WP. 25
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, 3-5 April 2006)

Topic (iv): Implementation

IMPLEMENTATION OF METASTORE AT THE OECD

Invited/Supporting Paper

Submitted by Russell Penlington & Lars Thygesen, OECD [1]

I. INTRODUCTION

1. MetaStore in the tool for managing statistical metadata in OECD. It supports a set of corporate principles for statistical metadata, giving guidance to how statistics should be described in order to allow users to use and understand the statistical data thus enhancing overall quality. The principles contain the following key notions:

- Attachment level: It must be possible to attach metadata to all levels of a statistical dataset: The whole dataset, a dimension, a dimension member, a time series, and one or more observations. Metadata should be attached at as high a level as possible to avoid duplication and inconsistencies.
- Metadata concepts. Metadata should be labelled under a common set of 41 metadata concepts which have been designed to cover all aspects of data characteristics, and to be aligned to similar concepts used in other statistical organisations; however, it is by no means mandatory to use all of these concepts for every dataset.
- Coherence across datasets: As far as possible, metadata for "the same" data should be managed in one place and referenced from other datasets.

2. MetaStore is a tool which fully supports these principles. It has been implemented progressively since early 2005. It is not a mandatory tool but dataset owners may use their own tools as long as they provide consistent metadata to the common data warehouse OECD.Stat.

3. This paper discusses the degree of success of this implementation; the difficulties encountered and measures taken; the coverage and coherence of the resulting metadata; which drivers have proven successful in the implementation; and secondary effects on visibility of OECD statistics.

[1] Prepared by Russell Penlington (russell.penlington@oecd.org) and Lars Thygesen (lars.thygesen@oecd.org)

II. OECD'S STATISTICAL INFORMATION SYSTEM

4. OECD statistical activities are carried out in several Directorates by statisticians with different backgrounds and work experience. More than 100 statistical activities are carried out by the OECD, the vast majority of which are devoted to collecting, processing and releasing data. Statistics are used daily by a wide range of OECD analysts, whose primary needs are to access data and metadata in the most efficient way. Finally, both internal and external users increasingly need to navigate across databases to carry out multi-domain analyses and comparisons.

5. Although the decentralised model has advantages, there are also disadvantages. The main problem areas are related to the efficiency of individual statistical processes and to the overall quality (in particular, coherence and methodological transparency) of OECD statistics from the user's perspective. To address these problems, a corporate Statistical Information System has been built during the period 2003-2006. It contains tools for data and metadata collection, manipulation, storage, dissemination, discovery and retrieval, preserving the independence of data producers while making their data and metadata part of a coherent and seamless corporate system.

6. In setting up of the system, the OECD has made use of best practices among peers in order to integrate such practices and as far as possible, avoid duplication of work.

7. The overall architecture of the Statistical Information System consists of three layers:
- The production layer for the collection, validation, processing and management of statistical data and metadata;
- The storage layer where validated statistics and related metadata are stored;
- The dissemination layer for producing statistical publications and online/offline interactive statistical products.

8. The system comprises five independent but inter-operating components, one of which is the metadata management system MetaStore.

III. THE METADATA PRINCIPLES

9. The metadata of international organisations basically aim to give the information necessary to understand if it is meaningful to compare macro data between countries, and to understand how much weight can be attached to such a comparison. It must therefore describe what the data mean (concepts), the overall quality of each of the data elements presented, as well as differences in quality and differences in the meaning of the data between different subject matter areas as well as between countries.

10. The metadata must primarily illuminate the following areas:
- Concepts, definitions of concepts, including such phenomena as measurement units, transformations.
- Delimitation of populations.
- Dimensions of quality, related to the original production, such as sample sizes, standard errors on estimates, and other kinds of known sources of errors such as non-response.

11. In 2004, OECD adopted a set of corporate principles laid down in *Management of Statistical Metadata at the OECD*[2]. The following principles are the most important ones.

[2] http://www.oecd.org/dataoecd/26/33/33869551.pdf

A. **Consistency**

12. Statistical metadata must be consistent. This means that:

- The same variable name, definition, and other description should be connected to the same statistics, no matter where it is and who is the "owner";
- The same variable name should not be used for statistics that are not identical;
- Terms and concepts should be consistent throughout;
- All OECD metadata, particularly reference metadata, should be made readily and freely available to external users.

B. **Redundancy**

13. Metadata on one element (a statistical collection or dataflow, a concept) should only exist as one instance; no matter how many times the same element is reused in different contexts. Ownership to the metadata should be clearly defined.

C. **Common metadata items**

14. A set of 41 metadata items are defined. All metadata from different subject-matter areas must be grouped under these headings. These are similar to the SDMX Cross-domain concepts and have been developed concurrently with those. The ambition is to have the closest possible consistency between the two sets.

D. **Attachment levels**

15. Metadata can be attached at any level of detail of the statistical data: at the global level (pertaining to all datasets), at the dataset level, at dimension level within a dataset, at dimension member level, time series, and individual observations. To ease understanding and avoid repetition of data, it is recommended to always attach metadata at the highest possible (or reasonable) level; exceptions will then have to be stored for those lower levels where they apply.

IV. METASTORE AND ITS FEATURES

16. MetaStore is a general toolkit for accessing and managing reference metadata for statistics. Statistical metadata can be separated into two categories:

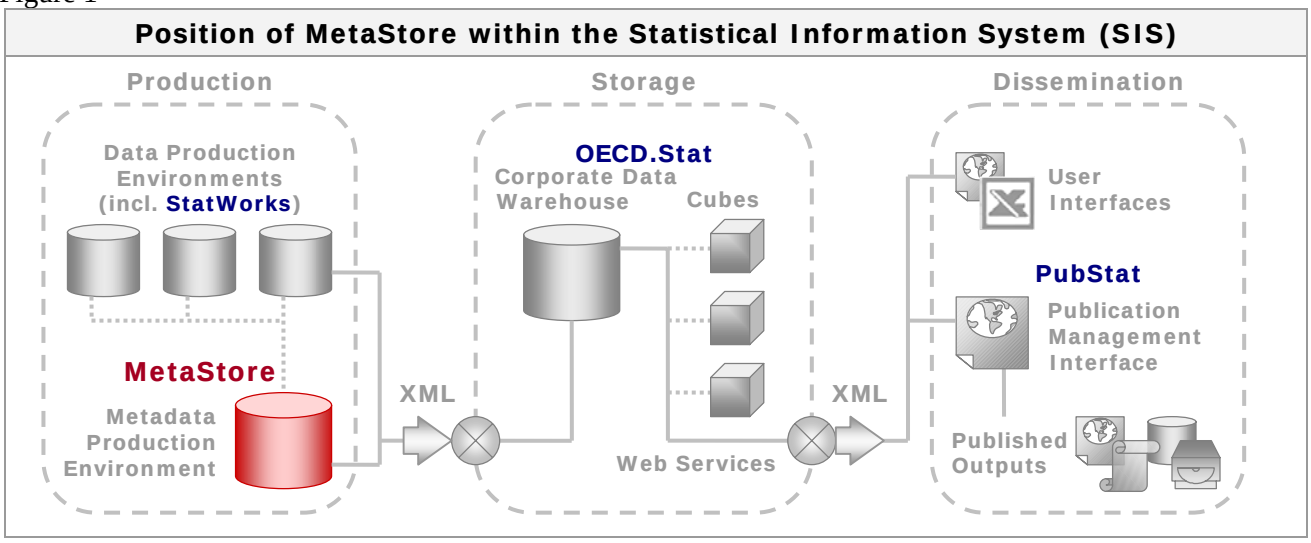
1. Structural Metadata describes the multi-dimensional structure of a dataset. Such metadata includes codes and names of dimensions and corresponding dimension members.
2. Reference Metadata is textual information documenting characteristics of the data within a dataset. Such metadata typically includes information concerning the collection, manipulation, purpose and quality of data at different levels of a dataset's multi-dimensional structure.

17. MetaStore inherits Structural Metadata from the data production systems of the datasets. While MetaStore allows the storage of customised data attachment coordinate descriptions, it does not allow the management of Structural Metadata in the datasets' production systems. MetaStore's sole purpose is the management of Reference Metadata.

18. The ability of MetaStore to remain separated from the datasets' data production systems allows it to be a common repository for managing Reference Metadata for datasets of different structures.

19. MetaStore is positioned in the production layer of the OECD Statistical Information System (SIS) for managing production reference metadata content. The following diagram outlines the position of each of components of the Statistical Information System.

Figure 1



A. **Key Features**

20. There are features and components of the MetaStore system for management of metadata both from within the web interface and from remote applications. The more notable features underpinning the core purposes of the system provide solutions to several common metadata management issues.

21. **Improving the quality of my statistical reference metadata**: While quality in statistics often refers to the data, the related metadata is an important vehicle to propel each of the dimensions of data quality. Furthermore, the same framework of quality dimensions can be equally applied to the metadata itself. Under the guidance of the Quality Framework and Guidelines for OECD Statistics (http://www.oecd.org/statistics/qualityframework), MetaStore aims to address issues for metadata accessibility, timeliness of metadata management, as well as coherence, interpretability, credibility, relevance and accuracy of metadata content.

- Metadata Attachment - MetaStore can attach metadata at any level of a dataset's structure. This is accomplished by allowing metadata to be attached at dataset and dimension levels and to all possible combinations of dimension members.

- Rich Formatting - Metadata can be entered, copied and edited with rich HTML formatting. The system's rich text formatting interface and integration with Microsoft Office allow existing content to be copied from all Office applications, as well as HTML sources such as web pages, with the formatting intact.

- Standard Classification – Metadata content must be classified into a set of common metadata item types which have been identified to improve the comparability and interpretability of metadata across datasets. Such types cover source information, collection, data characteristics, scope and coverage, statistical concepts, classifications, manipulation methods, and other aspects. Furthermore, several of the common metadata types must be chosen from a list of predefined texts or values rather than entered manually.

22. **Making metadata management more efficient**: Whilst the OECD does not regard cost-efficiency as a dimension of quality, it is a factor that must be taken into account in any analysis of quality as it can affect quality in all dimensions. If metadata can be produced and managed more efficiently with the same quality, then resources released can be used to improve quality in other areas.

- Content Sharing – Metadata content can be shared both within a dataset and between datasets. Ownership management prevents unauthorised update of content and the system automatically detects exact text matches to prevent duplication of metadata text within a single dataset.

- Links & Flags – Metadata can be further enriched by adding related hyperlinks with defined titles and links to the OECD Glossary of Statistical Terms (http://stats.oecd.org/glossary/). Items can also be flagged (on/off) for publication (whether shown in dissemination formats), archived (legacy versions), and private (draft versions for authoring user).

23. **Making metadata accessible to interested parties**: It is essential that reference metadata is accessible to both internal and external users to facilitate in the interpretability of the data for which is attached. The range of different users leads to such considerations as multiple dissemination formats and selective presentation of metadata.

- Accessibility – Metadata within the system can be accessed via a number of different methods. From remote applications, content can be directly accessed with the data coordinates loaded into the URL or extracted into the application via an ODBC query. Within the MetaStore web interface content can be accessed via broad search on coordinates and body text, dynamic filters for which criteria can be saved, and direct coordinate selection with the ability to find related coordinates with metadata attached.

- Reporting & Exporting – Metadata reports can be constructed for single data attachment coordinates or for a predefined number of levels below a specified data coordinate and output to HTML, Microsoft Word, Excel and XML for publishing. There also exists functionality to define, save and execute export criteria to feed XML formatted metadata into the OECD.Stat corporate data environment for dissemination.

24. **Integrating data production environments with MetaStore**: MetaStore is designed so that the functionality provided by the system uses components that can also be called from remote data production environments and applications. This means that the production databases can rely on MetaStore for metadata management. For detailed information on how to accomplish remote integration, please see the Remote Access in the MetaStore documentation available from the web interface.

- Connectivity - MetaStore is able to connect to and inherit the structure of any ODBC compatible database format (This includes databases such as SQL Server and Microsoft Access). Alternatively the system also has the possibility to read dimension information from text based files where a data production's dimension structure cannot be accessed via an ODBC connection.

- Remote Interaction – MetaStore is designed so that the functionality provided by the web interface uses components that can also be called from remote ODBC compliant applications. The web interface is also designed with dataset and dimension details (inherited from the data production system) embedded into key page URLs. This facilitates integration of remote applications with the actual web interface.

- Bulk Processing – MetaStore also allows for the bulk copying, moving and deleting of metadata content within a single dataset. Flexibility provided in bulk processing criteria includes wildcarding for both source and destination coordinates, and limiting to subsets of common metadata item types.

V. GOVERNANCE PRINCIPLES

25. The basic rules governing the management of metadata and the implementation of MetaStore stress the local ownership and responsibility for metadata. It is the unit responsible for collecting and managing the data – the data provider – who must also take care of metadata.

26. The use of MetaStore as the tool for managing the metadata is highly recommended in the metadata principles but is by no means mandatory. When populating the data warehouse OECD.Stat, the data provider

has to provide the metadata in the prescribed form (attachment levels, metadata items). The data provider may decide to transmit metadata directly from the proprietary production environment or from other documents to OECD.Stat. This means that in order to populate MetaStore, we are using carrots rather than sticks. Data providers must be persuaded to follow this route by attractive systems and good results. Therefore, demonstration of successful implementation of metadata for certain datasets is very important.

27. While it is evident that there is a considerable investment in populating MetaStore and organising the metadata in a new way, it must be demonstrated that, eventually, the management of the metadata can be more efficient with the facilities offered by the corporate system. Another important driver is the possibility of producing a better quality of metadata, leading to satisfied users and maybe a decrease in the efforts to support users who do not understand the data. One of the quality dimensions which can be greatly improved is the coherence of metadata between different databases. Finally, data providers should be interested in increased visibility on the Internet, see section VIII.C.

VI. POPULATING METASTORE

28. The MetaStore system is designed to facilitate populating and updating metadata content from production systems. In order to satisfy different user requirements, metadata can be either updated automatically from data production systems or entered manually through a web based interface.

29. The remote update of metadata through a parameter based procedure serves also to act as the agent for content entered manually through the interface. One of the more important parameters of this procedure for remote access is the sharing level. When metadata content is sent to the procedure one of three sharing levels can be initiated:

> 0 = No Sharing. There is no check to see if the exact text already exists in the system.
> 1 = Sharing Within Dataset. A check is made to see if the exact text already exists within the same dataset. The text is shared if a match is found.
> 2 = Sharing Across Datasets. A check is made to see if the exact text already exists within any dataset. The text is inherited if a match is found in another dataset and shared if found within the same dataset.

30. The most prevalent sharing level chosen when updating metadata from remote applications is 1 (within the same dataset). This ensures sharing of metadata texts while maintaining ownership of the content within the dataset. Such functionality also serves to automatically “clean up” repetitions within legacy metadata systems when migrating metadata to the new system. An Excel based system has been created to facilitate migration from either an Excel worksheet or delimited text file extracted from the legacy metadata system.

31. The web interface has also been constructed using a system of “friendly URLs” with the dataset identifier and data attachment coordinates built in. This enables a straight forward approach to integrating data production applications with the web interface. In practice, as users select data coordinates within the data production system, a URL can simply be built using the same coordinates to direct the user to manage metadata content at the correct level within MetaStore’s web interface.

32. The input fields of MetaStore’s web interface consist of rich text WYSIWYG (What you see id what you get) editors. These allow a small set of formatting commands including bold, underline, italic, bullets, and tables. The editors utilise a comprehensive set of client side code to automatically transform formatted content into valid XHTML. Such a “clean up” transformation is also launched upon pasting content from external sources such as Microsoft Office documents and HTML sources. This enables content to be migrated efficiently from offline sources. XHTML was chosen as a rich storage format so that it can also be rendered into XML for publishing purposes.

VII. SHOWING METADATA TO USERS

33. The metadata is always presented, in whole or partly, along with the statistical data itself through all the different media used for dissemination. Thus, all publications and off-line electronic media are provided with some edited form of the metadata.

34. The way in which metadata are presented on-line is crucial to the usefulness of OECD statistics. There are basically two ways in which users inside and outside the organisation get access to the metadata: l. stand-alone metadata that users may want to search in order to learn about potentially interesting data, and 2. along with the statistical data.

35. Stand-alone presentations of the full metadata of many datasets will be made available on the Statistics Portal on the Internet in the second quarter of 2006. This will make it searchable and increase visibility of the data behind them (see section VIII.C below). Already today, metadata for Main Economic Indicators have been made available through a special system on http://stats.oecd.org/mei/default.asp?lang=e&subject=15.

36. Together with the data, metadata is presented in the OECD.Stat Browser. Here the attachment levels will be reflected, so that metadata are shown at the level of detail where they belong. The following principles have been elaborated to present metadata in a way that will be immediately understandable to a wide audience.

37. In the OECD.Stat Browser, metadata availability is marked in the table view always with a red "i" icon ("i" stands for "information"), clicking this icon will reveal the metadata.

38. Besides metadata at the dataset and the dimension level, there exist three different metadata types:

1. Metadata for single dimension members (most often "definitions", e.g. Population coverage, Key statistical concepts used),

2. Metadata for particular observation values (for any complete combinations of dimension members; most often "exceptions") and

3. Metadata at higher-levels (for any incomplete combinations of dimension members; most often "exceptions").

39. These types will have the metadata availability mark in different places:

1. Metadata for single dimension members have a red "i" in the cell of the dimension member.

2. Metadata for particular observation values have a red "i" in the cell of the value

3. Metadata at higher levels will be marked in the following way (see figure 2 below): In order to avoid having a red "i" in all cells of a row when metadata pertain to all observations in that row, an extra column is introduced in the table view, containing a red "i" when there is a piece of metadata pertaining to all observations in the corresponding row. Reciprocally, an extra row is introduced, containing a red "i" when there is metadata pertaining to all observations in the corresponding column. These extra column and extra row will always be displayed independently on the concrete presence of such metadata.

40. An attribute "IsInheritable" to be set by the data provider is added at Dimension level. For hierarchical dimensions, when set to “true”, the presence of metadata at parent level is indicated also at all child levels, and this applies for any of the above three types of metadata.

41. A red "i": will also be shown, when relevant, in the other windows of the OECD.Stat Browser: theme/dataset selector, dimension selector and dimension member selector.

Figure 2: Extra column and row indicating higher-level metadata

Dataset: 1--Gross domestic product

		Country	United States				
		Measure	C: National currency, current prices, millions				
		Frequency	Annual				
		Time	2000	2001	2002	2003	2004
Transaction							
B1G: Gross value added; total activity			9 100 200	9 402 600	9 710 400	10 200 000	..
B1G: Gross value added; total activity	B1GA_B: Agriculture; hunting and forestry; fishing		112 100	110 600	100 500	121 300	..
	B1GC_E: Industry; including energy		1 767 200	1 697 500	1 684 800	1 776 200	..
	B1GF: Construction		430 900	464 300	473 400	495 100	..
	B1GG_I: Wholesale and retail trade; repairs; hotels and restaurants; transport		1 795 400	1 851 300	1 924 700	1 998 300	..
	B1GJ_K: Financial intermediation; real estate; renting and business activities		2 879 100	3 023 600	3 121 700	3 268 500	..
	B1GL_P: Other service activities		2 115 500	2 255 400	2 405 300	2 540 500	..

VIII. LESSONS LEARNT

A. Drivers to acceptance of MetaStore

42. MetaStore was developed to be sufficiently flexible allowing users to manage their metadata in the same way as they would within their existing system. This was done to minimise the cost for users to migrate metadata and associated management to the new system. Coupled with a range of value added features, the architecture of MetaStore has resulted in a broad acceptance and willingness to use the system.

43. Given the decentralised nature of statistics at the OECD, it was important to offer and clearly communicate the efficiency gains and quality benefits for statistical activity owners to choose MetaStore as a metadata management system.

B. Coherence and Sufficiency of metadata in MetaStore and OECD.Stat

44. The extent to which users have utilised the new features of MetaStore has varied across datasets; however the quality of metadata has significantly improved overall for datasets that have moved to the new system. The most significant improvements, as recommended within the metadata management guidelines, fall into three main areas.

1. Increasing the volume of metadata to make it more comprehensive, sufficient and relevant for users of the data to which it is attached.
2. Splitting metadata content and attaching it to more accurate and relevant data coordinates that was not possible in previous metadata management systems.
3. Increasing coherency by standardising of metadata via sharing of content both with and across datasets as well as employing a common classification of metadata concepts.

45. MetaStore manages metadata for about 145 datasets out 240 within OECD.Stat.

C. Effects on visibility

46. The growing trend to freely disseminate statistical data online presupposes that it is sufficiently visible to be accessed by interested users. The accessibility of statistical data on the web by interested users depends largely on the ability of leading search engines to index relevant metadata associated with the data. The visibility of statistical data on the web is therefore highly dependent on the content, structure, and availability of the reference metadata that describes it.

47. There are a number of important factors that search engines take into account to rank web content in relation to search phrases entered by users. The MetaStore metadata management system has been developed and optimised for these factors so that the visibility of the resulting online data content is maximised. Such strategies required for optimising data visibility online also ensures that the content and structure of statistical reference metadata serves to enhance understanding of the data by end users.

IX. PLANS FOR THE FUTURE

48. As users continue to utilise the MetaStore system and seek ways to increase the quality of reference metadata, the value added features of the system will be employed more extensively.

49. So far, users have been able to use MetaStore to mimic, to some degree, the metadata management behaviour employed within older systems. Moreover, there is no official requirement for statistical activity owners to migrate metadata and management to the new system. As a critical mass of statistical activities migrate to MetaStore, more pressure will be put on remaining statistical activity owners to move to the new system. In parallel with this, more rigid rules will be put in place within the system to further enhance the standardisation and coherence of reference metadata across datasets. Implementation of more rigid metadata management rules in the future are likely to include:

- Enforcing the reuse of exact text matches within the same dataset (currently an option).
- Enforcing content to be filled for certain combinations of attachment levels and metadata concept (for higher levels of a dataset’s structure in particular).
- Undertaking metadata quality reviews by dataset and proposing resulting actions to be taken by dataset owners.