QÜESTIIÓ, vol. 21, 1 i 2, p. 233-240, 1997

STATISTICAL DATABASES: THE REFERENCE ENVIRONMENT AND THREE LAYERS PROPOSED BY EUROSTAT

ROGER DUBOIS*

Eurostat

The functions of the Eurostat information system are divided into four sectors which correspond to the various stages in the processing of data from their collection to their diffusion:

- *Production: collection, validation and storage of the data and meta-data;*
- *Storage of the reference data (acceptance of the information);*
- *Use of the reference data (visibility/security and find/deliver);*
- *Diffusion.*

*The system of acquisition and validation represents for Eurostat a major workload given the diversity of the partners and of the different data processing situations encountered. The upshot of the diversity and change in both the production and dissemination environments is the need to de-couple them. If there are n production systems and m dissemination systems, then there will be a need for n*m interfaces between them. This will lead an explosion of work to long term to have the data available to end-users. If a reference layer is inserted to de-couple the production and dissemination environments, the interface problem reduces to n + m and therefore becomes much manageable. Once the interfaces to and from the reference layer are defined, new tools, products or domains in either environment are easily added. By de-coupling production from dissemination, a reference layer enables producers the freedom that they require to choose the most appropriate tools and techniques to do their job. Producers also gain a common interface into the reference area, rather than the many different interfaces that they have to cope with at present. By providing well documented data in a standard form, the reference layer provides disseminators with better raw material for their products This enables them to provide better and more diverse products in a more timely and customer-driven, as desired in Eurostat s mission statement.*

Keywords: Architecture, reference environment, statistical information system.

*Roger Dubois. Administrator Unit A3. Information Data-Shop. Eurostat. L-2920 Luxembourg.

– Article rebut l'octubre de 1996.

– Acceptat el maig de 1997.

1. FUNCTIONAL OBJECTIVES OF THE INFORMATION SYSTEM

The functions of the Eurostat information system are divided into four sectors which correspond to the various stages in the processing of data from their collection to their diffusion:

- production: collection, validation and storage of the data and meta-data;
- storage of the reference data (acceptance of the information);
- use of the reference data (visibility/security and find/deliver);
- diffusion.

Co-operation between these four specialised sectors involves generalised functions of cataloguing and control (manipulation of data/resources management).

In terms of their nature, the functions can be classified into six categories:

- acquisition and validation of the data into six categories;
- management of the production data;
- manipulation of the data and meta-data;
- external presentation;
- associated services.

The system of acquisition and validation represents for Eurostat a major workload given the diversity of the partners and of the different data processing situations encountered. The objective for the development of this system is to minimise the workload required despite the increase in demand for statistical data. The use of EDI techniques (exchange of computerised data) is the direction to be followed in an attempt to eliminate the varieties of magnetic media, of format and of interpretation in the collection of data.

The management of the production data has to allow the migration of the various types of objects which combine to form Eurostat s information systems. The Nomenclatures are called to play an essential role in the creation, the homogenisation and then the documentation of the statistical data; they constitute de facto the heart of the information system.

The manipulation of the data must, in the first place, support the process of data transfer between the various environments. A number of services appears which allow the establishment of groups of users and of their associated privileges; certain meta-data become control rules in the information system.

The diffusion must consist of quality products which are well-known and are based on a platform which is easy to acquire and use (databases, statistical documents and publications, extractions from databases, magnetic media, CD-ROM, etc.).

The last two systems are activities of a general service nature which are necessary for any data-processing architecture, while emphasising particularly two very important activities for Eurostat: the STRINGS project for electronic publishing and the security of EDP applications.

2. TYPOLOGY OF THE DATA

Statistical data can be divided into two interdependent subjects:

- Statistical observations which correspond to all quantitative (and qualitative) measures associated with economic phenomena. This group is primarily composed of numerical values associated with information on their temporal evolution.

  The principal types of projects which make it possible to organise and to structure this set of data are:

  – vectors and matrices of numbers;

  – time series;

  – tables with one or more dimensions.

- All the statistical meta-data relating to the characterisation of economic phenomena and of statistical observations.

  An item of statistical meta-data is characteristically associated with one or a collection of statistical observations so as to specify the significance of the related data.

The principal types of objects which characterise meta-data are:

– The Nomenclatures and the dimensions, which allow the organisation, the classification and the association of economic phenomena and statistical observations.

– The properties and documentation, which allow the characterisation and the documentation of the phenomena and the observations.

Models: which define the processes (of derivation or of validation) on the statistical observations.

It is very difficult to give a single definition of the meta-data concept and its representation for the following reasons:

– The meta-data serves as a link between a concrete entity, the statistical observation, and an abstract entity the economic phenomenon associated with the observation.

  Note: the temporal evolution of these meta-data is the complexity factor.

– The meta-data concept may be used either «to classify» or «to seek» a subsets of observations. The hypotheses which underlie these two activities can be different and give rise to specific meta-data.

– In addition to the «link» aspect, meta-data characterise the «properties» of a subset of information.

For example, the meta-data associated with a time series are:

$\Longrightarrow$ the economic phenomenon whose evolution is defined by the series;

$\Longrightarrow$ the frequency, the source, the type of series (flow, stock, etc.), the unit, the order of magnitude, etc.;

$\Longrightarrow$ one or more comments or descriptive texts on one or other aspect of the series.

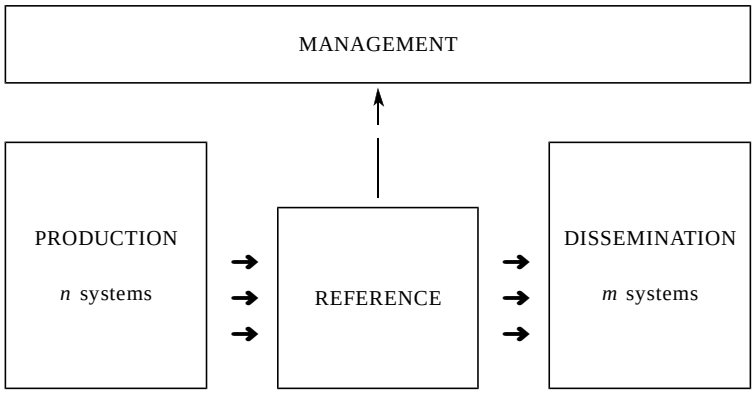
3. ARCHITECTURAL SOLUTION: A REFERENCE LAYER

3.1. Architecture

The upshot of the diversity and change in both the production and dissemination environments is the need to de-couple them. If there are n production systems and m dissemination systems, then there will be a need for n^*m interfaces between them. This will lead to an explosion of work to long time to become available to end-users. This is not in line with Eurostat s stated objective of timely provision of data.

If a reference layer is inserted to de-couple the production and dissemination environments, the interface problem reduces to $n+m$ and therefore becomes much more manageable. Once the interfaces to and from the reference layer are defined, new tools, products or domains in either environment are easily added. This enables

Eurostat to manage the evolution and change inherent in the European statistical environment, because changes to one part of either production or dissemination will not have detrimental impacts to any other part.



3.2. Potential benefits

A. For producers

By de-coupling production from dissemination, a reference layer enables producers the freedom that they require to choose the most appropriate tools and techniques to do their job. It provides well documented data to producers when they are using other domains. It also helps them to understand their own domains when they look at old data in their own domain, not only because their own memories may not be completely accurate, but because the movement of producers within the office means that expertise is not always available.

Producers also gain a common interface into the reference area, rather than the many different interfaces that they have to cope with at present. They have more mobility for other jobs with Eurostat, because the common working environment will be familiar to them in a new post.

B. For disseminators

By providing well documented data in a standard form, the reference layer provides disseminators with better raw material for their products. This enables them to provide better and more diverse products in a more timely and customer-driven manner, as desired in Eurostat s mission statement.

C. For customers

While not directly involved in the reference layer, the reference layer should provide customers with data that is:

- Visible. They will know what is available.
- Understandable. It will be documented to an agreed minimum standard.
- Accessible. They will be able to get the data more easily.
- Comparable. the standardisation of meta-data will help to enable cross-domain comparison.
- Timely. The data should be available more quickly.
- Accurate. The authorisation of data moving out of production should ensure its high quality.
- Appropriate. Feedback on usage by users should help Eurostat provide what users need.

D. For managers

A reference layer assists managers because it can provide:

- Central security and access control to Eurostat data. This ensures that data is not left on insecure and vulnerable PC platforms, and can be safeguarded as the vital resource that it is.
- Harmonisation of nomenclatures and code lists leading to less duplication of effort in capture and translation of meta-data. Producers will be able to re-use previous work by others. There will also be increased comparability of data between different domains because of harmonised definitions of their dimensions.
- Flexibility of choice in production tools and methods and in dissemination products.
- Stability of Eurostat data and meta-data over time.
- Co-ordinated market feedback to enable Eurostat to disseminate what their customers want and need. This information will come from both the reference and dissemination environments.

4. TYPOLOGY OF THE PROCESSES

The processes corresponding to one or to a number of functionalities, must occur in one of the defined environments: production, reference, diffusion.

An EDP application is seen as an ordered set of processes; the latter need not necessarily be on the same computer, but can be on two (or several) machines of the same or different platforms.

The distribution of the processing within an application over the different platforms is a difficult decision for which some rules have to be fixed:

- $\Longrightarrow$ general rules applicable throughout the European Union;
- $\Longrightarrow$ volume considerations concerning the data and the population of potential users;
- $\Longrightarrow$ security considerations requiring a more or less protected environment, access control functions, backup and recovery, archiving etc.;
- $\Longrightarrow$ the availability of suitable software on the different platforms.

4.1. Production

The production layer divides into several subjects according to the type of objects managed; precise description of the objects is needed and depends on the construction of standard applications giving a framework of utilisation to Eurostat s various producers. Eurostat has selected, on this basis, FAME for series and ACUMEN for multidimensional objects.

4.2. Reference

The «acceptance of information» poses the problem of dynamic definitions and the automated control of the passage between production and the reference. It depends on the production structures and on the complexity of the rules to be implemented. In view of the volume of existing information, only a completely automated procedure is viable.

The concept of quality of managed information is vital for Eurostat: the separation into two environments must preserve the wealth of existing information, i.e. allow more thorough interrogation of the set of information available if that is necessary.

The management of reference is concerned with a homogenous solution allowing uniform access to the set of managed information.

The visibility and security layer must ensure control of the activities of users privileges and of data confidentiality. It is important to guarantee that no leaks of information cross this layer of software.

The «FIND and DELIVER» layer allows a non-specialist user easy navigation around Eurostat s information and a rapid and exhaustive reply to his questions.

4.3. Diffusion

Eurostat disseminates statistical information by means of 3 types of products:

- $\Longrightarrow$ Publications or statistical reports, which need to mix texts, graphs and tables, using desk-top publishing products.
- $\Longrightarrow$ Databases.
- $\Longrightarrow$ Other electronic diffusion services (download, CD-ROM, etc.).