WP. 23
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

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 9-11 February 2004)

Topic (i): Functions of metadata in statistical production

OPERATIONAL METADATA FOR FEDERATING STATISTICAL REFERENCE SYSTEMS AT EUROSTAT

Contributed Paper

Submitted by Eurostat, European Commission[1]

Abstract: Eurostat is using various reference systems, both for internal purposes and for external dissemination, concerning statistical data, publications and metadata. In this paper, two of these systems are considered: REFIN and Site 3. REFIN, which stands for Internal Reference, concerns all data, public or confidential, handled by Eurostat applications once they have been validated and until they will be published or disseminated. REFIN can provide access to production and reference data as they are currently stored in physical databases (physical datasets) or to derived data (virtual datasets). Site 3 is the forthcoming Eurostat Portal currently under development, which will make available on-line not only Eurostat publications (official or not) but also data tables (as pre-defined datasets or as open datasets) as well as statistical metadata (e.g. definitions, code-lists, nomenclatures, statistical guidelines, etc.). The paper describes the kinds of metadata that are handled by both systems and how these metadata are used to support the operations of these systems. In both cases, the metadata structure has been designed in a generic way to favour exchange and even interoperability with other systems within the European Statistical System (ESS).

I. INTRODUCTION

1. Statistical information in governmental agencies is one of the privileged tools used to assess economic, social, cultural or environmental situations and to build policies for the benefits of the whole society. However, data per se are not sufficient. Metadata in statistical information are necessary and play a crucial role, i.e. defining more precisely the semantics of datasets by providing contextual, usage and interpretation information about data tables and time series. Statistical metadata are data that are needed for proper production and usage of statistical data. They are essential features to make statistical data comparable, to ensure a certain level of data quality and to make statistical data more efficiently searchable [5, 6].

2. Eurostat, the Statistical Office of the European Communities, is engaged in a programme aiming at improving the way it produces, utilizes and manages metadata throughout the statistical data life cycle (herein called CVD). This especially concerns data collection, data production and data dissemination. This applies to datasets but also to publications (official or not) and to statistical metadata (e.g. statistical definitions,

[1] Prepared by G. Pongas and F. Vernadat, Eurostat, European Commission, L-2920 Luxemburg

nomenclatures, methodological notes, code-lists, etc.). There is also a need to differentiate statistical metadata from IT-oriented metadata and from dissemination-oriented metadata.

3. The paper describes current efforts deployed on two important systems under development at Eurostat: REFIN or Internal Reference, which is a reference environment federating the production databases of nearly all statistical production systems of the Office, and Site 3, the forthcoming Web-based dissemination portal.

4. The organisation of the paper is as follows. First, we describe the CVD process or data life cycle and production environment used at Eurostat and their associated metadata. Then, we focus on the Internal Reference, which gives uniform access to nearly all operational data stored at Eurostat wherever it is stored. Finally, we discuss the dissemination environment and its metadata before concluding.

II. STATISTICAL DATA LIFE CYCLE, PRODUCTION ENVIRONMENT AND METADATA

II.1. CVD

5. In Eurostat jargon, CVD stands for "Cycle de Vie des Données" that means Data Life Cycle. It is a generic business process describing all the production steps of statistical information, starting with data collected from data providers in Member States or Acceding Countries and ending with data dissemination to various audiences. The CVD process involves the following main steps [5]:

- Data Collection: It consists in transmitting datasets from Member/Acceding States and storing the data in relevant data repositories at Eurostat. Automated transmission tools such as Stadium can be used to support secured data transfers, but data can also be sent by e-mails with attachments (less than 3 Gb) or CD-ROMs. Data can be exchanged in Gesmes format or Excel, SAS, relational tables or CSV files.
- Data Validation: This step deals with checking the validity, consistency and integrity of data, and in some cases with making corrections. This is a heavily time-consuming task. Eurostat provides corporate tools (DPS, SAM) to support validation of a wide range of validation rules, but most of the production systems have their own validation module.
- Data Production and Internal Reference: This step deals with seasonal adjustment, production of the statistical aggregates (EU-12, EU-15…) and indicators (structural indicators, euro-indicators, etc.), data analysis and ensuring data confidentiality procedures.
- External Reference: This step deals with producing and loading aggregate datasets (i.e. multi-dimensional data tables) in the reference databases (NewCronos, Comext) for dissemination purposes and direct data extractions by external users of Eurostat.
- Dissemination: This covers all activities related to the production of publications (e.g. yearbooks, statistics in focus, pocket-books, etc.) and press releases, information dissemination via Internet (Eurostat Website) and product sales via the Office for Publications (OPOCE) or information relays.
- Statistical Metadata Production: Throughout the CVD process metadata defining and describing statistical information are produced. These are stored in systems such as Coded (for definitions), Ramon (for nomenclatures), NPS (for time dependent nomenclatures and correspondence tables), in production and reference databases (e.g. abstract, keywords, footnotes, flags…) and *ad hoc* descriptive text files (e.g. guidelines, legal notices, SDDS templates…).

II.2. Metadata

6. From the CVD point of view, three major families of metadata need to be differentiated [5, 6, 7]:

- Statistical metadata: This is meta-information about the terminological, contextual, legal and methodological baseline for statistical surveys and production of indicators. It deals with semantic aspects of statistical data and is targeted for the use of statisticians. Statistical metadata include: statistical definitions and glossaries, thesauri, legal notices, methodological guides, nomenclatures (e.g. NACE, NUTS, Nomenclature Combinée…), code-lists, footnotes, descriptive texts, etc.
- IT Metadata: These are characteristic items or metadata attributes used to support data processing and management of the data life cycle from an Information Technology (IT) perspective. They

include publication or dataset identifiers, date of last update, file size, mapping between logical names and physical names of files, dataset input flows, access methods to databases, etc.

- Dissemination metadata: These are also descriptive data mostly of qualitative nature. They provide the necessary information basis to classify data into taxonomies (e.g. collections, themes, categories…) to index statistical data, to manage the publication and dissemination process (who has created what, who can access what and when…) and to support search within data. They cover properties such as title, keywords, abstract, price, publication frequency, publication date, etc.

II.3. Production Systems

II.3.1 Production Systems Characteristics

7. A recent census of the applications used in Eurostat revealed the potential use of more than 250 applications, 110 of them asking for an annual development or maintenance funding. Such an environment is highly heterogeneous and complex, hence generating important maintenance costs [3].

8. The software packages or DBMSs in use include: Access, Excel, Oracle, MySql, Fame, Oracle Express, SAS, TPL, C, C++, Java, Javascript, Coldfusion or VB. Indeed not all of them are of equal importance. In fact, for historically reasons, the majority of the applications are written in Fame or Oracle Express (which replaced in the past, as a result of a 'big bang' operation, respectively, the Cronos time series DBMS and the APL language). These applications are developed using a metabase layer, which is described in the next paragraphs.

9. Applications in Eurostat are usually organised in a vertical manner. The same application insures the data validation and imputation, confidentiality treatments, aggregations, new variable creation and finally exports in various formats.

10. Provided that applications are tailor made to a given statistical domain, they cover very well the specific user needs. Nevertheless, they present obvious drawbacks such as model heterogeneity, lack of metadata standards application and important metadata redundancy.

II.3.2 Various Metadata Implementations in Production

a) The Oracle Express metabase layer

11. Oracle Express is used in more than 20 statistical applications of Eurostat. All applications are based on a common schema, as described hereafter. The metabase object types are implemented using multi-dimensional arrays.

12. The main definitions used in the metabase are:

- **Database set**: Array having a single dimension whose elements are the database names and variables (measures) are the database location, database description, database administration userid, etc.
- **User rights set**: Two-dimensional array (database X userid) giving per user the access rights for all the available databases.
- **Dimension set**: The dimension names of a given database are contained in a one-dimensional array whose name is the database name.
- **Dimension element sets**: The attributes (various labels) of the dimensions are stored in arrays whose name is DatabaseName.DimensionName and measures are the label names.
- **Flag and footnote sets**: Data may have associated footnotes and/or flags. Both are implemented as dimensions. Their association with the measures is done using arrays whose dimensionality is the same as the measure dimensionality plus the footnote or flag dimension.
- **Hierarchy sets**: In Oracle Express each dimension can be characterised by one or more named hierarchies. Hierarchies are implemented as one-dimensional measure whose name is the hierarchy name and array values that are the sons of the tree.

- **Measure set**: Each database contains one or more measures, each one being stored in a multidimensional array whose dimensionality is held by Oracle Express in native mode (not declared in the user metabase).

c) The Fame metabase layer

13. Fame applications are the most numerous at Eurostat (more than 50). Fame databases cover almost all macro-economic domains, with moderate data volumes and increased time handling requirements. As for Oracle Express every production database contains a metabase whose components describe the database. Fame databases are described using case series (i.e. ordinary vectors whose index is 1, 2, ..., n).

14. In native mode a Fame database is a file containing objects identified without ambiguity. Precise definitions follow:

- **Statistical domain**: Set of Fame databases and also a set of data collections.
- **Collection**: A set of time series of various periodicities belonging to a statistical domain and having the same dimensions. A collection can be considered as a set of one or more multi-dimensional arrays whose only difference is the time dimension.
- **Dimensions, dimension elements and labels**: These three constructs use the same mechanism for storage than Oracle Express, i.e.:

  Collectionname = {dimension names}
  Dimensionname = {dimension codes}
  Dimensionname.labelname = {labelvalues}

- **Footnotes, flags**: Footnotes and flags are implemented as parallel series using the following structure:

  Series values: Val.TimeseriesKey
  Flags, ootnotes: Flg.TimeseriesKey, Fnt.TimeseriesKey

d) Relational Database applications

15. This category includes applications that use Oracle, SAS or Access. Conversely to Oracle Express or Fame these applications are characterised by a large variety of metadata implementation features, varying from sophisticated meta-schemata to a complete lack of metadata. Whenever metadata exist they are implemented in a particular manner. Hence, the application has to be reversed engineered before its interface is written. Let us give some examples of dimension implementation:

- Code, level1, level2, level3. This represents a dimension with three levels. level1, level2 and level3 take the value 1 or 0 depending on if the code belongs or not to the corresponding level. The parent code is recovered by the code structure.
- Code labels are embedded in the fact table enhancing browsing facility and avoiding the join with a dimension table (clearly creating tables with update anomalies).
- Code, dim1, dim2…dimk. dim_i takes the value 1 or 0 depending on if the code belongs to the dimension dim_i.
- Partitioned as follows:

  Dimensionname, dimension code
  Followed by
  Dimensionname, dimensioncode, labelname, abelvalue

III. EUROSTAT's INTERNAL REFERENCE (REFIN)

16. The objective of the REFIN system is to offer an environment for virtual and loose integration of all the statistical databases independently of the model or DBMS used, presenting all the data through a common interface and insuring data naming uniqueness [3]. The advantages of such an approach are:

- Inter-database connectivity may be insured without perturbation of the production environment.
- Meta-data harmonisation is done with minimal cost.
- Complex algorithms for the same class of problems can be concentrated on a single system.

17. REFIN is a multiple component system made of:

- **Two clients** (C++ and Java)
- **REFIN server** responsible for dispatching messages (data fetch messages to particular data servers or computational messages)
- **Computational engine** responsible for performing all statistical computations
- **Data base drivers** (APIs written in C responsible for fetching data from Oracle, Fame, Oracle Express, Access and ad hoc systems that do not use any commercial software)
- **A Metabase** implemented in Oracle DBMS containing the following:
  - The location of all the remote objects and their physical type.
  - The correspondence between remote names and unique REFIN names.
  - The dimensions and dimension elements of all the remote objects.
  - C or PL/SQL program sequences in case the standard API is not enough (for example create a dimension and label by taking the projection of the corresponding columns of a table).
  - Metadata refresh information. Given that REFIN uses metadata to dynamically form queries toward the production systems, it is important to have a statistical “crowler” which uses this information in order to refresh the metadata information whenever necessary.
  - Replication information. Some databases are seldom updated. These databases are described as “mappable” to Oracle tables and refreshable upon update only. Also in some micro-data native databases, data are mapped into multiple tables (a version of transposed storage to ensure data access efficiency).
  - Segmentation information. REFIN offers the possibility to merge logically native database parts with computational results of REFIN. For example National Accounts data may be stored in Oracle Express by country, aggregates for EU are stored in REFIN tables and the whole information is presented as a multi-dimensional array under REFIN.
  - Mapping information from the various data sources to virtual cubes. Cubes are followed by dimensions, footnotes, flags and other documentation items.
    A data element may play simultaneously two roles, i.e. being a dimension but also a measure under two different names.
  - Calendar information for time series. Calendars follow the Fame frequency implementation.
  - User rights and security information. For confidential domains the IP address of the user PC is checked before allowing access.

18. Figure 1 gives an overview of the REFIN architecture with its different layers.

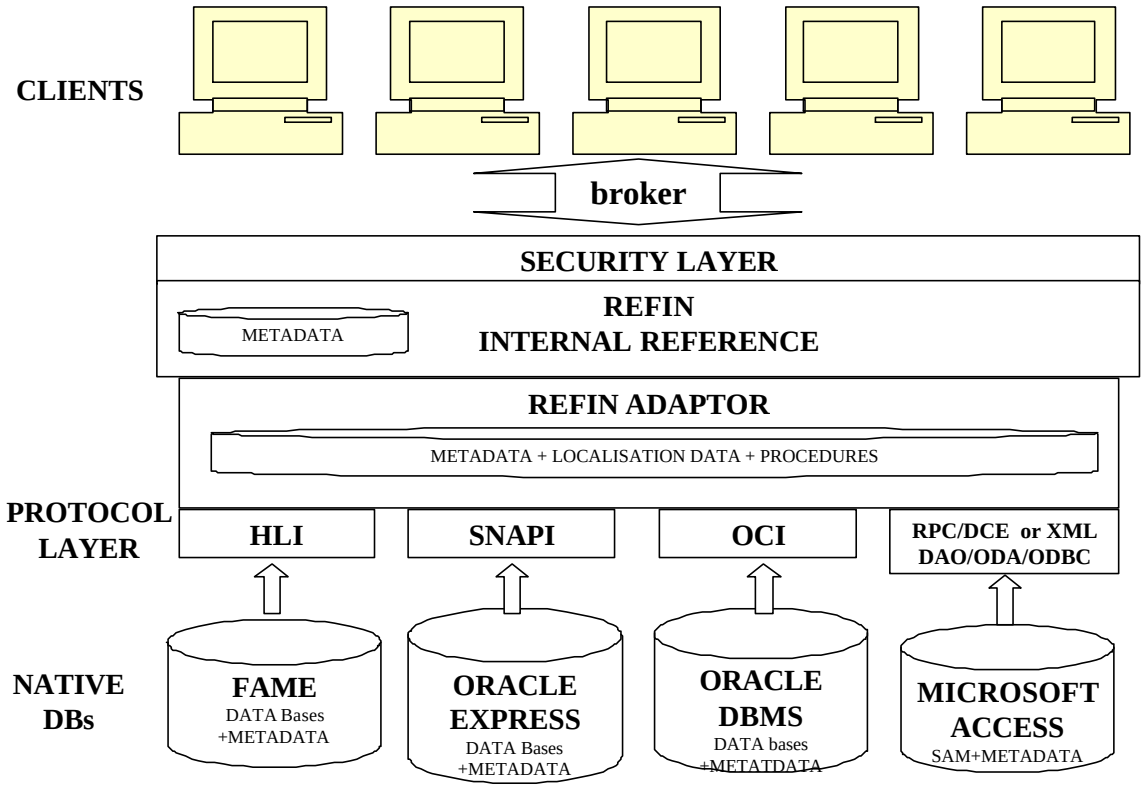

Figure 1. REFIN architectural diagram

IV. EUROSTAT DISSEMINATION PORTAL (Site 3)

IV.1. Dissemination Environment

19. The purpose of the Eurostat's dissemination environment is to make widely available EU statistical information either directly as datasets from the reference databases or as packaged finished products in the form of publications or synthesis documents. The Eurostat Website is the main gateway to this information offered to the public (see http://europa.eu.int/comm/eurostat). The future portal (Site 3 to be opened by mid-2004) will give access to:

- a large variety (about 3000 in 3 languages) of official and non-official *publications* (e.g. yearbooks, pocket-books, statistics in focus, panoramas or press releases)
- *datasets* in two forms: fixed datasets, i.e. pre-defined and pre-formatted data tables to be accessed as a whole as they are (over 1300 tables), and open datasets, i.e. selected data tables extracted from the reference databases and dimensions and subsets of which can be selected
- *statistical metadata*, i.e. statistical definitions from the CODED database, nomenclatures from RAMON/PRAMON, methodological notes and legal texts from text repositories or descriptive texts from the reference databases in the form of SDDS templates
- *links*, i.e. either publications or datasets for a given statistical theme that are stored in remote systems and only known to the Site 3 by their URL

20. All these items are collectively referred to as content objects. Each object must belong to a collection (be it a yearbook, panorama, pocket-book, dataset, etc.) and can belong to one or more themes (agriculture, fishery, trade and commerce, economy, population and society, etc.). Each content object is characterised by a fixed set of metadata attributes used for its management and retrieval.

21. Typical categories of users of the Eurostat Website include:

- EU officers (European Commission or European Parliament) and politicians

- European Central Bank (ECB), specifically interested by Euro-indicators
- National Statistical Offices (NSO) and local government agencies
- Journalists
- Researchers, academics and students
- Citizens

IV.2. Dublin Core

22. For the official and non-official publication part of Site 3 content, it has been decided to be compliant with the Dublin Core recommendations. The Dublin Core Metadata Initiative (DCMI) is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. The Dublin Core is a standard applicable to descriptive notices used in documentation information systems.

23. This international standard recommends a set of 15 metadata attributes used to characterise (textual or electronic) publication documents for management and filing purposes. These attributes concern: the document content (coverage, description, type, relation, source, subject and title), the intellectual property (contributor, creator, publisher and rights), and the document instantiation (date, format, identifier and language). Details on these attributes can be found in reference [1].

IV.3. SDDS

24. SDDS (Special Data Dissemination Standard) is a data format for explanatory metadata promoted by the International Monetary Fund (IMF) and supported by other international organisations including BIS, ECB, OECD and Eurostat [2]. It has therefore been decided to start documenting datasets stored in the reference database of Eurostat with SDDS templates for the purpose of dissemination on Internet and for a better integration with metadata disseminated by other statistical institutes in the same format. SDDS data will become metadata in the dissemination portal for fixed datasets and will give reference to statistical metadata items via hyperlinks.

25. An SDDS template is organised in a certain number of sections. Each section is made of a number of fields to be filled in with text by users. SDDS sections include:

- General information (contacts and statistical domain)
- Data (coverage characteristics, periodicity and timeliness)
- Access by the public (advance dissemination of release calendar, simultaneous release to all parties)
- Integrity (dissemination of terms and conditions under which official statistics are produced, including confidentiality, identification of internal government access to data before release, etc.)
- Quality (information the user needs to assess data quality) and
- Dissemination formats (hardcopy, electronic)

IV.5. Role of Dissemination Metadata

26. The aim of dissemination metadata in a statistical information portal is threefold: (1) to provide relevant descriptive information about products to be disseminated and to support product management, (2) to facilitate search through the content of the dissemination information system, and (3) to favour exchange and even content sharing within interoperable dissemination systems involved in a networked organisation.

27. Product description and management: Products to be disseminated need to be characterised and classified by a number of attributes in terms of their description and management. These include attributes such as product code, ISBN or ISSN numbers, title, creator, responsible unit, price, themes and collection, language, presence of tables, graphs or maps in the content, who created the product and who is responsible for the content, who is responsible for the publication. Other important information items in terms of management

concern various dates of the product life cycle, especially effectivity dates to manage embargoed content or content for which the visibility by outside users must be controlled. The list of dates includes:

- Release date: the date at which the content was released by its author(s)
- Creation date: the date at which the content was uploaded on the portal
- Start effectivity date: the date at which the content becomes effective, i.e. visible by outside users
- End effectivity date: the date at which the content can remain on the website but is no more visible by users
- Expiration date: date at which the content must be removed from the portal

28. Search: Basic search (based on text strings) and advanced search (based on combination of text strings and/or attribute values) are essential features of websites and portals, especially for information-oriented portals. In the case of Site 3, it has been decided that searching the website content will only be based on metadata attributes, not on the document content to improve the quality of answers to queries. Indexing will be based on content of a predefined list of metadata attributes (see list in the next section). Among these, important ones are product title, author(s), description or summary, keywords, themes as content categories, product type and coverage. Regarding keywords, the system does not allow more than 10 keywords. 5 keywords must come from official lists of keywords (such as the EC thesaurus named EUROVOC) and 5 can be freely defined by the authors or the authoring unit.

29. Federation of statistical dissemination systems (ESS): In a networked organisation like the European Statistical System (ESS), it is important that information stored in dissemination website of the partners of the ESS could be made interoperable to be used by various processing systems or to be syndicated to be presented by third party systems extracting contents from different sources. This is only possible if some common rules, standards and metadata attributes (such as Dublin Core or SDDS) are used by partners of the network.

IV.6. List of Metadata Used in Eurostat Web Site 3

30. The following table provides a list of metadata attributes found essential for the management, search and proper dissemination of Eurostat end-products. One column (Dublin) identifies the attributes that are compliant with the Dublin Core while another column (Mandatory) indicates attributes that must have a defined value. The last column (Domain) indicates the data type of the attribute (LOV means list of values, i.e. a fixed pre-defined set of possible values). The list of attributes is intended to apply to any type of content objects (namely for Site 3, publications, datasets, statistical metadata and links).

V. CONCLUSION

31. The paper provides a discussion of operational metadata used in two reference statistical systems at Eurostat, one to federate production systems and one to federate the dissemination environment. According to our experience and to Eurostat users, any statistical metadata information system must pay a lot of attention to:

- Importance of tightly linking data and metadata: This means that from a dataset it should be possible to have direct access to all its metadata via hyperlinks. This may concern the cells in the dataset (footnotes, flags), but also the variables and their dimensions (e.g. definitions, code-lists, nomenclatures…), the dataset itself (e.g. theme, keywords, descriptive texts…) and its domain (e.g. survey, methodological notes…).
- Importance of having a logically integrated metadata environment: This means that the various database systems used for metadata organisation and storage should be organised in a consistent and interoperable way so that they appear as a unified system to the user.
- Importance of clearly differentiating statistical metadata from IT metadata and publication/dissemination metadata. The three types of data serve different purposes and must be managed accordingly.

Site3 Attribute	Description	Dublin	Mandatory	Domain
product_code	Unique identifier of the content object	✓	✓	string
ISBN_ISSN	Official ISBN or ISSN publication code	✓	✓	string

author	Author's name(s) of the content object	✓		string
responsible_unit	Identification of Unit responsible for the availability of the content object	✓	✓	LOV
coeditor	Name of co-editing organisation, if any			string
creator	Name of user who uploaded the content object	✓	✓	string
approved_by	Name of person who approved content object upload/creation on the Website	✓	✓	string
current_version	Current version of the content object		✓	string
release_date	Issue date of content object by authoring unit	✓	✓	date
creation_date	Date of creation/upload of the content object	✓	✓	date
start_effectivity_date	Date and time at which the content object becomes visible on the Website		✓	full date
end_effectivity_date	Date and time at which the content object becomes invisible on the Website		✓	full date
expiration_date	Date at which the content object must be deleted/purged from on the Website		✓	date
theme	Theme name of content object		✓	LOV
collection	Collection name of content object		✓	LOV
language	Default language of the content object	✓	✓	LOV
other_languages	List of other languages in which a version of the content object is available	✓		LOV
table_of_contents	Name of file containing the table of contents			string
title	Official title of content object	✓		string
summary	Content object summary or official abstract	✓		string
keywords	Unordered list of keywords (maximum 10)	✓		string
freetext	Free text to add comments if needed. Not visible on the Website			string
graphs	Indicate if there is any graph attached to the content object		✓	Boolean
tables	Indicate if there is any data table in the content object		✓	Boolean
maps	Indicate if there is any map attached to the content object		✓	Boolean
cover_image	Name of file containing data for the cover image, if any			string
filename_url	URL or file name of the content object (if not physically stored on the Website)			string
related_products	Name(s) of related products/datasets	✓		string
type	Type of content object (publication, dataset, metadata, link)	✓	✓	LOV
support_format	Medium type (electronic or paper)	✓	✓	LOV
other_formats	List of other available mediums			LOV
layout_size	Size of publication format (e.g. A4, A5…)			string
page_nb	Number of pages of the content object		✓	number
table_number	Dataset identifier (logical name)			string
price	Selling price of the content object in Euro			number
out_of_stock	Indicate if publication is out of stock		✓	Boolean
update_frequency	Update frequency (e.g. daily, weekly, monthly, quarterly, yearly, biannual)			LOV
coverage	Time period covered by publication or dataset	✓		string
status	Content object status (visible, embargoed…)			LOV

References

[1] Dublin Core Metadata Initiative (DCMI), Dublin Core Metadata Element Set, Version 1.1: Reference Description, June 2003, http://dublincore.org.

[2] International Monetary Fund (IMF), Guide to the Data Dissemination Standards, Module 1: The Special Data Dissemination Standard, Washington, May 1996. http://dsbb.imf.org/Applications/web/sddshome.

[3] Pongas, G., Vernadat, F. A New Architecture for Eurostat Information Systems, Version 1.1, Unit A1, EC Eurostat, Luxemburg, 27 April 2002.

[4] Pongas, G., Vernadat, F. Data Life Cycle object model for statistical information systems. Joint ECE/Eurostat/OECD meeting on the management of statistical information systems, Geneva, Switzerland, 17-19 February 2003.
[5] UN/ECE, Guidelines for the modeling of statistical data and metadata, UNSC/ECE, CES Methodological Material, United Nations, 1995.
[6] UN/ECE, Guidelines for statistical metadata on the Internet, UNSC/ECE, CES Statistical Standards and Studies – No. 52, United Nations, 2000.
[7] UN/ECE, Best pratices in designing Websites for dissemination of statistics, UNSC/ECE, CES Methodological Material, United Nations, 2001.