

Distr.
GENERAL

CES/SEM.47/5
7 February 2002

ENGLISH ONLY

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE
EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint UNECE/Eurostat Seminar on Integrated Statistical
Information Systems and Related Matters (ISIS 2002)**

(17-19 April 2002, Geneva, Switzerland)

Topic I: Application of web technology to integrate statistics

**METADATA REQUIREMENTS FOR THE INTEGRATION OF A DECENTRALIZED
STATISTICAL SYSTEM AT THE OECD**

Invited paper

Submitted by the Organization for Economic Cooperation and Development¹

Abstract

Statistical activities are an essential input for the OECD's analytical work. They are decentralized in subject matter areas and coordinated by a Statistical Directorate. Over the years, decisions on the organization of statistical processes have been taken in response to the requirements of individual statistical units. This has resulted in making individual statistical activities very efficient. In general, the need for coordination between statistical activities for cross subject studies has not been considered a priority in the past. Many of the current horizontal issues addressed by the OECD involve data coming from various subject matter areas and the need for coordination is now more important. New information technologies can help in this objective but there are also a number of prerequisites in terms of cooperation tools and metadata to increase the integration of statistics. This paper describes how the OECD is putting together the information necessary to seamlessly integrate the various components of its decentralized statistical system and to help collaboration between individual statistical units. The elements described include administrative information on individual statistical activities, a glossary of statistical terms and a collection of reference statistical data. The paper concludes by presenting the OECD' experience as a case study for the integration of statistics in e-government initiatives.

I. INTRODUCTION

1. The OECD Secretariat collects and compiles a wide range of statistics for its ongoing work of helping governments of its member countries achieving their objectives. The Organization has a decentralized structure similar to that of central administrations in its Member countries. Statistical activities at the OECD are decentralized, statisticians and analysts working closely together on a subject-by-subject basis. This process has resulted in a virtuous circle of quality improvements in both statistics and analysis. However, an increasing number of policy issues studied by the OECD involve cooperation

¹ Prepared by Gérard Salou (gerard.salou@oecd.org).

between analysts and statisticians from different subject matter areas. The present statistical infrastructure of the organization makes those studies more difficult because of the lack of cooperation tools and common metadata. This paper describes how the Organization is putting together the information necessary to seamlessly integrate the various components of its decentralized statistical system and to help collaboration between individual statistical units.

II. STATISTICAL ACTIVITIES IN THE OECD INTERNAL STRUCTURE

II.1 Statistical activities of the OECD

2. The OECD is an intergovernmental organization, with 30 Member countries, which has as main objective to help member governments achieve their policy objectives. Statistics represent an essential input into the analytical work of the Organization. The primary objective of OECD statistical activities is to collect data, mainly from governments of Member Countries, and to convert them, as much as possible, onto internationally comparable basis. Statistics are then made available for internal use in policy analysis. Statistics are also an important output by themselves. Most of OECD statistics are published in both printed and electronic forms so that they are made available to the general public. In addition, the OECD co-operates with statisticians and other experts from Member Countries and other international organizations in the development of statistical systems and standards to respond to policy concerns. The OECD provides also technical assistance on statistical issues to non-member countries. The following section describes how statistical activities are integrated into the structure of the OECD.

II.2 Internal organization of statistical activities at the OECD

3. The structure of the OECD is very much similar to that of governments in Member Countries. It is divided in Directorates devoted to subject matters such as agriculture, economics, environment, social affairs, etc. Most of them have a statistical unit responsible for statistical activities related to the main policy issues covered by the Directorate. Although there is also a central Statistics Directorate, mainly responsible for economic statistics and for internal and external coordination in the area of statistics, the internal organization of statistics at the OECD is for a major part decentralized. With about 150 staff, statistical activities of the OECD are relatively modest in size when compared to those in Member Countries. The decentralized structure of statistics makes that statisticians and analysts work closely on each subject. This close relationship creates a virtuous circle of quality improvements. It ensures that statistics are kept relevant and that statisticians are responsive to the needs of policy analysis in each subject area. However, many of the current policy concerns of Member countries involve experts and statistics from several subject matter areas. The following section describes the risks and difficulties faced when working across subjects.

II.3 Risks associated with a decentralized structure

4. The decentralized structure of the organization of statistical activities at the OECD creates risks of inefficiencies along the statistical process, from the preparation of data collection to data usage and dissemination. The risks are due to the fact that, in the past, coordination between statistical activities was viewed as a burden penalizing the efficiency of individual processes, as well as to the very limited availability of resources. The most visible difficulties, from data collection to data usage and dissemination, resulting from that situation, are described in the following paragraphs.

5. Data collections are not formally coordinated which creates the risk that definitions of variables are not fully harmonized across questionnaires sent to countries. Little information is centrally available and accessible on data already collected. The existing information is not organized consistently across collections and this decreases further its corporate value. Data collection methods have evolved over the years independently of each other. Those elements create the risk of duplications, inconsistencies or

inefficiencies in data collections, from a corporate point of view.

6. Data is stored in databases collection by collection and there are few standards on variable and dimension naming. Because of the decentralized decision process, data sets have been implemented over the years in a variety of systems. Decisions on IT implementation for statistics have been taken giving priority to the efficiency of individual processes and to requirements of statisticians in individual subject matter areas. Major elements in the software infrastructure used across OECD analytical and statistical areas today are:

- MS SQL-Server 7 for data and metadata storage and cataloguing;
- ORACLE Express for multi-dimensional data-manipulation, 4GL programming and some data storage;
- FAME for time-series data manipulation, graphics, 4GL programming and data storage;
- SAS for analysis of disaggregated data;
- MS-Excel/Access for common data manipulation and some data storage.

7. This diversity of systems has resulted in a variety of different implementations. As a result, users meet the following difficulties:

- Multiple software packages and data formats create inefficient access to data
- Insufficient metadata associated with data and no metadata standard makes usage of data from different collections difficult
- Data duplication creates confusion and a risk of inconsistencies

8. On the dissemination side of the process a lot of progress has been made to increase the global value of OECD statistics. Most electronic dissemination is standardized on a unique software package and data format, used on the web as well as on CD-ROM's. However, the implementations vary and few standards are used for naming of dimensions and variables. The production processes used to create the necessary input files are developed independently of each other. The content of data files is not harmonized, with consistent data and metadata presentations. Furthermore, because data files are independent, some data series need to be duplicated and some apparent inconsistencies can occur because of differences in data vintages and of lack of metadata, for example.

III. NEW COMPONENTS OF THE INFRASTRUCTURE AND METADATA REQUIREMENTS

III.1 A vision for future statistical systems at the OECD

9. As part of a new strategy for OECD statistics developed by the OECD Chief Statistician, a vision has been developed to keep the benefits of a strong link between statistical and policy analysis activities and increase the importance of the corporate value of OECD statistics. The vision has been elaborated jointly by the Statistics Directorate and the Directorate for Information Technology and Networks with a wide consultation of OECD data users and data producers. User requirements, the role of IT and the technical considerations are described in the paper *Improving Access to Statistical information at OECD in Response to Users' Requirements*². The vision involves the development of a corporate infrastructure for statistics, including common information systems and coordination tools. The rest of the paper describes the coordinations tools identified as primary requirements in the strategy.

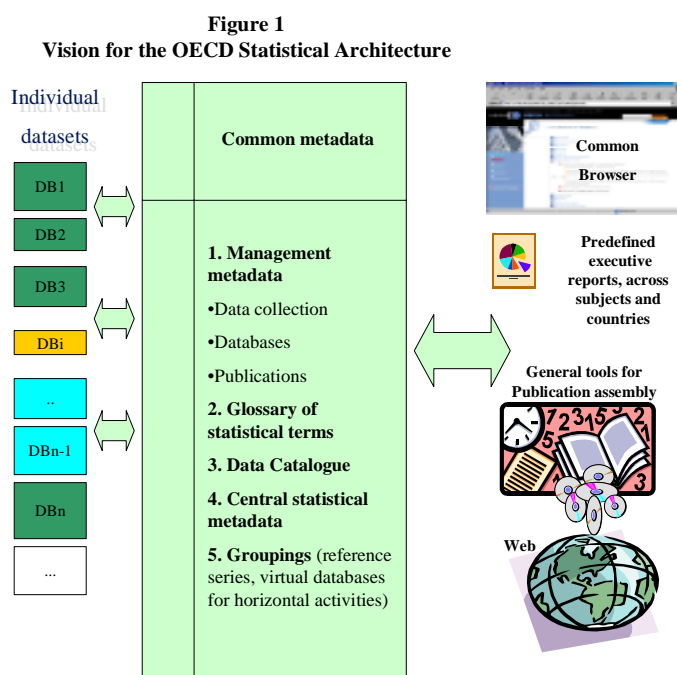
10. Figure 1, below, presents a vision of architecture for future statistical systems with emphasis in the metadata elements needed for facilitating data usage and for coordination of statistical activities. They are

² Paper by Peter Lubkert presented at the ISIS 2002 Seminar

represented in the central box and include:

- Management metadata giving high level information on statistical activities;
- A glossary of statistical terms for the harmonization of terminology and concepts;
- A central data catalogue for the location of data in the collection of OECD datasets;
- A central metadata repository for the storage of metadata elements that are independent of individual data items³; and
- Information on groupings of data: the most commonly used are publications, but we also need to identify the most commonly used data series referred to as Reference Series, or virtual databases for horizontal studies.

11. Given the important number of statistical activities and the variety of IT implementations it is too ambitious to imagine a “big bang” approach in which all data would be stored and documented in a common data warehouse in one go. That approach would probably be too IT centric and take too much time before delivering any output, which would jeopardize the credibility of the whole strategy. It has been decided to create a number of sub-projects that can rapidly deliver useful results. The rest of the paper describes the status of those sub-projects.



12. First, in order to co-ordinate activities, management metadata, including the highest level information on statistical activities is needed. This includes, for all statistical activities, the title of activity, general purpose, objectives for the current year and the year to come, person responsible, position in the OECD internal structure, position in the classification of statistical themes described below, associated data collections, databases and details on dissemination.

³ Those items are referred to by the IMF as catalogued metadata, see www.sdmx.org

13. Then, the common Glossary of Statistical Terms has been identified as a necessary condition for improving the coherency of independent activities, from collection to dissemination.

14. A classification of statistical themes has been developed to help the navigation by themes, independently of the Organization's internal structure.

15. At this stage, to maximize the usefulness of the system for the larger number of analysts, it has been decided to focus on the time series that are the most commonly used across Directorates. They are called "Reference series". Associated with Reference series, a common repository for statistical metadata is identified as a facilitator for the development of coherent metadata across the organization. The following sections describe those various coordination tools

III.2 High level information on statistical activities

16. The OECD program of work is focused on analytical activities. There is no systematic information on the program of work of the Organization in the area of statistics. This would be extremely useful for the internal coordination of collections and research activities and for the coordination with other organizations, national or international. A project has been started to compile the higher end information on OECD statistical activities. It includes general information about the activity and its future as well as technical information on associated data collections, databases and details on dissemination. This will represent an essential entry point into the future OECD common statistical system. It gives analysts and managers an overview of the scope of OECD statistical activities. Also, because it contains information on datasets, the corresponding description, purposes and persons responsible, it permits to get access to the persons who know about specific statistics. In the present infrastructure this is an essential metadata item. This sub project has made possible to produce the first integrated program of work of the OECD in the area of statistics. This is an essential tool for internal and international coordination. Also, an application has been developed to give users the possibility to navigate and search through the activities and their attributes.

17. The main difficulty met during this project was the harmonization of responses. This is due to the variety of statistical activities and their number, about 100. This work has permitted to obtain basic information to start building the statistical infrastructure. Now information on datasets and their corresponding database systems, data collections and publications together with contact persons are centrally available to any analyst in the organization.

III.3 Classification of statistical themes

18. A classification of statistical themes has been defined in order to permit access to the knowledge about OECD statistics and statistical activities that has been gathered by the sub-project described above. The classification permits to navigate the information independently of the OECD internal structure. It is also used on the OECD Statistics Portal.

III.4 OECD Glossary of statistical terms

19. The OECD Glossary contains a comprehensive set for target definitions of the main variables collected by the Organization for use in its statistical and analytical output. In addition, the Glossary contains definitions of key terminology and concepts used in OECD publications. Finally, the Glossary contains commonly used acronyms⁴.

⁴ The OECD Glossary is available on the Internet at <http://cs3-hq.oecd.org/scripts/stats/glossary/index.htm>

20. The OECD Glossary draws extensively from existing international statistical guidelines and recommendations from international organizations such as the United Nations, ILO, OECD, Eurostat and the IMF. In the main, the definitions are quoted word for word from these sources and a detailed reference provided to enable the user to refer to the complete source document to obtain further information or context where needed. The source information provided relates to the source from where the definition was extracted for inclusion in the OECD Glossary. It should be emphasized that the definitions contained in the OECD Glossary are, in particular for those relating to variables collected by the OECD, “target” definitions based on existing international statistical recommendations and guidelines. National definitions used in the actual compilation of data by OECD Member countries may (and frequently do) depart from these standards for a number of reasons. Information on national definitions, concepts, etc, for specific data collected from Member countries are normally presented in relevant OECD publications. In the future OECD statistical system, the difference between national definitions and target definitions will be stored in the central metadata repository.

21. The main aims of the OECD Glossary of Statistical Terms are to provide a:

- highly visible and readily accessible source of definitional information for use by OECD author areas in the development of questionnaires and other data collection instruments, and for inclusion in published output. The Glossary is intended here to facilitate the collection of consistent data by the various Directorates and Committees of the Organization. The Glossary is also intended to lend greater transparency to OECD data requirements from national agencies that provide information to the Organization. Obviously, such uses would also be of interest to people working in national agencies;
- set of target definitions based on existing international statistical standards that will ultimately be linked to actual data located on OECD databases. In other words, the Glossary will also be an integral part of the OECD common statistical system;
- catalyst for the development of consistent international statistical standards by international organizations working in cooperation with national agencies. The Glossary highlights areas of existing inconsistencies between existing standards and may help prevent similar occurrences in the future.

22. The main elements of the current version of the OECD Glossary are:

- unique title for the definition;
- text outlining the actual definition;
- detailed source information;
- classification of each definition into the common classification of statistical themes;
- internal cross-links to related definitions, etc., contained elsewhere in the Glossary;
- URL links to the complete source where these are currently located on websites.

23. Where more than one definition exists, a unique title has been provided through the inclusion of acronyms to identify the source of each definition in the title (SNA, Eurostat, ISIC, UN, ESA, ILO, etc.). Detailed reference information regarding the source of the definitions contained in the OECD Glossary has been provided with each definition. Furthermore, to facilitate user access to the complete source document to obtain more information about the definition, its context, etc, extensive use has been made of URLs where these documents have been located on the Internet.

24. The Glossary also includes search and interrogation facilities.

25. The next phase of the project will be to have links between data and their target definition in the Glossary. This is an important feature of the Reference series sub-project as described in the next section.

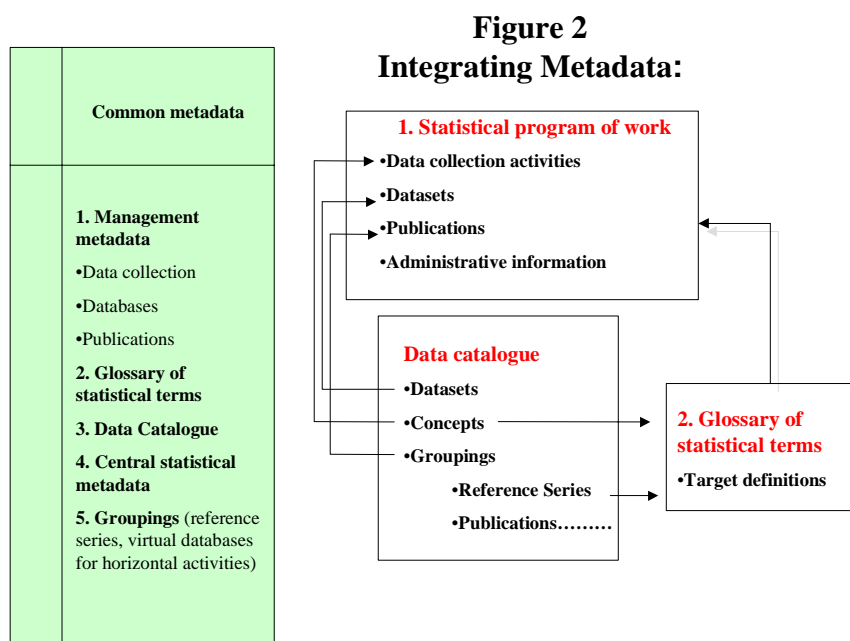
III.5 Reference series

26. In order to deliver rapidly tangible benefits to analysts in the Organization it has been decided to concentrate on data series that are the most frequently accessed by users who are not involved in the corresponding statistical activity. Analysts who are not experts in a subject matter area from which they need data have difficulties in locating those data. This is because the data they are searching are part of complex datasets containing thousands of data series. For example, GDP data are stored with all the rest of national accounts data in a complex accounting framework. In addition, individual data systems have been designed for experts, as described in the first part of this paper, and do not contain the metadata elements that non-experts need. In some cases, there is a risk that the wrong data is used. In the past, “Reference Series” were duplicated in individual databases to facilitate their use as background data to calculate ratios, per capita, etc. This was another factor of confusion and of risks of inconsistencies. An initial list of Reference Series has been obtained through consultation with analysts at the OECD. The list contains the following statistics: main aggregates of National Accounts, with history and forecasts; labour force and population data; exchange rates; purchasing power parities and price indices.

27. Reference series are also meant to define the standards in terms of associated documentation, with the objective to provide information to users who are not experts in the particular domain of the corresponding Reference series. Metadata are present at all levels of the data structure, including dimensions, elements in dimensions, crossing of dimensions and their elements. The corresponding target definitions are linked to the Glossary. An essential and new metadata element of the system is the information on usage. All metadata elements are taken from a central metadata repository.

28. Access to Reference Series is envisaged through the following means:

- ❑ Access through web browser interface;
- ❑ Access through a unique SQL stored procedure so that interfaces can be easily developed;
- ❑ Automatic link to ODBC compliant software (ODBC link in Excel for example).



29. The Reference Series Database is integrated with the other elements described in this paper, the high level information on statistical activities, the Glossary and the statistical classification, as shown in figure 2.

IV. CONCLUSION

30. The pragmatic approach described in this paper is part of the strategy developed to increase the integration of statistical activities at the OECD. The main difficulty arises from the decentralized nature of statistical activities at the OECD where statistics are mainly a resource. The same situation exists in central administrations of OECD countries with a decentralized statistical system. Statistical activities at the international level are also conducted in a decentralized manner, under the coordination of central bodies like the Conference of European Statisticians. It would be interesting to compare the OECD approach to that used at the national level in the context of e-government initiatives and to investigate how the OECD vision can be extended to help the international coordination of statistical activities.