4.	China, Peoples Republic of	5,550
5.	India	3,244
6.	Argentina	2,530
7.	Australia	2,129
8.	Mexico	1,804
9.	Pakistan	1,435
10.	Russian Federation	1,360
	Rest of the world	9,895
	World	56,988



Data source: USDA PS&amp;D

The reporting tool has functionalities similar to Excel 2010 with additional reporting features. For example, the elements can be sorted, filtered, and the order can be precisely defined (which is useful for a balance sheet). Empty cells are filtered. Tables' and cells' formats can be specifically defined. New names can also be given to the data, e.g. scenario names for a graphic representation, or the data of one model can be displayed with the codes of another model for direct use as input data after export. For example, for the calibration of the ESIM model to the AGLINK-COSIMO EU baseline a report with the AGLINK-COSIMO data using the ESIM codes is prepared.

Both the reporting tool and the workbook allow the visualisation of data from different databases or data analyses. However defining a new report is less user-friendly than using the workbook, yet it has more powerful visualisation features. However, report templates built by advanced users can be made available to all users and further modified more easily. For example, for the analysis of the AGLINK-COSIMO scenario results, three reports have been defined to analyse one, two or three scenarios and compare them to the baseline. In this case, to analyse a new scenario users can save the template with a different name and select the name of the new scenario in the control box. Ultimately it is up to each user to choose what is the most suitable for their work.

Reports can also be defined for further publication on the web for example or they can be used to provide useful information to users. For instance, the report presented in Figure 4.6 could be very useful to get a rapid overview of the most important players in a sector before starting an analysis.

## 4.5 The export and import of data

A bulk export of a database can be done from the main form in Excel, .mdb, .txt, .csv or .gdx format. The data to be exported can be easily selected and users can choose to keep the years in columns or to transpose them into

rows. This bulk export does not require the database to be opened and can be saved as a task if a similar export is to be repeated on a regular basis. With this functionality only datasets from a unique database can be exported whereas from the workbook data from different databases can be exported together with graphs.

Model data can be imported thanks to a customised user interface for importing model datasets (the automatic 'data loaders'). They are adapted to each model and .gdx files can also be imported. This is particularly useful to import the results of different scenarios. In the event of a new model version with slight changes, new elements are directly added to the nomenclature in the Metadata definition tool.

Datasets in Excel, .txt, .csv, .mdb or .gdx format can also be imported into DataM. In this case different steps are to be followed. First, the structure of the dataset is to be defined in the Metadata definition tool (number and name of the dimensions with the list elements). In a second step the database is to be created from the 'Management Console' (see 0). To ease this step it is possible to use existing databases as templates or existing dimensions' definitions. The data import is the last step. If the import is to be repeated regularly it can be stored as a task for further execution.

## 4.6 Transformations

Transformations are used to perform large-scale calculations on complete datasets, e.g. a change in units. Different kinds of calculations can be implemented including conditionals or using meta-information like the unit of measure. These calculations are run automatically before a data update is delivered, so calculated results are directly available with the data update. The transformations can be implemented for all commodities, countries and attributes or specific elements can be targeted. It is very useful to aggregate the results of a group of countries for example. This module is also essential for the data harmonisation in order to change units (from ha to 1000 ha) or create new items (net exports), etc.





# 5 Databases structure

## 5.1 A tool designed for time series

The frequency is a compulsory dimension as in the underlying software platform all datasets are organised as time series. Different frequencies can be defined: daily, monthly, quarterly and annual. For databases like Oil World which display production by crop year (e.g. from October Year N to September Year N+1), it was decided to display the production in calendar Year N. For ESIM, the base year (an average of two years) was linked to the first year of the average (e.g. in ESIM 2010, the base year is an average of 2006 and 2007 and it is displayed as '2006'). For the DG AGRI budget expenditure, users often tend to forget that most of the payments of a financial year 'N' relate to the production of the year 'N-1'. For this reason it was decided to allocate the budget expenditure of the year 2008 for example to the calendar year 2007.

## 5.2 Nomenclature and classification schemes

The original classification of each database is kept (name and code). Datasets are organised by dimensions. The dimensions are defined in the Metadata definition tool except for the frequency which is compulsory. Similar dimensions were defined for all the databases when possible. The four main dimensions are named: Attributes (production, consumption, etc.), Countries, Commodities and Scenarios for each model or Year of release for Outlooks. For the databases with bilateral trade the country dictionary is split into Reporters and Partners.

## 5.3 Specific databases organisation

Although most of the time the original classification is kept, there are several exceptions to this rule, for instance Eurostat data. In this case, the main issue was to integrate the different tables available on the website into a limited number of datasets. Therefore all the tables related to the same topics and with similar dimensions were integrated.

Nevertheless, in the workbook users can check the name of the 'original' Eurostat database. For example, the 11 tables related to the Economic accounts of agriculture are gathered in one dataset. The crop production survey and the land survey were merged. This enables the differences between the areas recorded in each one for the commodities covered by both surveys to be checked rapidly. This re-grouping sometimes necessitated the creation of a mapping between the commodity nomenclatures of the different Eurostat surveys.

In the case of CAPRI, the original files are very large cubes, so as a consequence the data were reorganised to ease users' utilisation of the data. A non-CAPRI modeller can therefore find the commodity balances in two or three clicks.

For the three databases on futures prices available in DataM, the nearby futures price is calculated for the agricultural contracts and stored in a fourth database. As a result time series on prices can be directly compared to futures prices. For the financial data of the EU budget expenditure, a mapping between the budget codes across the different financial years is under development to build time series which will be readily available for modellers.

If original nomenclatures mix commodity and attribute concepts, like in FAO Prodstat where to display the gross production the user must select the quantity of 'indigenous cattle' produced, the dictionaries may have had to be reorganised. In this example the gross production was added to the list of attributes and the commodities like 'indigenous cattle' were deleted to keep a unique commodity (cattle) for which the net and gross production figures are available.

## 5.4 Groups and hierarchies

Groups of countries, commodities or attributes can be defined in the Metadata definition tool. Users working mainly with balance sheets may be interested in creating a group with the balance sheet. It means that the elements of this group can be selected with a simple click in the workbook or in the reporting tool. Groups of countries like Mercosur or the EU-27 can also be added. To be able to visualise the aggregated results for these groups, 'transformations' rules must be defined (see chapter 4.6).

Creating hierarchies allows users to organise all or selected elements of a dimension in a tree without modifying the original structure of a database. These hierarchies can be displayed in the workbook and in the reporting tool. For example, to aid the selection of the countries a hierarchy called 'by continent' was added to all databases. Using the original country nomenclature it classifies the countries by continent with a specific entry for the EU: Africa, Americas, Asia, EU, Other Europe and Oceania. Using this hierarchy, users can rapidly select the EU Member States or deselect

them if for example they want to select the 10 biggest wheat producers in the world, considering the EU-27 as a unique region. Another example is that for Eurostat databases with information at NUTS2 level, a hierarchy called 'only MS' was added for users not interested in the regional level. Hierarchies can also be used to create a list of the most frequently used elements. A 'simple AGLINK' hierarchy was added to the very long list of attributes to help non-AGLINK modellers to find the information they are looking for.

# 6 Data harmonisation

## 6.1 A pragmatic approach

Harmonisation is achieved by linking database nomenclatures to a common dictionary. The common dictionaries enable the user to work with data from different sources under one unique nomenclature (the common one). In building up the common dictionaries (name, code and unit) a pragmatic approach is followed.

First, the common dictionaries were limited to the items mostly used in the AGRILIFE Unit but progressively the dictionaries are being complemented. It may occur that the elements do not correspond exactly to each other, e.g. the marketable milk production and the total milk production; or China including Hong Kong and excluding Hong Kong. Or the price of wheat of bread-making quality may be linked to an average wheat price. Annual prices calculated over the marketing year may also be linked to the calendar year average. The link is nevertheless established because users are interested in comparing what is available in the different databases. Sometimes some calculations can be defined to ensure a more appropriate correspondence, like for the units, but it is not always possible.

As a consequence of the pragmatic approach, the user needs to be reminded that they should not always compare apples with apples. Thus meta-information is crucial and DataM allows it to be collected and displayed, e.g. with remarks in the attribute descriptions.

It is also possible to map different elements of the original database to the same element of the common dictionary. It can be useful when the attribute does not have the same name according to the commodity. For example, in the DG AGRI outlook the common 'Production' element is mapped to the 'Usable production' used for crops, the 'Production' used for meat and the 'Total production' used for dairy products. Moreover, one original element can be mapped to two common elements; it is used when the 'Food use' and the 'Domestic use' are identical because certain databases display only one of the two attributes and no other 'Use' attribute is available.

## 6.2 The common dictionaries

There are 11 common dictionaries: commodities, attributes, countries, NUTS, price indicators, trade flow, trade facts, HS commodities, CPA<sup>9</sup> 2002, year of release and units. In the Metadata definition tool the elements of each dimension are listed, together with the name, the code, the unit and meta-information.

Countries are organised by continent with a specific entry for the EU (as for the 'by continent' hierarchies) and the names of the countries are as per the UN classification. The NUTS nomenclature is separated from the country dictionary because most of the databases contain information only at country level and NUTS are specific to the EU, as well as because selecting countries in a simpler list is easier. Trade flows are only two: imports and exports. Trade facts are the quantity and the value. The Harmonized Commodity Description and Coding System (HS) of tariff nomenclature of 2007 is used as common dictionary for all the trade databases by tariff nomenclature.

The common attributes are the most important elements of the balance sheet, as well as yields, prices, macroeconomic data and income/labour indicators. This dictionary is displayed in Annex 1. A common unit for the attributes is defined in the common dictionary (1000 t, 1000 ha, t/ha, etc.). The common unit for biofuels is the million t.o.e.<sup>10</sup> and for sugar w.s.e.

The common list of commodities (see Figure 6.1) already covers close to 300 elements. Nevertheless, only a few fruit and vegetables are covered so far. The names and codes of the common attributes and commodities were decided by the administrators with the intention of adopting the labels most frequently used in the different databases.

Another criterion in defining labels was to keep them short to avoid further formatting work in the tables and graphs, e.g.

<sup>9</sup> CPA is a statistical classification of products by activity. The CPA products are assigned to the economic activities of the NACE classification. CPA is the underlying nomenclature of input-output tables of Eurostat, which is used in many cases to construct social accounting matrices, the first building block of CGE models. CPA is currently mapped with AgroSAM's product list.

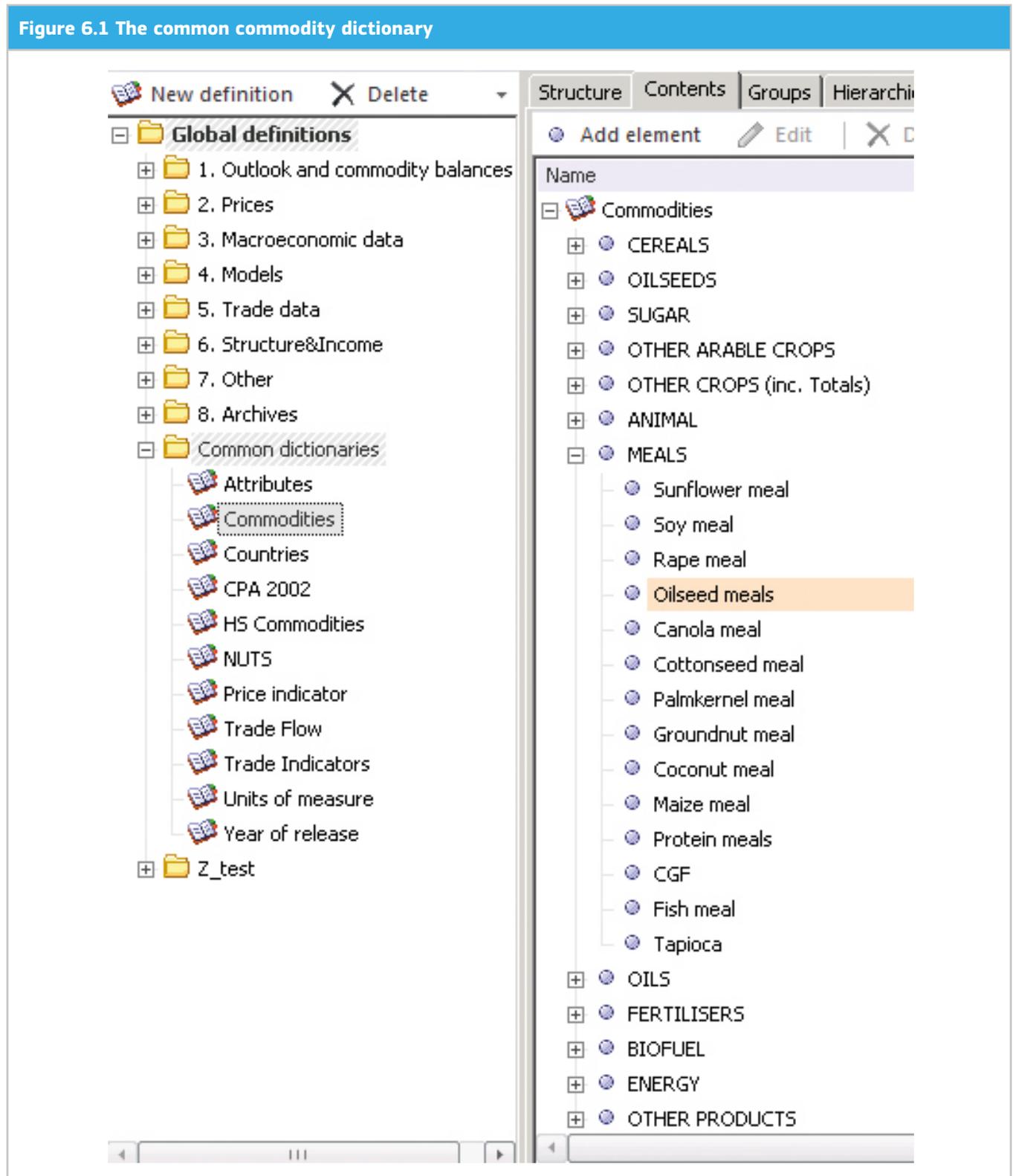
<sup>10</sup> The coefficients of conversion used for ethanol are 0.645 t.o.e./t and 0.5016 t.o.e./m<sup>3</sup> and for biodiesel 0.884 t.o.e./t and 0.7882 t.o.e./m<sup>3</sup> respectively.

'EU-27' is used instead of 'European Union (27 countries)'. In addition, improving the efficiency of the search is an important aim. As a consequence manioc is labelled 'Manioc (cassava)', the code of 'Maize' is 'Corn' or the code of 'EU-27' is 'EU27'.

The common list of price indicators was created from the metadata of different databases to link world prices. For a given commodity, it allows different world prices commonly used as reference for a given commodity to be linked. The most used price indicator for a commodity is labelled 'World reference price' and these specific price indicators are displayed in Annex 2.

The HS of tariff nomenclature of 2007 is used as the common dictionary for all the trade databases by tariff nomenclature.

Figure 6.1 The common commodity dictionary



## 6.3 Linking databases to the common dictionaries

In the Metadata definition tool, the harmonisation is managed by dimension. A double click on the original element of the database opens a window where the common element can be selected from the list of elements of the common dictionaries. In case the link is not direct, meaning a calculation such as a unit conversion is required, elements are added to the original list under a sub-folder called 'Common items'. For example in DG AGRI Market Outlook, the 'Imports' are not mapped directly to 'Imports' of the common dictionary because the units have to be changed. Therefore a new element called 'Imports (common)' is created and linked to the common dictionary and in the comments the implemented calculation is explained. The final step is then to define the transformation rule (refer to chapter 4.6).

In order for the linked elements to be seen from the workbook or the reporting tool under their common name, a common hierarchy has to be defined for each dimension of each database. This hierarchy is created automatically.

In order to map COMEXT and COMTRADE to the databases focusing on commodity balances, groups of tariff nomenclatures are being created progressively. It started with the COMEXT database for which groups' definitions were collected close to the DG AGRI units. The trade by commodity is calculated in value and product weight but also in carcass weight equivalent and grain equivalent.

For prices, at first no conversion to a single currency like the Euro was foreseen as users such as researchers may prefer to decide which exchange rate to use (average, end of year, source, etc.). Nevertheless, in the continuing data harmonisation process, in the most important price databases conversion to Euro and USD is gradually being introduced. As far as possible the annual AMECO and the monthly Eurostat average exchange rates are used for the conversion. In addition when monthly data is available like in the World Bank price database, the average prices over the marketing year are added. In that case a dimension is added with three elements: calendar year, marketing year and not available.

## 6.4 Examples of benefits for using the common dictionaries

The most obvious and direct benefit of the common dictionaries is the efficiency of the data search. For example a search of the 14 Outlook and commodity balance databases for 'EU-27 maize production<sup>11</sup>' (or 'EU27 corn production') using the common nomenclature will return results for 8 databases while a search in the original classifications returns no results for 'EU-27 maize production' and 3 results for 'EU27 maize production'. In addition, after a search on the common nomenclature the datasets are displayed in the Workbook directly in the comparable common unit.

In the workbook and in the reporting tool users can display the common hierarchies to select the required elements more rapidly. Moreover, the common labels can be displayed in the data area of the workbook.

After the original nomenclature is linked to the common dictionary, a 'common' data cube can be created from the original database. The available elements are limited to the common items which are listed under their common names. Using these 'common cubes' in the reporting tool allows dynamic reports from several databases to be created. For example, in a graph comparing the USDA projections, the OECD-FAO Outlook and FAPRI projections two control boxes on the countries and commodities can be added so that users can choose the elements to be graphed.

Finally, the database linkage enables the creation of 'harmonised datasets (HD)'. The difference in comparison to the 'common cubes' is that in a HD the database becomes a dimension whereas a 'common cube' contains data for a unique dataset. Furthermore, the number of dimensions is limited to the common ones whereas it is specifically defined for each HD. For example in the HD dedicated to the models, for AGLINK-COSIMO the dimension showing if the variables are endogenous or exogenous is dropped. These HD are the databases accessible from the DataM web application which is currently under development.

<sup>11</sup> In the common nomenclature the maize commodity is labelled maize and the code is corn.



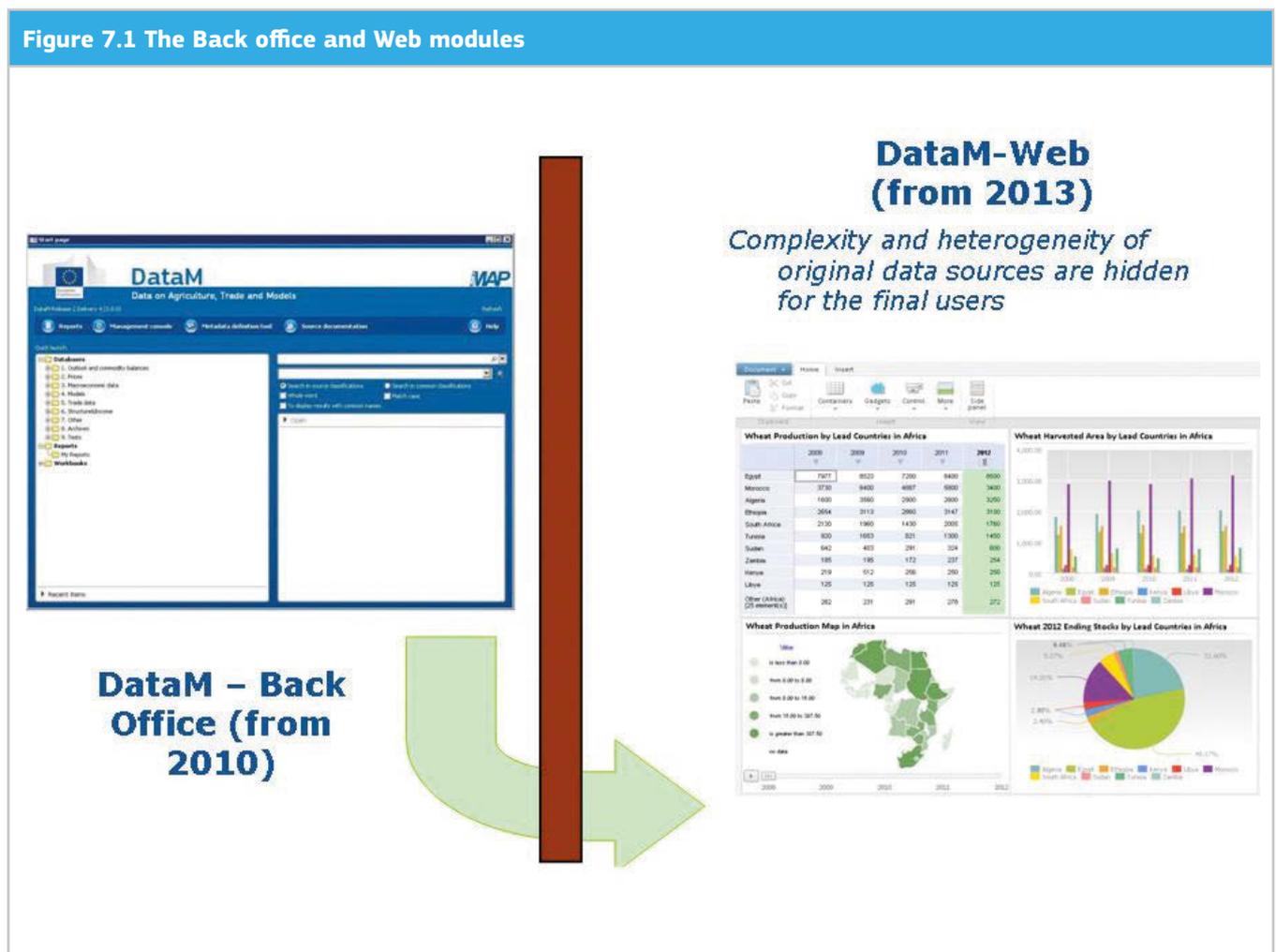
# 7 Technical infrastructure

DataM is a computer application based on a meta-software layer, the so-called 'platform', a creation of the software company PROGNOZ, enabling them to develop similar customised applications for different clients, with reduced production costs.

In DataM there is a distinction between the two main modules:

1. "DataM Back Office", a powerful analytical suite to be used for the researcher and analysts within the Commission offices
2. "DataM Web", an online portal sharing enriched views of the data created with the BackOffice application.

Figure 7.1 The Back office and Web modules



## 7.1 Back office application

The back office application is entirely hosted by the JRC IPTS. The PROGNOZ platform runs on Windows and needs to interact with a Database Management System (DBMS). In the case of DataM, the database engine is Oracle running on a Linux machine, which is in line with the standard technologies in use at the European Commission.

DataM is in principle a two-tier client-server application; the presentation and application logic layers are concentrated in the Windows executable whereas the data layer is managed by the database server. After various experiences with different configurations, the optimal one that has been adopted consists of installing the platform on a powerful application server, accessed by the client PCs via RDP (Windows' Remote Desktop Protocol). This solution emulates the three-tier pattern of software engineering that separates the software layers responsible for the presentation (User

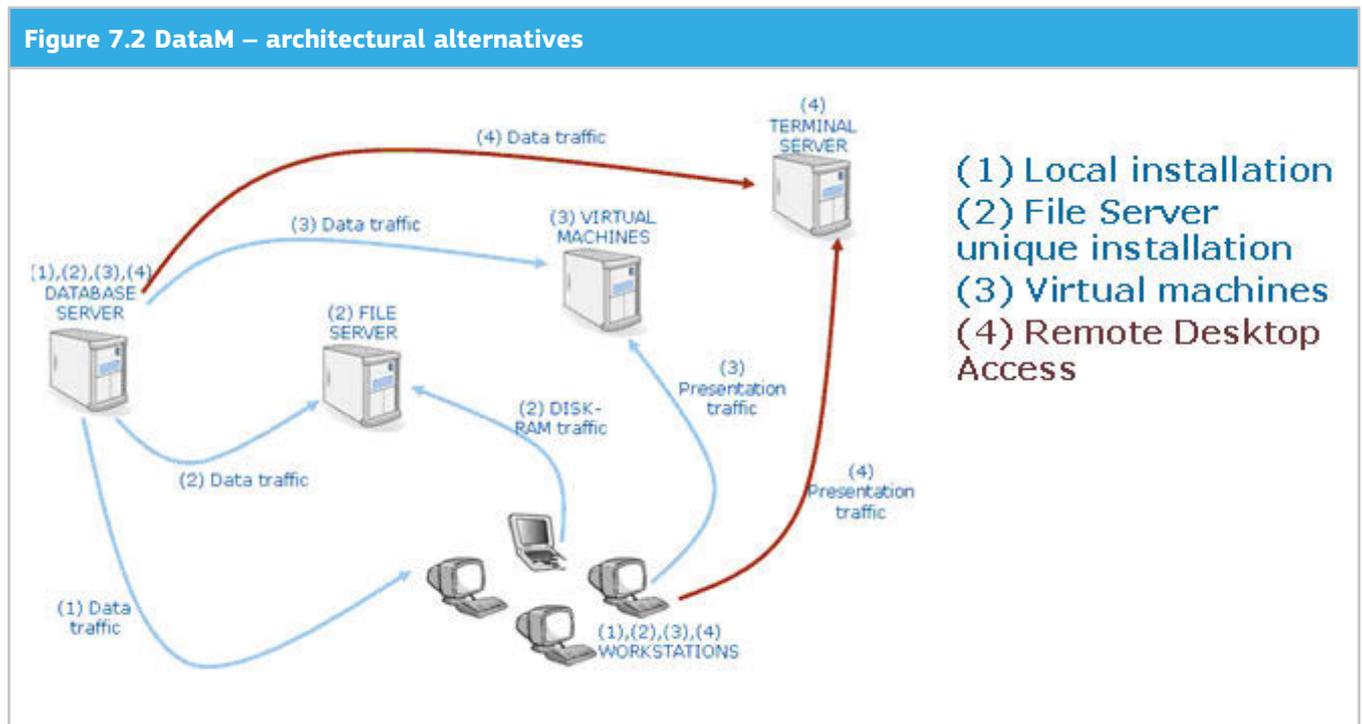
Interface), the application logic (Platform) and the data access (DBMS).

Through the Remote Desktop access, the 'terminals' of the final users take over the elaboration workload of the user interface. The resulting advantages are not only in terms of performance but also given by the optimised maintenance of the installation, concentrated on the application server. Excellent performance is also ensured for connections from remote locations over the Commission network, e.g. for the DataM users located in DG AGRI in Brussels.

## 7.2 Web application

The hosting of the web application is completely outsourced. The database underlying the web application is an extraction, periodically refreshed, of the full database of the back office application.

Figure 7.2 DataM – architectural alternatives



- (1) Local installation
- (2) File Server unique installation
- (3) Virtual machines
- (4) Remote Desktop Access

# 8 Data flow in DataM

## 8.1 A data warehouse solution

Computer applications are classically divided into OLAP and OLTP, and DataM falls into the category of the 'data warehouses' or OLAP (On-line Analytical Processing), meaning software oriented to analytical purposes.

The OLAP family differs from the other, wider, class: the OLTP (On-Line Transactional Processing). OLTP encompasses the computer applications created for operational purposes, serving day-to-day business processes (e.g. accounting applications, banking systems, etc.). In OLTP applications, data is normally created through data entry, and is rapidly used following the served business workflows.

OLAP, conversely, accumulates data from external sources (that can be other OLAP, OLTP or random outcomes of other information systems) after specific processing. Data in OLAP is in general used much later than its creation. OLAP often maintain historical data over large time periods. Typical OLAP processes involve the research and/or the condensation of information of huge data volumes (2GB in the case of DataM).

OLAP software such as DataM are characterised by certain standard concepts:

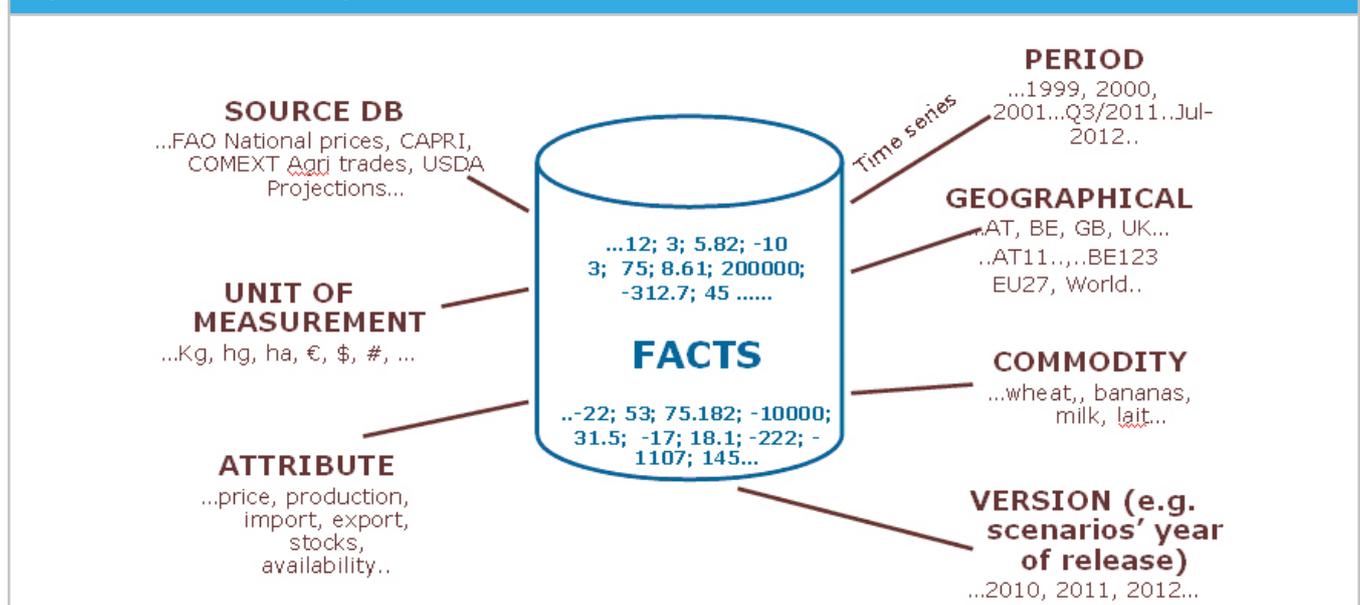
1. Hyper-cubes and star diagram
2. Data marts
3. The ETL (Extracting - Transforming - Loading data)

These three concepts will be rapidly presented, by starting from the first one, concerning what the final users see, up to the last one, concerning the data processing, in order to give a rational overview of the data life-cycle in DataM.

## 8.2 Hyper-Cubes

Every data warehouse structures its data in a specific form that favours analysis, the so-called 'star diagram' where the 'facts' (figures) are the points of a space with a number of dimensions.

Figure 8.1 DataM star diagram



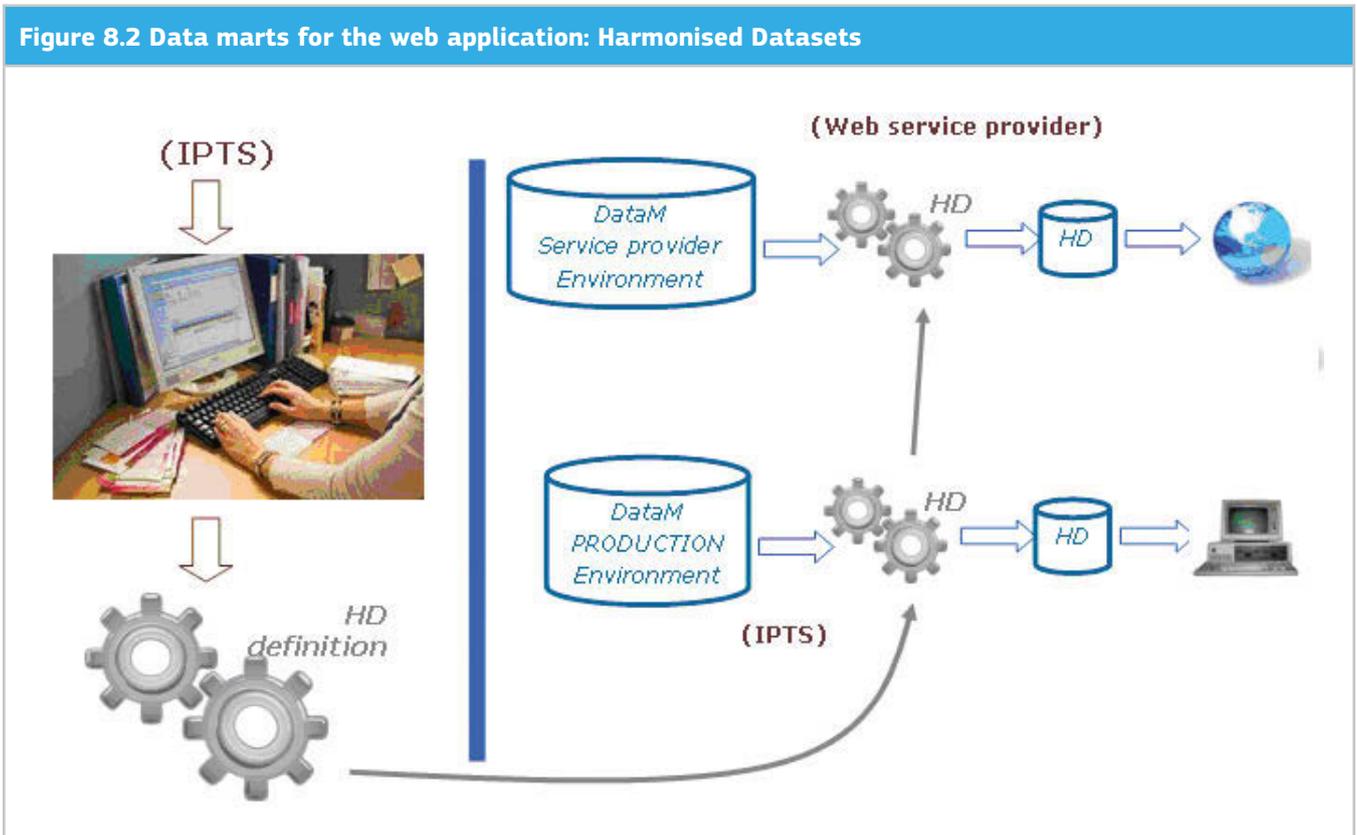
In DataM workbooks, the navigation on data is based on this concept of an n-dimensional cube, where the dimensions are presented as the exchangeable levels of a tree structure. In tabular representations of data, dimensions are grouped in a hierarchical way as headers of rows and columns. In the workbooks and related charts, the most manipulated object is the 'time series', the mono-dimensional vector of facts across the only time dimension, obtained by fixing the value of all the other dimensions.

### 8.3 Data marts

OLAP systems present users with 'data marts' that are 'views' of the data oriented to the needs of the final users. These views are obtained by elaborating the input data. DataM's major strength consists of presenting data from many sources in a comparable way, thanks to the meticulous data entry work carried out at the JRC IPTS for their harmonisation (see chapter 6).

Thus, the data marts of DataM are the original data sources integrating the knowledge data in order to show common attributes, nomenclatures and units of measurement in addition to the original values. A special case of data marts are the 'Harmonised Datasets', simplified selections of data combining various sources referring to an homogeneous business domain. The web portal of DataM is fully based on these marts.

Figure 8.2 Data marts for the web application: Harmonised Datasets



## 8.4 ETL, Extracting - Transforming - Loading

Every data warehouse needs a mechanism to digest data from outside.

Conceptually, three phases are clearly identifiable:

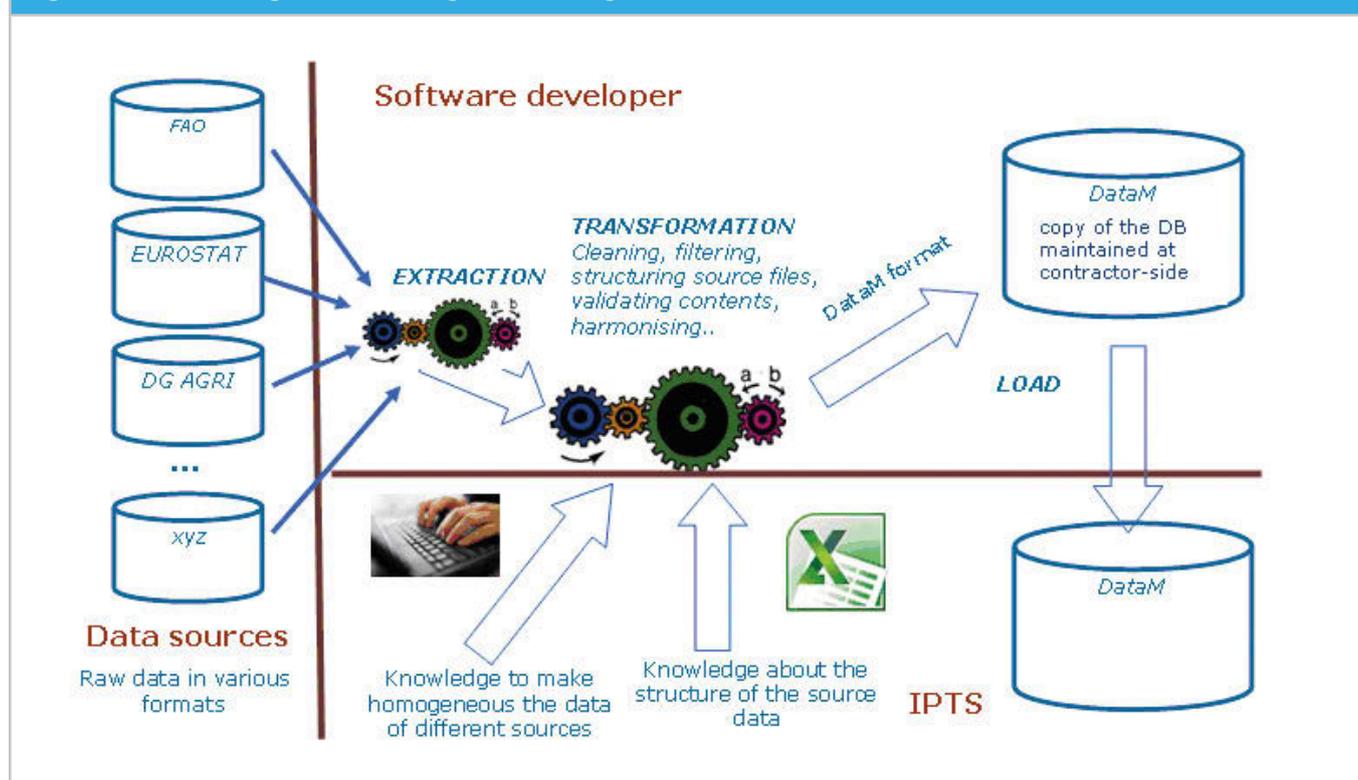
1. **Extraction** - The data comes from external sources that exist independently from the data warehouse. It is necessary to write computer programs to extract all and only the data necessary for the data warehouse.
2. **Transformation** - A data warehouse has its own data structure that foresees specific fields, levels of aggregation and standard codifications. In general, these do not coincide with the structure that can be obtained by the extraction from the source. A computer program for the transformation is therefore necessary.
3. **Load** - Once data is obtained and adapted to a format adequate for the data warehouse, a third piece of software integrates it with the existing data in the application.

In DataM, the software developer's service includes responsibility for a big share of such efforts. PROGNOZ retrieves data from variegated original sources, often in the form of web pages, and adapts it for DataM. The only exception is when data are not publicly available but the JRC IPTS owns the licence: in this case only the JRC IPTS can make the 'extraction'. Beside this, the JRC IPTS executes only the data load. This is based on data packages that the software developer transmits after having executed the whole ETL chain from their side.

## 8.5 Synchronisation

The outsourced execution of the first phases of the ETL is in general terms advantageous; even though it implies certain complications. Besides retrieving data subject to licence, the JRC IPTS does not execute anything else for the Extraction and Transformation; however these phases are based on crucial JRC IPTS contributions for interpreting the structure of data at source and for the harmonisation rules. The service provider is therefore obliged to maintain a synchronised copy of the DataM database in order to constantly integrate the evolutions in the knowledge data produced by the JRC IPTS prior to executing the transformation processes.

Figure 8.3 Extracting, Transforming and Loading data





## 9 DataM – summary and next steps

DataM has proven to be a useful database management tool simplifying the daily data work of analysts and also supporting economic modellers. In particular, it eases access to manifold datasets, facilitates data visualisation and dissemination, and can also support users in feeding economic models with data, checking statistical information or analysing results.

A specific feature of DataM is its ability to link different databases by their nomenclatures, so that users can quickly retrieve information from different sources without specific knowledge of the nomenclature used by the data providers. This consolidation, and to a certain extent harmonisation, of data shall contribute to the call for transparent data. In this context the link between data and (economic) models adds another dimension for the analysis – information which is often not available and a prerequisite for ex-ante policy assessment.

DataM was conceived and developed to mainly serve the analysts and modellers of the agro-economic modelling platform iMAP. However, during the development and pilot phase the advantages of DataM for many other users such as researchers, market analysts and policy-makers in the public and private sector became clear. For that reason the entire JRC IPTS AGRILIFE unit and pilot users from the Quantitative analysis unit in DG AGRI were envisaged as new users.

The global coverage of databases in DataM also offers various fields of applications to users outside the European Commission. To this end, a strategic partnership with the

African Development Bank (AfDB) has been agreed. As part of the Capacity Building Initiative for African Agricultural Statistics, the AfDB and JRC IPTS, in cooperation with PROGNOZ, are releasing an analytical application of DataM for all African countries. The application provides free access to most of the commonly used data sources on agriculture and tools for analytical purposes.

The DataM desktop application is a rather sophisticated tool conceived for client-server use within an enterprise's local area network or extranet. The adaptation to cloud computing and the access for external subjects might be envisaged in the future taking into account the related issues of security, ergonomics and performance. Therefore a web-based version<sup>12</sup> of DataM is being developed which is directly linked to a subset of databases built into the desktop. The web portal eases the access for registered DataM users to the main datasets via the internet and creates the potential for a multi-user environment. Moreover, in a further step it could allow the dissemination of data, e.g. of research projects, to a wider audience.

The DataM web portal is a browser-based application for analysing and visualising data from a variety of different datasets. It also aims to satisfy the needs of different purposes. A broad range of information and data can be disseminated to different users, distinguished by open or restricted access. Figure 9.1 shows the pilot interface of the DataM web portal and some thematic datasets, which group corresponding datasets. A search function enables the user to respond to ad-hoc data queries.

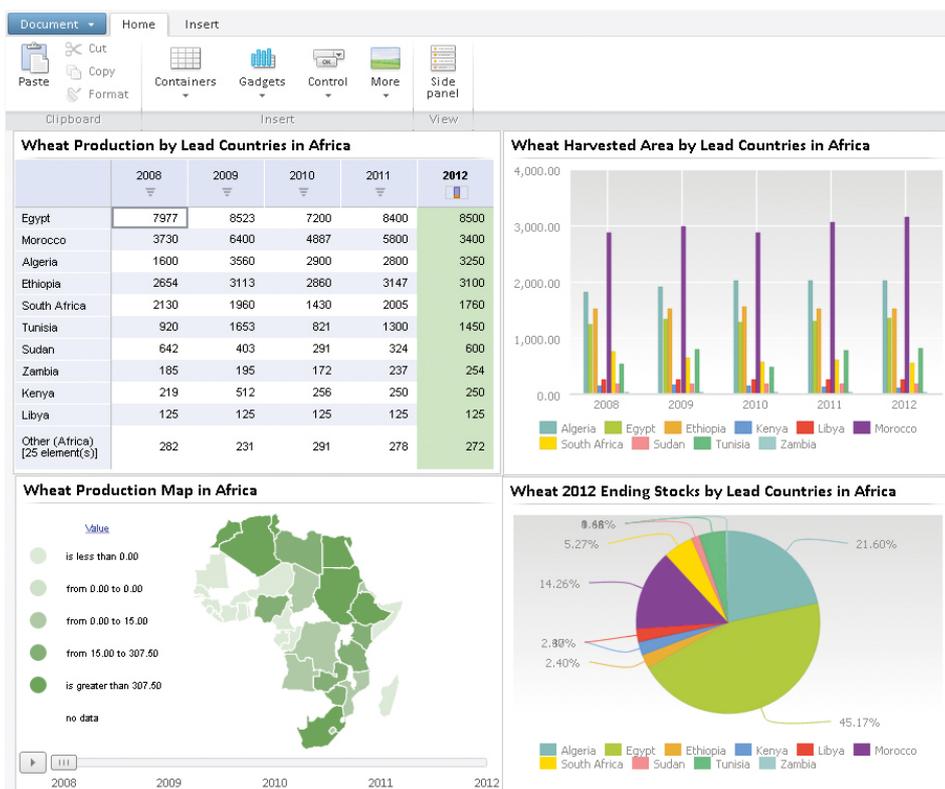
<sup>12</sup> [www.datamweb.com](http://www.datamweb.com).

**Figure 9.1 DataM web portal and thematic datasets**



The web application provides end users with a user-friendly interface to produce 'print ready' outcomes without having strong reporting skills (so-called 'dashboard designer'). The data can be visualised in maps, tables, charts and text (see Figure 9.2).

**Figure 9.2 Visualisation of data**



Predefined reports will gather not only a subset of data series from several original data sources around a specific topic, but will also provide the information in an attractive and ready-to-use way (see Figure 9.2). Among the reports, there will be for example Projections for agricultural markets, allowing the user to contrast different outlooks for production, consumption, prices, exports, and imports of different regions. A series of reports is under construction, for example on Food security-related issues which could comprise food balances as well as information about crop yield growth by country and commodity, self-sufficiency in food production, and information on future market prices.

DataM could also serve as a Visualisation portal for results of research projects. The Seventh Framework Programme (FP7) of the European Commission also finances research projects dealing with food security and international agro-food markets where the DataM web portal could be used as a stable platform for the dissemination of the project's outcome.

The DataM desktop and web portal have been developed in a public-private partnership as an institutional setting which combines research-oriented focus and inspiration

with tailored IT engineering and professionalism. In practice, the JRC as DataM's owner provided a licence to PROGNOZ for commercialising the application for public and private institutions. User licences for the tool will be offered to interested parties, together with additional services (e.g. training, documentation, maintenance and technical support). This approach should enable DataM to be maintained and further developed.

As a tool co-developed by a public institution, DataM is already available to pilot researchers within the European Commission and the African Development Bank and should also be accessible under preferential terms to users in other public institutions. Current and future users are invited to contribute to expand the link of different databases through common nomenclatures and improve the meta-information of data. Evidently, and as is the case for all intellectual property and as recognition for several years of work, DataM's use as a tailor-made software hub for databases as well as the individual databases must be correctly cited. Scientific data work is the basis for sound analysis and evidence-based policy-making. DataM, as a product of the JRC's agro-economic modelling platform iMAP, is conceived to contribute to this call.



# Annex 1: List of the balance sheet common attributes

NAME	UNIT	META-INFORMATION
Harvested area	1000 HA	
Inventory	1000 HEADS	at the end of the year
Heads slaughtered	1000 HEADS	Only for Meat
Production	1000 T	For meat gross production
Live Exports	1000 T	Only for Meat
Live Imports	1000 T	Only for Meat
Net production	1000 T	For animals = slaughterings
Food use	1000 T	
Seed use	1000 T	
Feed use	1000 T	
Biofuel use (or processing)	1000 T	
Farm use	1000 T	
Crushing	1000 T	For oilseeds crushed in oils and meals
Other use	1000 T	Item introduced to close the balance when it is not feed, food, seed or biofuel use
Domestic use	1000 T	= Production + Imports - Exports + Stock change
Exports	1000 T	in carcass weight equivalent or grain equivalent, for animals excluding live exports
Net exports	1000 T	Total exports - Total imports
Imports	1000 T	in carcass weight equivalent or grain equivalent, for animals excluding live imports
Beginning stocks	1000 T	
Stock variation	1000 T	
Cons. per capita	KG/CAPITA	
Ending stocks	1000 T	
Gross production	1000 T	For meat; Gross production = slaughterings - live imports + live exports (in cwe)
Marketable production	1000 T	Production - farm use
Deliveries	1000 T	Milk deliveries
% Fat content	%	Only for Meat
% Protein content	%	Only for Meat
Meat Exports	1000 T CWE	
Meat imports	1000 T CWE	
Stocks-to-use ratio	%	
Energy share	%	For biofuels



# Annex 2: List of the world reference prices

---

World reference price for coarse grains - Maize, US No.2, Yellow, U.S. Gulf
World reference price - WMP, Oceania, f.o.b.
World reference price - Wheat, US No.2, Hard Red Winter ord. Prot, US Fob Gulf
World reference price - Tea
World reference price - Sunflowerseed oil, f.o.b. NW Europe
World reference price - Sunflowerseed meal, 37/38%, Argentina, c.i.f Rotterdam
World reference price - Sunflowerseed, EU, c.i.f Rotterdam or Lower Rhine
World reference price - Sugar, Raw
World reference price - Soybeans, US, c.i.f Rotterdam
World reference price - Soybean oil, f.o.b. Rotterdam
World reference price - Soybean meal, 44/45%, Argentina, c.i.f Rotterdam
World reference price - Sorghum, US No.2, Yellow, U.S. Gulf
World reference price - SMP, Oceania, f.o.b.
World reference price - Rice (Thai 100% grade B), f.o.b.
World reference price - Rapeseed Meal, f.o.b. Hamburg
World reference price - Rapeseed, c.i.f. Hamburg
World reference price - Pork (producer price), Barrows and gilt, dead weight, Minnesota
World reference price - Palmkernel meal, 21/23%, Malaysia, cif Rotterdam
World reference price - Maize, US No.2, Yellow, U.S. Gulf
World reference price - Lamb, New-Zealand
World reference price - Groundnuts, US, Runners 40/50, c.i.f. Rotterdam
World reference price - Groundnut oil (any origin), c.i.f. Rotterdam
World reference price - Ethanol, Brazil
World reference price - Coconut oil, Philippines/Indonesia, bulk, c.i.f. Rotterdam
World reference price - Cocoa, ICCO
World reference price - Chicken (producer price), US
World reference price - Cheese, Oceania, f.o.b.
World reference price - Casein, New-Zealand, f.o.b.
World reference price - Butter, Oceania, f.o.b.
World reference price - Biodiesel, Germany
World reference price - Beef (producer price), Fed Steer Price, dead weight, Nebraska
World reference price - Bananas, Latin America, Major brands, f.o.t. Europe
World reference price - Rapeseed oil, Dutch, f.o.b. ex-mill
World reference price - Palm oil, c.i.f. NW Europe



European Commission

EUR 25976 - Joint Research Centre - Institute for Prospective Technological Studies

Title: Data on Agriculture, Trade and Models. A tool for flexible management, extension and integration of (model) databases

Author(s): Sophie H elaine, Mihaly Himics, Robert M'barek, Arnaldo Caivano

Luxembourg: Publications Office of the European Union

2013 - 42 pp. - 21.0 x 29.7 cm

EUR - Scientific and Technical Research series - ISSN 1831-9424 (online)

ISBN 978-92-79-29915-5 (pdf)

doi:10.2791/17240

#### Abstract

In 2007, JRC-IPTS started to develop concepts for data consolidation, harmonisation and management together with external partners. Building on the experience of this project, JRC-IPTS began a second initiative in 2009, together with the PROGNOZ software company to design and implement a software platform for data harmonisation and management. The resulting product, DataM, is a database management tool intended to simplify the daily data work of analysts and modellers. DataM facilitates the input of data into economic models, checking statistical information or analysing results.

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new standards, methods and tools, and sharing and transferring its know-how to the Member States and international community.

Key policy areas include: environment and climate change; energy and transport; agriculture and food security; health and consumer protection; information society and digital agenda; safety and security including nuclear; all supported through a cross-cutting and multi-disciplinary approach.