Distr.
GENERAL

Working Paper No.9
14 March 2008

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2008)
(Luxembourg, 7-9 April 2008)

Topic (ii): Statistical information systems architecture

METADATA USE IN THE STATISTICAL VALUE CHAIN - THE EUROSTAT CASE

Invited Paper

Prepared by Georges Pongas and Adam Wronski, Eurostat

Abstract

Statistical information is of paramount importance in assessing economic, social, cultural or environmental issues and building policies for the whole society.

The efficiency, precision and readability in collecting, processing and disseminating the statistical information depend at an important degree of another type of information, called metadata information.

Metadata information includes all kinds of data, used to handle meaningfully and correctly other data (data about other data). Obviously what is data and what is metadata is contextual.

This paper concentrates on the metadata needed through the major steps of the Statistical Value Chain (**SVC**), in which the statistical information is already being created, omitting hence the steps of needs' expression, data collection design, and informatics tools development. These considered steps, objective of this paper, are the secondary data collection (in the Eurostat case the primary data collection is done by the National Agencies), data editing, data processing and data dissemination.

I. INTRODUCTION

1. Metadata occupy an important place in all the phases of the SVC. Coherent and active metadata use, linked with the statistical data improve data quality, insure consistent data interpretation and finally minimize human resources necessary to achieve the collection, production and dissemination of statistical data.

2. The SVC (definition taken from: Re-engineering projects focusing on metadata and the statistical cycle, Metis 2006) is defined as a sequence of the seven following steps:

(i) Expression of the need;
(ii) Data collection design;
(iii) Specification and development of the tools needed for the data collection;
(iv) Data collection;
(v) Data editing and imputation;
(vi) Data processing;
(vii) Data dissemination;

In the paper, first we specify the technical characteristics of the metadata. Then we focus on the metadata types used in the various SVC steps and also their links with the statistical data.

II. GENERAL CONSIDERATIONS

3. In general metadata are data describing other data or processes. As such, they may describe database content, database access paths, statistical variables, complex structures such as multidimensional arrays, time series, registers, actors involved in the statistical production process, dataset construction directives, express quality judgments on content, specify confidentiality directives or even qualify or describe other metadata.

4. According to the Eurostat Metadata Common Vocabulary[1] (MVC), metadata may be classified according to the following two categories:

- Structural Metadata (SM) used to identify, describe or retrieve statistical data, or all the information needed for finding statistical information such as dimension names, variable names, dictionaries, dataset technical descriptions, dataset locations, keywords for finding the data etc. Structural metadata may be subdivided into
  - technical metadata dealing with the technical aspects of data retrieval (e.g., the names of the datasets and the tables they are stored, the dimension values to present to the users, the display formats for the variables, etc.) and
  - discovery metadata dealing with the '*statistical indexation*' of the datasets (e.g., keywords indexing the statistical variables).
- Reference Metadata (RM), describing statistical concepts (e.g., definition of variables), statistical methodologies (e.g., the way estimates are calculated) and giving information on quality of data (e.g., response rate).

5. While statistical data are fairly regular structures with, predominantly, numeric content, metadata have irregular structures with mixed content (e.g., text, numbers, files of various formats), variable attribute number, and interrelations (usually complex) between data and metadata and metadata items themselves.

A. Operational characteristics of the metadata

6. The most important operational characteristics of metadata to be taken into account when or before traversing the SVC are:

- The relatively static nature of metadata (for example once the SDDS are well written they are very slightly and rarely modified afterwards; most classifications tend to be static).
- The relatively long production duration (generally an interactive system is less important in such cases).
- The primary material is frequently located in other resources (word documents, methodology manuals, etc.) As a matter of fact, the principle of '*capture once and share afterwards*' is worth to take into account and apply.
- The criticality of some metadata entities which define the access to the statistical data, e.g., valuesets, correspondence tables. The critical entities are strongly coupled with the statistical data they are

[1] http://www.sdmx.org/docs/2006/Content_04_Draft_Guidelines_Metadata_Common_Vocabulary-%20MARCH%202006.pdf

associated with, and hence it is of paramount importance to take this fact into account when developing a system.
- The linkage between statistical data and metadata may be volatile, due to the variability of statistical content. In order to avoid *'broken links phenomena'*, it is worth considering an active metadata system (i.e., integrated with the statistical data in the sense that update operations to metadata and data are done simultaneously to assure coherence).
- The large number of different metadata entities needed in the SVC.

B. Technical Characteristics of the metadata

7. A quick examination of the metadata entity names used in various papers reveals a great and intimidating diversity. Taking the set of classifications as an example, we observe metadata entities named as: case law notes, explanatory notes, classification indexes, classification history, classification variants, valuesets, valuepools, keywords, groupings, hierarchies, annotations, node labels, units, etc.

8. We propose to look at this diversity in a simpler way by analysis of the (technical) structure, leaving out the statistical notions that the metadata entity names convey. It is believed that the statistical notion is more an issue of context and of the role the metadata are used in, rather than a factor implying a structural diversification.

9. We suggest classifying all the metadata under the following categories, according to the technical structure point of view.

B.1 Simple metadata entities (SME)

10. These entities are characterised by a simple key, and a variable number of attributes. Metadata such as classifications, glossaries, lists of users, processes' identifiers, web portal metadata (categories, perspectives, items, administrative entities, program names, publication names, Dublin Core items, SDDS documents) can be classified as simple metadata entities.

11. All entities of this type can be modelled according to the following classical vertical schema used in the Content Management Systems or e-Catalogs:
- *Entity domain name, (e.g., NACE)*
- *Entity instance identifier, (e.g. ,N1000)*
- *Entity attribute name, (e.g., English label)*
- *Attribute value, (e.g., " mining and quarries")*

B.2 Binary relationships (BR)

12. Binary relationships may be defined in two ways:
- between two different metadata entities, e.g.,
  - Statistical Agency (entity 1) and file names received from the Statistical Agency (entity 2),
  - links between dataset names (entity 1) and keywords in English (entity 2),
  - correspondence table between the CPA classification (entity 1) and the HS classification (entity 2),
  - access rights in a database defined as a qualified relationship database items (entity 1) and user names (entity 2)
- recursive, e.g.,
  - Thesauri, classification hierarchies, glossaries, index entries, relationships between regulations, directives and other legal entities, links between statistical documents..

13. Binary relationships are usually modelled according to the following schema (the typical schema of semantic network implementation):

- *Relationship Identifier, (e.g., supervisor relationship)*
- *key identifier of first entity, (e.g., John Williams)*
- *key identifier of second entity, (e.g., Brian Welles)*
- *reason of linkage (link qualifier), (e.g., approval of absences)*
- *role name of first entity, (e.g., issue)*
- *role name of second entity,. (e.g., validation)*

B.3 Clustered metadata entities (CME)

14. These entities are mostly aggregates (containers) of type BR and type SME entities as Gesmes, SDMX structure descriptions (SDS or statistical dataset structure), data set descriptions, publication descriptions, links between annotations and dataset cells, confidentiality directives, etc. The modelling of such entities may be done by combining the two previous ones with techniques of tagged hierarchies (XML like).

15. For example, in a data set containing the variable revenue with the dimensions country, year, and industry, we attach a footnote to the data cell defined by the coordinates industry = "mining", country = "PL", year = "2006" containing the text "use with caution". In this case we have a text in a relationship qualified as *footnote* attached to specific combination of the three classifications and a specific pair of data set and variable.

III. METADATA ENTITIES PARTICIPATING IN THE SVC

A. Metadata for collection.

16. In Eurostat the collection step registers the incoming data, performs basic validation, dispatches the data to the responsible unit in Eurostat and finally informs the data suppliers in case of delays or errors discovered.

17. The metadata entities involved are:
- Description and attributes of the source agency (entity of type SME).
- Metadata describing data files in SDMX terms (entity of type CME).
- Description of the deliverables as a whole including timing information acting as trigger for reminders (entity of type SME).
- Description of contacts implied in the data exchange (entity of type SME)
- The relationships between delivery instances and organisations and persons (entity of type BR).
- Basic validation directives (entity of type SME).
- Relationship between deliverables (or delivery instances) and validation directives (entity of type BR).
- Codelists necessary for basic validation (entity of type SME).
- Format descriptions (entity of type SME).
- Relationships between deliverables (or delivery instances) and formats (entity of type BR).

B. Metadata for editing and processing

18. For data processing all the information from the collection phase is needed plus the information needed to prepare the inputs for the dissemination step (data and metadata).
This includes: editing, aggregation rules and other data and metadata processing steps, extractions for publications, etc.

19. At this phase the most important metadata entities used are:
- Mapping information between collection environment and production. Usually this is achieved by metadata of type CME. It involves the mapping of the input structures to the internal database structures, the update of the workflow chain (processing of data), and the submission of the data to the editing and derivation procedures.
- Dataset definition metadata (entity of type CME).

- Valuesets definitions as subsets of classifications or valuepools (entity of type BR). (Please note that the equivalent terminology in MCV is respectively codelist and value domain (strictly speaking a valuepool is a coded domain containing the official codes of a classification together with all the derived or exception codes).
- Formulas and programs associated with the datasets, e.g., confidentiality treatment scripts, procedure to call editing software (entity of type BR).
- Descriptions of statistical data, e.g., annotations (CME), resource description such as Dublin Core (entity of type SME), or methodological content description such as SDDS documentation (entity of type SME).
- Metadata qualifiers of concepts describing the summary variables such as how to aggregate (stock, flow, value per unit), seasonal adjustment indicator, base period for indicator / estimate , measurement units, measurement period (e.g., longitudinal or flow estimate), desired output symbols (e.g., quality indicators), algorithms used for computation, etc. (entity of type SME).
- Data for dissemination (Publication) descriptions (entity of type CME).
- Transcodification information (entity of type CME).

C. Metadata for dissemination

20. The dissemination environment is of a very challenging nature given that it represents a multi-domain environment and gives access to users of various backgrounds, possibly imprecise specification of requirements and varying computational and visualisation needs.

21. In order to visualize the complexity of the dissemination environment let us have look at the following questions:

- What are the datasets including a specific variable?
- What datasets use a specific dimension and a specific classification?
- What datasets can be merged with a specific dataset?
- What datasets are linked to a specific methodology document?
- What datasets have been used to produce a specific publication?
- Which queries are impacted by the update of a specific dataset? Or, does the query need to be rerun?

22. The metadata need to provide answers to those and similar questions. Only an active and integrated (tightly coupled) statistical data and metadata environment is able to satisfy to such requirements. The use of metadata in the dissemination environment accomplishes three distinct services:

- Search and navigation services (what is available, how can it be found)
- Interpretation services (what can be done, how can the data be used or combined)
- Post-processing services (recurrent demand satisfaction, helper applications to further manipulate data).

C.1 Search and navigation metadata

23. A non exhaustive list of dissemination metadata entities is:

- Sitemap. This entity enumerates the domains for which the site offers statistical information (entity of type CME)
- Frequently asked questions entities. As the glossaries, the FAQ entities may be global or attached to a statistical domain (entity of type SME).
- Site news (entity of type SME).
- Description of the subject areas contained in the site (entity of type SME).
- Description of the statistical collection systems (entity of type SME).
- Publication descriptions in terms of contents (entity of type CME).
- Contacts per subject (entity of type BR).
- Release calendar (entity of type SME).
- Press release texts ((entity of type SME).

- The Links to other sites (entity of type BR).
- Links to active publications through means of stored queries (entity of type CME)
- Multilingual keyword lists associated to various levels of detail of statistical entities, e.g., dataset, dimension, summary variable (entity of type BR).
- **Indexes organising the site content according to different facets (entity of type BR).**
- Ready made (the most frequently used) queries per dataset (entity of type BR).

C.2 Interpretation metadata

24. These metadata constitute the information for the appropriate use of statistics. They are associated with the statistical data at various levels of detail:

- statistical domain,
- dataset,
- dimension,
- dimension element,
- cell,
- summary variable.

25. For example, the following shows a selection of metadata definitions used in dissemination, the linkages and dependency among them. It also demonstrates the possibilities of navigation in a statistical site:

- The dataset is named ‘Annual production’.
- Its dimension named Product uses the nomenclature Prodcom in its variant of 2001.
- Alternative dimensions for the Product are CPA, CPC.
- The historical evolution of the codes for this dimension is named HEPROD (a recursive relationship).
- The questionnaire that the companies fulfil is in the location http://quest.eu/prod.html.
- Technical questions may be answered by writing to john.doe@estat.eu.
- Available ready made queries are:
  - q1 - top 10 produced industrial items;
  - q2 - top 10 domestically produced items in export terms.
- The Publications: Publication 1, Publication 2, use the dataset ‘Annual production’.
- The data for the product P1 and country C1 are provisional.
- The SDDS in pdf format can by found in http://quest.eu/prod.pdf.
- The example can be extended to include the interpretation metadata:
  - The titles and usage notes of the statistical objects (entity of type SME);
  - The description of the statistical coverage (entity of type SME);
  - Observation units description (entity of type SME);
  - Classifications, valuesets (entities of type SME, BR);
  - Correspondence tables thesauri, relation ships explaining historical evolution of a classification (entity of type BR);
  - Units of measurement (entity of type SME);
  - Various annotations (entity of type SME);
  - Sources descriptions (entity of type SME);
  - Symbol explanations (entity of type SME);
  - Links between press releases and data (entity of type BR);
  - Description of methods, errors, variables (entity of type SME);
  - On-line domain specific user guides and help (entity of type SME);
  - Confidentiality related information (entity of type SME, BR).

C.3 Post-processing metadata

26. These metadata support the processing following the data extraction operations, e.g.,

- Description of available formats for download (entity of type SME, BR).
- Description of potential imports to other software (entity of type SME, BR).
- Software description itself (entity of type SME).

IV. EXPERIENCE

27. Our experience in Eurostat has shown that from the software development point of view, concentrating on the physical properties of the entities minimizes the global development effort. For example, the generic implementation of BR type entity and SME type entity are provided. The same generic BR entity functionality can be used for both correspondence table between CPA and HS and correspondence table between the list of users and list of datasets names for access contol. The same generic SME entity functionality can be used for both list of datasets and their attributes (multilingual labels, content descriptions, dates of update) and for the statistical classifications.

V. CONCLUSION

28. In the paper we have attempted to give a simple overview of the metadata entities necessary for fulfilling the needs of the computational parts of the SVC. While there is no doubt that having an active metadata system integrated with the data is a very complex task, we believe that this task can be simplified by separating the statistical notions and the structure of the metadata, the former defining the use in a specific context and the latter defining the necessary functionality to be implemented. Consequently, such approach minimises structural metadata types and thus it makes easier to build and implement a complex statistical (metadata and data) system.