

Distr.
GENERAL

CES/AC.71/2003/8
28 January 2003

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 ECE/Eurostat/OECD meeting on the management of statistical information systems
(Geneva, 17-19 February 2003)

Topic II: Impact of technical measures and standards on data quality

ENHANCING DATA QUALITY THROUGH DATABASE INTEGRATION AT OECD

Invited paper

Submitted by OECD¹

1. INTRODUCTION

1. The OECD has a good reputation for both the quality of its analytical work and for the statistics that underpin that work. In some areas, OECD's statistics are internationally recognised as the "best" in terms of coverage, timeliness, and comparability. However, new challenges have emerged for the OECD over the last few years due to the development of new Information and Communication Technologies (ICT) for statistics, the development of the international statistical system, budget pressures on the Organisation and the need for improving the efficiency of its statistical activities in order to devote more resources to new statistical developments. In particular, with the expansion in the Organisation's cross-cutting activities and products, there is a need for significant improvement in internal coordination across the whole range of its statistical activities – data collection, transformation, dissemination, and development.

2. To face these challenges, the Organisation launched a new strategy for statistics at the beginning of 2001. Since then, several initiatives have been undertaken to improve the quality of OECD statistics and the efficiency of statistical processes.

3. The main aims of the strategy are to improve the quality of OECD statistics and the efficiency of its statistical activities. The strategy is based on the following:

- the reputation of the OECD is also based on the quality of its statistics and that, as for other activities, the quality of statistical products depends on investments being made to improve processes used to satisfy user needs;

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

- the expression “good statistics” implies not only the usual characteristics of “good quality” statistics (coverage, timeliness, comparability, accessibility, accuracy, etc.), but also the adoption of a set of modern methods and standards for data and metadata collection, storage and dissemination;
- statistics is a multidimensional and cross-cutting activity, which must be managed and organised following an approach based on a modern “statistical information system”;
- the OECD has to provide statistics to the civil society, as part of its general political role in the current global “information society”.

4. The two pillars of the strategy are: the establishment of a quality framework and the implementation of a new statistical infrastructure. This paper describes the background at the OECD, the quality framework and describes how the new infrastructure supports the implementation of the quality framework.

II. COORDINATION WITHIN A DECENTRALISED ORGANISATION

5. The OECD is an intergovernmental organisation, with 30 Member Countries, which has as its main objective to help Member Governments achieve their policy objectives. Statistics represent an essential input to the analytical work of the Organisation. The primary objective of OECD statistical activities is to collect data, mainly from Governments of Member Countries, and to make them, as much as possible, internationally comparable. Statistics are then made available for internal use in policy analysis. Statistics are also an important output by themselves. Most OECD statistics are published in both printed and electronic media made available to the general public. In addition, the OECD cooperates with statisticians and other experts from Member Countries and other international organisations in the development of statistical systems and standards to respond to policy concerns. The OECD also provides technical assistance on statistical issues to non-member countries.

6. The structure of the OECD is very similar to that of Governments in Member Countries. It is divided into directorates devoted to subject matter such as agriculture, economics, environment, social affairs, etc. Most directorates 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 co-ordination in the area of statistics, the internal organisation of statistics at the OECD is for a major part decentralised. With about 150 staff, statistical activities of the OECD are relatively modest in size when compared to those in NSOs.

7. This organisation, based on a decentralised model, has both advantages and disadvantages. If close contact between statisticians and analysts is very beneficial for fostering dialogue between users and producers, with a positive impact especially when new statistics have to be developed, the management of a decentralised system is quite problematic. In particular, problems and risks can arise in the areas of co-ordination and efficiency in data and metadata collection and management, insufficient coherence of the statistical information published, inappropriate inferences from cross-country data, unnecessary overlap between OECD and other international organisations.

8. The OECD strategy for statistics primarily addresses those risks with the development of a corporate Quality framework and of a new statistical infrastructure. Those two elements support each other as will be shown in this paper.

III. THE OECD QUALITY FRAMEWORK

9. For an international organisation, the quality of statistics disseminated depends on two dimensions: the quality of national statistics it receives and the quality of its internal processes for collection, processing, analysis and dissemination of data and metadata. OECD statisticians have always devoted a significant part of their efforts to quality improvement at an individual level. However, the absence of a framework within

which the OECD could systematically assess, compare and improve statistics was a weakness in its statistical system as a whole.

10. The potential benefits of a common quality framework are considerable. First, it provides a systematic mechanism for ongoing identification and resolution of quality problems; second, it gives greatly increased transparency to the processes used by the OECD to assure quality; and third, it reinforces the political role of the OECD in the context of an information society.

At the beginning of 2002, the project of developing an “OECD Quality Framework” was launched and a task-force established². OECD Statisticians adopted the Framework in July 2002. Its implementation will start in 2003. The following section describes the Quality concept in the OECD context.

IV. THE QUALITY DIMENSIONS FOR STATISTICS IN THE OECD CONTEXT

11. The OECD views quality in terms of eight dimensions: relevance, accuracy, credibility, timeliness, punctuality, accessibility, interpretability and coherence. Another factor is that of cost-efficiency, which although not strictly speaking a quality dimension, is still an important consideration in the possible application of one or more of the eight dimensions cited previously to OECD statistical output. A description of the eight dimensions, applied to the OECD context, is made below.

A. Relevance

12. The relevance of data products is a qualitative assessment of the value contributed by these data. Value is characterised by the degree to which the data serves to address the purposes for which users seek them. It depends upon both the coverage of the required topics and the use of appropriate concepts. Value is further characterised by the merit of users’ purposes in terms of the OECD mandate, the agreements with Member Countries and the opportunity costs of producing the data. In the OECD context, users include the Secretariat, Committees, Member governments and other external users. The Secretariat and Committees are primary users and determine priorities, but data are also produced for external users according to the political role of the Organisation vis-à-vis the civil society.

B. Accuracy

13. The accuracy of data products is the degree to which the data correctly estimate or describe the quantities or characteristics that they are designed to measure. Accuracy refers to the closeness between the values provided and the (unknown) true values. Accuracy has many attributes and, in practical terms, there is no single aggregate or overall measure of it. Of necessity these attributes are typically measured or described in terms of the error, or the potential significance of error, introduced through individual major sources of error. In the OECD context the accuracy of the data published is largely determined by the accuracy of the data received from the contributing organisations. On the other hand, the activities carried out by the Secretariat can influence the overall accuracy of data published. This influence can be positive because the quality checks adopted by the OECD may detect errors and result in improvements to the estimates previously provided by national agencies. Or it can be negative, due to errors that may result from the collection, processing, derivation, or dissemination procedures adopted by the Secretariat.

² A lot of work has been done in recent years to apply the concept of quality to statistical data. For example, the IMF, Eurostat, Statistics Canada and other NSOs have identified various sets of data quality components and have adopted quality frameworks to improve their organisations and the quality of data produced. The OECD Quality Framework benefited from this work, adapting existing definitions and approaches to the OECD context, and developed internal procedures for assuring the quality of new activities and to improve the quality of already existing activities.

C. Credibility

14. The credibility of data products refers to confidence that users place in those products based simply on their image of the data producer, i.e., the brand image. Confidence by users is built over time. One important aspect is trust in the objectivity of the data. This implies that the data are perceived to be produced professionally in accordance with appropriate statistical standards, and that policies and practices are transparent. For example, data are not manipulated, nor their release timed in response to political pressure. In the OECD context the Secretariat has to decide if the publication of poor quality data received from countries affects the overall credibility of the OECD as high quality data provider. If the answer is affirmative, the Secretariat should refuse to publish the data. Furthermore, it must ensure that, once agreement between the Secretariat and countries has been reached on collection of specified data, the data subsequently collected cannot be withdrawn in response to political pressure.

D. Timeliness

15. The timeliness of data products reflects the length of time between their availability and the event or phenomenon they describe, but considered in the context of the time period that permits the information to be of value and still acted upon. The concept applies equally to short term or structural data; the only difference is the timeframe. In the OECD context the timeliness of the data published by the OECD is largely determined by the timeliness of the data it receives from the contributing organisations. The Secretariat itself is also a potential source of delays, which may occur during collection, processing, derivation, or dissemination.

E. Punctuality

16. The punctuality of data products implies the existence of a publication schedule and reflects the degree to which the data are released in accordance with it. A publication schedule may comprise a set of target release dates or may involve a commitment to release data within a prescribed time period from their receipt. Here “release date” refers to the date on which the data are first made publicly available, by whatever medium, typically but not inevitably, the web site. In the OECD context a publication schedule would help: external users, to improve their capacity for timely use of OECD statistics; internal users, by enhancing their capacity to plan their work based on the released dates; the Secretariat, by enhancing its capability to resist pressure to tamper with release dates for political reasons. On the other hand, there may be occasions where the OECD cannot adhere to its schedule, for example due to changes in priorities. These changes should be clearly communicated to users.

F. Accessibility

17. The accessibility of data products reflects how readily the data can be located and accessed from within the OECD data holdings. The range of different users leads to such considerations as multiple dissemination formats and selective presentation of metadata. Thus, accessibility includes the suitability of the form in which the data are available, the media of dissemination, and the availability of metadata and user support services. It also includes the affordability of the data to users in relation to its value to them, and whether the user has reasonable information that the data are available and how to access them. OECD internal and external users might have quite different perceptions of accessibility because of the differences in access tools provided for each category.

G. Interpretability

18. The interpretability of data products reflects the ease with which users may understand and properly use and analyse the data. The adequacy of the definitions of concepts, target populations, variables and terminology underlying the data, and information describing the limitations of the data, if any, largely determines the degree of interpretability. The accessibility of that information is another aspect of this criterion. The range of different users leads to such considerations as metadata presentation in layers of

increasing detail. Definitional and procedural metadata assist in interpretability: thus, the coherence of these metadata is an aspect of interpretability.

H. Coherence

19. The coherence of data products reflects the degree to which they are logically connected and mutually consistent. Coherence implies that the same term should not be used without explanation for different concepts or data items, that different terms should not be used without explanation for the same concept or data item, and that variations in methodology that might affect data values should not be made without explanation. Coherence in its loosest sense implies the data are “at least reconcilable”. For example, if two data series purporting to cover the same phenomena differ, the differences in time of recording, valuation, and coverage should be identified so that the series can be reconciled. Coherence has four important sub-dimensions: within a dataset, across datasets, over time, and across countries. For an international organisation, ensuring coherence across countries is one of the major sources of value added. The role of metadata in explaining possible changes in concepts or methodologies over time and across countries is absolutely fundamental. Unexplained inconsistencies across datasets can seriously reduce the interpretability and credibility of OECD statistics.

20. The OECD has developed this Quality Framework at the same time as the development of new statistical infrastructure to support it. The following sections describe the main aspects of the infrastructure.

V. A NEW STATISTICAL INFRASTRUCTURE AT THE OECD

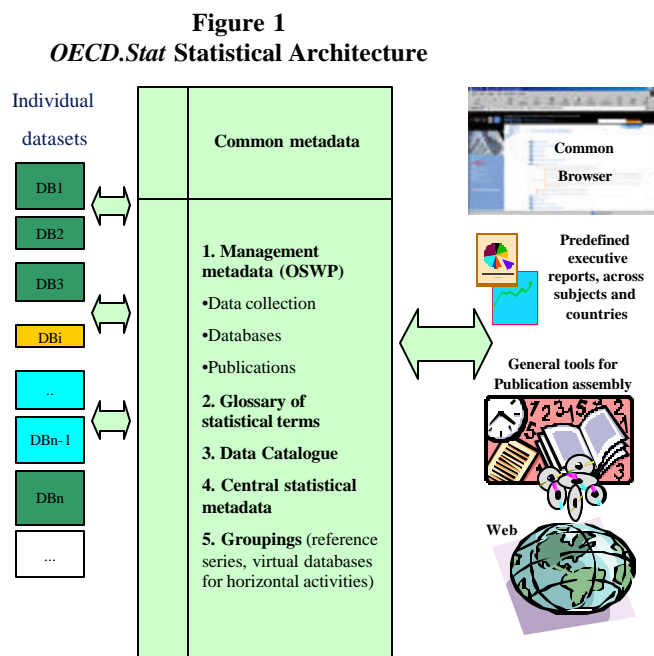
21. The new strategy for statistics is placed in the contexts of the continued modernisation of IT tools made available to OECD users and the continued search for efficiency in the usage of IT resources at the OECD. In particular, new application developments should take advantage, as much as possible, of technologies and skills already available at the OECD. The aim is to minimise the number of tools used in statistical activities in order to make training, new developments, maintenance and support more efficient. In addition, in order to reduce maintenance burden and to create synergies between different statistical domains, the strategy aims to reduce the number of specific statistical applications developed and in use throughout the Organisation.

22. In parallel with the work on Quality the OECD has started the reform of internal statistical production processes. A new statistical infrastructure has been elaborated jointly by the Statistics Directorate and the Directorate for Information Technology and Networks with wide consultation of OECD data users and data producers. The new system, called *OECD.stat*, has been designed to manage a wide range of decentralised statistical activities in a general co-ordinated framework. It tries to resolve the conflict between the traditional “single process-oriented” (or “input-oriented”) organisation of the institution and the cross-cutting information needs of statistics users.

23. In *OECD.stat*, each Directorate contributes to a corporate “data warehouse”, where all relevant data are stored with their metadata, in order to give full and easy access to all authorised internal users. Data collection activities are carried out using electronic questionnaires (often shared with other international organisations) or through direct access to databases developed by national statistical agencies. In this way, each Directorate remains independent in determining its own statistical and analytical processing on raw and final data, but the inputs and the outputs of these processes are part of the corporate data warehouse. In practice, because of the complexity and variety of statistical activities, the system is based on a “constellation” of datasets or data cubes, connected to each other and considered various parts of the single data warehouse. Central metadata elements are key in this architecture. Figure 1, below, presents a schema of the architecture for OECD statistical systems. Metadata are represented in the central box and include:

- ❑ management metadata giving detailed information on statistical activities, the OECD Statistical Program of Work (OSWP);
- ❑ a glossary of statistical terms for the harmonisation 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: publications, the set of most commonly used data series (referred to as Reference Series), and virtual databases for horizontal studies.



The next section gives more details on the main metadata elements, which are the “Heart” of *OECD.Stat*.

VI. METADATA

A. Management metadata: the OECD Statistical Programme of Work

24. The OECD Statistical Programme of Work application (OSWP) was developed as a tool for internal co-ordination of statistical activities and for external communication on OECD statistics. The Statistics Directorate has the responsibility of preparing the Programme, while the relevant Directorates provide the necessary information through an electronic questionnaire for each line of activity. It includes general information about the activity and its future as well as technical detailed information on associated data collections, databases and publications. In the present decentralised institutional infrastructure this is an essential metadata item. The OSWP information is used in two ways. In a top-down approach it permits to drill down to activities to data items, using navigation or search. In a bottom-up approach it allows to relate data items to their respective statistical activities and subsequently to all details on related data collection and dissemination. Therefore, one could say that the OSWP represents the “heart” of the OECD statistical information system

B. Common Vocabulary: OECD Glossary of statistical terms

25. The OECD Glossary contains a comprehensive set for target data element definitions of the main variables collected and derived by the Organisation for use in its statistical and analytical output. In addition,

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

the Glossary contains definitions of key terminology and concepts used throughout OECD statistical activities. Finally, the Glossary contains commonly used acronyms⁴.

26. The OECD Glossary draws its definitions from existing international statistical guidelines and recommendations from international organisations such as the United Nations, ILO, Eurostat, the IMF and the OECD itself. Definitions are quoted 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 emphasised 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 will be stored separately in *OECD.Stat*.

27. The Glossary is an integral part of the OECD statistical systems. It is used stored centrally in *OECD.Stat* as the source for target definitions associated to variables. It is then possible to relate target definitions to actual data items located on OECD databases. It is also possible to navigate *OECD.Stat* through the Glossary.

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

- unique title for the definition;
- the actual definition;
- for some definitions, text providing further background on the definition, its application and relation to similar or related concepts. This field may also contain URLs to relevant documents describing appropriate use of the variable defined, etc;
- detailed source information;
- classification of each definition to a broad statistical theme;
- internal cross-links to related definitions, etc., contained elsewhere in the Glossary;
- URL links to the complete source document containing the definition where this is currently located on the websites of international organisations or national agencies.

29. 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.

C. Groupings: Reference series

30. In order to rapidly deliver tangible benefits to analysts in the Organisation it has been decided to concentrate on data series that are the most frequently accessed across the whole Organisation. Because individual data series are often part of complex and voluminous datasets, users who are not experts in the corresponding subject matter area, have difficulties in locating the data series they search. 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 and, often, do not provide easy access to the metadata elements that non-experts would need. 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

⁴The OECD Glossary is available on the Internet at <http://www.oecd.org/statistics/glossary>

another factor of confusion and of risks of inconsistencies. An initial list of Reference Series has been obtained through consultation with analysts throughout 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.

31. 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 information on usage. All metadata elements are taken from a central metadata repository.

32. The three elements described above are key in the integration of OECD data sets in the sense that they link them together. A common IT infrastructure has been put in place to realise the concrete integration. It is described in the next section.

VII. TECHNICAL IMPLEMENTATION

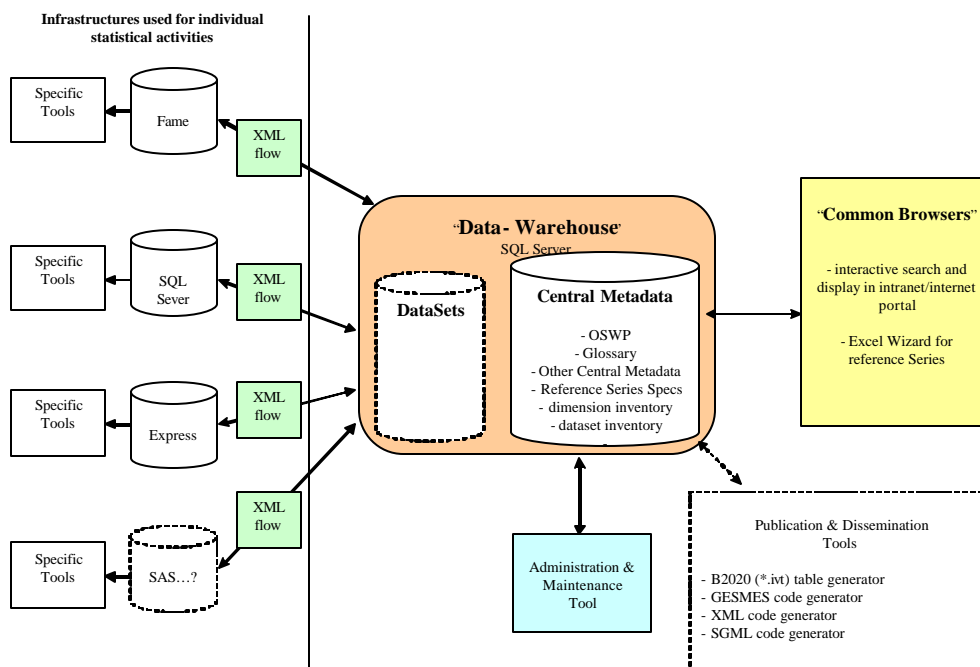
33. In technical terms, *OECD.Stat* is developed around Microsoft tools. MS SQL Server 2000 and its OLAP component are used as central data and metadata repository. Datasets are stored as cubes in that structure, with common dimensions taken from a central repository. Each cube is related to the relevant item from the OSWP. Elements of dimensions, in general those orthogonal to country and time are related to the corresponding entry in the Glossary.

34. Data flows between systems used by individual production areas and the central systems use an internal XML format.

35. An Excel Wizard has been developed to permit easy access to Reference Series. A web-based interface is on development as general user interface to the entire system. Finally, web services are being developed for delivering data outside the OECD.

36. Figure 2, below, gives a schematic representation of the technical implementation of *OECD.Stat*.

**Figure 2,
OECD.Stat Technical Architecture**



VIII. REINFORCEMENT OF QUALITY THROUGH THE IMPLEMENTATION OF THE NEW ARCHITECTURE

37. The new architecture and the Quality Framework support each other in the following ways. In preserving the individual production areas in Directorates, the new architecture preserves the benefits of the decentralised organisation. In addition, and more importantly, it improves the capabilities for cross cutting studies by permitting navigation and selections across datasets and by permitting the creation of virtual datasets through a bookmark system. This aspect reinforces the relevance of OECD statistics.

38. Accessibility is the most obvious quality dimension immediately improved by the new statistical infrastructure. This is done through several ways:

- the OSWP permits to locate statistical activities through navigation and a powerful search engine;
- the new technology has unified access to data and no special software packages and skill are required any longer. Microsoft Excel has been used for Reference Series to make access to those series as easy as possible for non-statistical staff.

39. Interpretability is improved by central tools for data documentation and by improved accessibility to metadata. The Glossary, central metadata items and a common set of metadata items are the main instruments. Coherence of metadata is also an important factor for improving interpretability.

40. Coherence of data has four important sub-dimensions: within a dataset, across datasets, across countries and over time. Coherence within datasets is improved in *OECD.Stat* by the fact that data are related to the glossary which forces data concepts to be in-line with their official definition and other attributes, whether internationally agreed or not. Coherence across datasets is improved by the data confrontation permitted by the data warehouse in the improved accessibility and interpretability of data. Accuracy is also potentially improved through more data confrontation possibilities.

41. Finally, quality guidelines encourage the use of the new tools provided by the new infrastructure and, consequently, quality reviews will permit the full implementation of the new infrastructure for all activities.
