Distr.
GENERAL

CES/AC.71/2004/13
9 March 2004

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE (ECE) 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 (MSIS)
(Geneva, 17-19 May 2004)

Topic (ii): Development of IT strategies in statistical offices

BUILDING OECD'S NEW STATISTICAL INFORMATION SYSTEM

Invited Paper

Submitted by the Organisation for Economic Cooperation and Development (OECD)[1]

Summary

1. The vision of a modernized OECD Statistical Information System was set out in the Organisation's Statistics Strategy in 2002. The aims are to:

- improve the efficiency of the Organisation's statistical processes;
- enable reuse of data within the Organisation for multiple analytical and policy purposes;
- improve the quality of the Organisation's statistical data and metadata, notably its timeliness, coherence and availability
- enhance the accessibility of the Organisation's statistical resources to external users.

through innovative use of information technologies.

2. The new Statistical Information System consists of three inter-operating layers:

- a **production layer** for collection, validation, processing and management of statistical data and metadata;
- a **storage layer** where validated statistics and related metadata are stored;
- a **dissemination layer** for producing statistical publications and online/offline electronic statistical products.

3. The three layers are supported by a workflow system that automates statistical and publication processes wherever possible, and tracks the steps involved.

[1] Prepared by by Lee Samuelson and Lars Thygesen (lee.samuelson; lars.thygesen@oecd.org).

4. The new Statistical Information System thus encompasses all aspects of the Organisation's statistical processes. More importantly, it preserves the decentralized nature of OECD directorates' statistical activities, while making their data and metadata part of a coherent corporate system. Users of OECD statistical outputs will be able to enjoy an important improvement of coherence between products, as they will be directly interlinked and present themselves with a common look and feel. Contents will be presented with richer and more consistent metadata.

I. INTRODUCTION

5. The gathering and harmonization of international statistical data in a multidisciplinary environment are key to international comparative analysis and policy work. The availability of timely, accurate statistical information enables OECD committees, officials in Member countries and the OECD Secretariat to address a wide range of issues in today's rapidly evolving global economic and social landscape. The statistical products that OECD makes available to analysts and decision-makers in the general public help all better understand, and respond to factors shaping the global economy.

6. The OECD is developing a modern Statistical Information System to support its statistical activities. The objectives are to improve the efficiency of data and metadata collection, validation, processing, storage and dissemination; eliminate errors, incoherencies and duplication within in and across datasets; shorten statistical publication production cycles and streamline related work flows, and generally enhance the accessibility, timeliness and visibility of the Organisation's statistical outputs.

7. In order to achieve this, the Organisation is taking full advantage of advances in information technologies and standards (e.g., OLAP, XML, Web Services, GESMES/TS).

8. The structure and ideas behind the new Statistical Information System are inspired by best practices in the international community of official statistics. A basic source has been Information Systems Architecture for National and International Statistical Offices – Guidelines and Recommendations[2], issued by UN/ECE 1999.

9. The OECD Statistical Work Programme (OSWP), a database with high-level information on all of the Organisation's statistical activities, is an important complement to the new Statistical Information System. The information in OSWP is structured and activities are linked to themes, data collections, datasets, and publications. The contents of OSWP can be searched in a structured manner with key words, as well as with free text.

10. The overall architecture of the Statistical Information System consists of three layers:

- a **production layer** for collection, validation, processing and management of statistical data and metadata
- a **storage layer** where validated statistics and related metadata are stored
- a **dissemination layer** for producing statistical publications and online/offline electronic statistical products

11. The three layers (or pillars in the architecture model, see Figure 1) are supported by a workflow system which automates statistical and publication processes wherever possible, and tracks the steps involved.

12. The basic design principles underlying the development of the new Statistical Information System include:

[2] www.unece.org/stats/documents/information_systems_architecture/1.e.pdf

- a clear separation of production environments (data and metadata collection, processing, validation) from the data dissemination system;
- a central repository ("warehouse") for statistical data and related metadata as the unique source for publication and online dissemination.

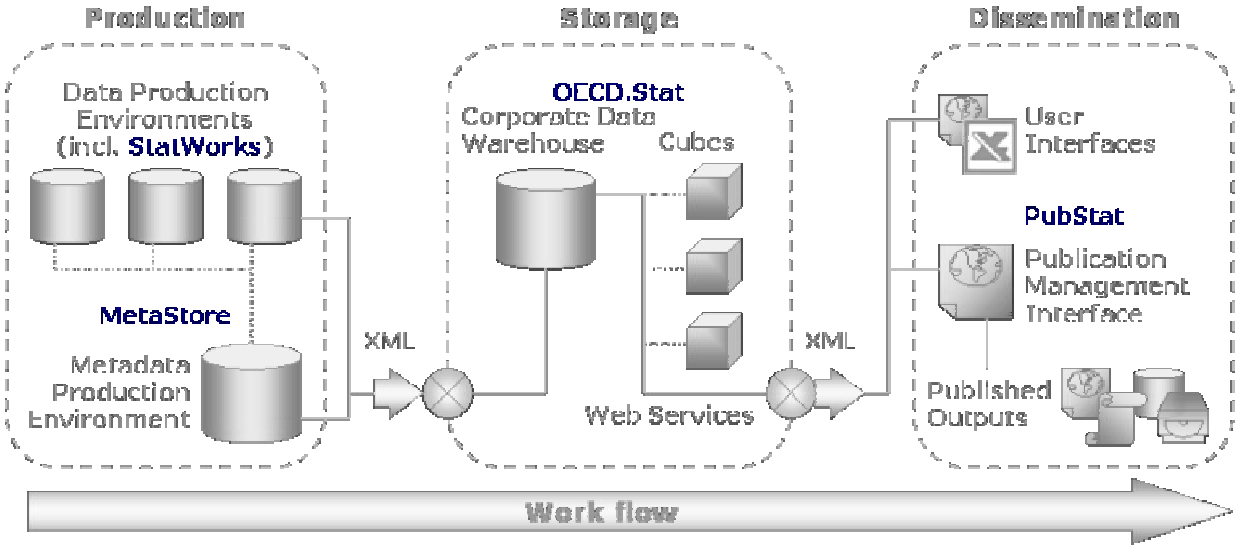

Figure 1. Architecture of the Statistical Information System

13. The technology foundations of the new Statistical Information System include MS-SQL Server and its OLAP component, the .Net development platform, ASP.Net for web applications, and MS Office Web Components. XML-based design features permit "loose coupling" of the System's components. Thus a change in one component and/or its data inputs and outputs can be made without necessitating changes in the others.

14. The new system is being developed through a joint effort of three of OECD's directorates (IT, Publishing and Statistics), with a continuous dialogue with users and producers in substantive directorates.

II. COMPONENTS OF THE NEW STATISTICAL INFORMATION SYSTEM

15. The five independent but inter-operating components of the new Statistical Information System are described below.

II.A OECD.Stat – The central repository for validated statistical data and metadata

16. OECD.Stat is the central repository ("warehouse") where validated statistics and related metadata are stored. OECD.Stat – is at the heart of OECD's new Statistical Information System. It will, in due course, be the sole and coherent source of statistical data and related metadata for the Organisation's statistical publication and electronic dissemination processes.

17. OECD.Stat enables the Organisation's analysts and statisticians to easily locate needed data from a single online source, rather than having to navigate multiple databases and data formats. They can do this with a familiar tool, Excel, rather than multiple query/manipulation systems. And the access to systematic metadata in OECD.Stat helps ensure appropriate selection and use of statistical information.

18. OECD.Stat has been designed to preserve the decentralized nature of OECD directorates' statistical activities, while making their data and metadata part of a coherent corporate system. Each directorate

contributes contents of its production databases to OECD.Stat. Updated statistical data are exported to OECD.Stat from StatWorks and other production databases, upon validation by a dataset manager. Similarly, updated statistical metadata are exported to OECD.Stat from MetaStore (and, during a transition period, other sources), when validated by a dataset manager.

19. OECD.Stat is being progressively populated from the nearly one hundred time-series and cross-section production databases managed by directorates across the House. OECD.Stat notably includes an agreed set of "Reference Series" (series frequently used in the calculation of other indicators), to help ensure that economic and social indicators are computed in a coherent and consistent way across the House.

20. An automated "import" gate provides a common basis for loading statistical data and metadata to OECD.Stat. The XML-based design of the import gate allows "loose coupling" of OECD.Stat and production databases. That is, a modification of the structure of OECD.Stat does not necessitate modification of a production database's export routines, and a change in structure of a production database does not require modification of the OECD.Stat import gate. Statistical data are stored in a number of datasets, in a relational database environment, and in multi-dimensional tables (often referred to "cubes"). A Web service provides a common basis for extracting data from OECD.Stat (i.e., for producing statistical publications, for electronic data dissemination, for interfaces to analytical applications, etc.)

21. Metadata can be attached at any level of an OECD.Stat dataset: dataset-level, dimension-level (e.g., "Variable means…"), dimension member-level (e.g., "GDP means…"), and data-level for any lower level (e.g., series level, observation level, etc.). OECD.Stat metadata values are themselves character strings in an XML format -- a standard yet flexible format for managing textual information, which can be easily reformatted (e.g., to HTML) for display.

22. The principal user interface to OECD.Stat -- through which internal users locate, retrieve and display statistical data and metadata -- is based on an Excel "add-in". This add-in, via a sequence of menus, helps users locate and retrieve data series of interest. The choice of Excel as the basis for the user interface was made largely because Excel is a familiar tool to OECD analysts and statisticians, and an Excel-based solution could be quickly developed and implemented. The Excel interface enables a user to pull data together from multiple sources. An Excel table may be stored, and when later opened, the data contained therein will be automatically refreshed from OECD.Stat.

23. Many users appreciate the advanced data manipulation features of Pivot Tables, which are available through the Excel interface. A web-based interface has also been developed, however, to meet the needs of those who simply wish to be able query and display elements from a dataset, in a manner similar to that when navigating the Internet, and possibly transmit them to their favourite analytical tool for further processing.

24. Access rights to datasets in OECD.Stat are managed by the corresponding dataset managers, who grant permissions to update an OECD.Stat dataset, and to access that dataset. Access rights are typically defined when a dataset is initially exported to OECD.Stat. By default, access to a dataset will be public (i.e. open to all staff at the OECD Secretariat), as most information in OECD.Stat consists of validated statistics and metadata destined for publication. A dataset manager can modify access rights at any time.

25. A number of additional features are planned for subsequent releases of OECD.Stat. These include a search facility, to make it easier to locate series of interest; "joint display", to make it possible to seamlessly view series coming from different datasets; "personal workspace", enabling a user to, *inter alia*, subscribe to e-mail alerts whenever data series of interest are updated; pre-defined queries and graphics (i.e., designed by dataset managers) that can be stored, and invoked from a list (i.e., to reproduce tables in a statistical publication, "OECD in Figures", "Education at a Glance", etc.); user-defined queries that can be stored and later invoked from a list box; and "versioning", so that data managers can provide a "snap-shot" of their dataset at a specific point in time (i.e., at the time data are frozen for a publication), thereby enabling users to either see the dataset with its most up-to-date data, or as it stood at the time of the publication.

II.B StatWorks -- Modernizing the production database environment

26. OECD statistical database applications have been developed using a number of different software platforms (Oracle Express, SQL Server, Excel, Paradox, Access, FAME, etc.) This is in part a reflection of the decentralized nature of the Organisation's statistical activities. StatWorks has been developed to provide a common hosting environment for production databases, based on the OECD standard database hosting platform, MS-SQL Server. The objectives are to minimize the number of tools used in statistical activities, and correspondingly reduce training and support requirements. OECD directorates are being encouraged to migrate production databases hosted in non-standard environments to the OECD standard (MS SQL Server), and to StatWorks in particular. However, there is no intention of migrating all the present production environments to StatWorks, as some of the production systems provide very sophisticated support for the specific statistical area, which could only be migrated through a major development effort.

27. StatWorks has been designed to host statistical databases irrespective of the number of their dimensions (country, time, subject, etc.). Database dimensions can be defined as "private" and specific to a database, or as "public" and shared among several databases hosted in StatWorks (e.g., country, time). A system of referential integrity common to all databases protects against inadvertent corruption. Access controls ensure that each database manager "sees" only his/her database in StatWorks.

28. StatWorks includes a toolkit for managing statistical data, which minimizes the need for developing and supporting database-specific programs. This toolkit includes facilities for:

- initial data migration;
- database administration (managing user access rights, defining dimensions, adding a country, etc.),
- security management;
- data collection via interactive questionnaires;
- data import / conversion routines;
- data validation (e.g., checking: for missing data, breaks in series, re-basing of series, significant departures from previous values, etc.);
- data manipulation (e.g., calculation of regional totals);
- query via Excel pivot tables and web interface;
- dynamic links to MetaStore;
- export of datasets to OECD.Stat.

II.C Coherent management of metadata – MetaStore

29. OECD's corporate metadata facility, MetaStore, is designed to improve the efficiency of metadata preparation, storage, access, management and dissemination for the Organisation's statistical products. MetaStore provides dataset managers with a common interface and common set of tools for managing metadata, and supports adherence to common standards for statistical metadata across the House. MetaStore addresses problems of fragmented metadata located in numerous databases and text files maintained by different Directorates, duplication of effort in metadata preparation, gaps in metadata availability (particularly metadata explaining differences between similar or related series residing in different databases), and inconsistent metadata across databases.

30. MetaStore has been designed to support a set of principles which apply to metadata validated statistical data to be shared or disseminated, whether internally or externally.

Box 1. Metadata principles

The following principles apply to metadata validated statistical data being shared or disseminated, internally or externally.

1. All statistical data being shared or disseminated, internally or externally, must have appropriate metadata, preferably including all of the following:

- *name(s)*
- *discovery metadata,* allowing users to search for statistics corresponding to their needs
- *conceptual metadata,* describing the concepts used and their practical implementation, allowing users to understand what the statistics are measuring and, thus, their fitness for use
- *methodological metadata,* describing methods used for the generation of the data (e.g. sampling, collection methods, editing processes)
- *quality metadata,* describing the different quality dimensions of the resulting statistics (e.g. timeliness, accuracy)
- *technical metadata,* making it possible to find and retrieve the data

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

- the same name, definition, and other descriptions should be connected to the same statistics, no matter where it is and who is the "owner"
- the same name must not be used for statistics that are not identical
- terms and concepts must be consistent throughout

3. All metadata must be created only once, for efficiency reasons and also in order to avoid confusion, incoherency and mistakes.

4. For each dataset there must be one responsible (unit) who is also responsible for the metadata. In cases where more than one unit includes the same statistics in the data they disseminate, it must be decided who is responsible; by default, the responsible is the one who originally collected it.

31. MetaStore provides a database manager with:

- a storage area for statistical metadata;
- an interface for managing production metadata;
- facilities for enriching current metadata content through external links to standard classifications, glossary terms, SDDS concepts, etc.;
- export of metadata to OECD.Stat.

32. MetaStore can accommodate any kind of metadata related to any level of detail of the corresponding statistical data, from subject matter area (data set), country, time series, down to the single observation; or an arbitrary set of these, such as data series observations in a specified time interval. For example, metadata may pertain to a database in general (i.e., purpose of the database, database manager, most recent update, next update, dimension members, methodology used in compiling statistics, seasonal adjustment methods, publications derived from the database, etc.), or be specific to the data themselves (i.e., breaks in series, missing observations, estimates, revised figures, etc.).

II.D PubStat – modernizing statistical dissemination processes

33. Technology and publication standards for the OECD statistical products have been defined, enabling the development of modernized tools and processes for producing traditional statistical publications and interactive data products. This is made feasible by the existence of a single source for validated statistical data and metadata, OECD.Stat. Thus PubStat is being developed with the objectives of:

- increasing the efficiency of statistical dissemination processes;
- reducing the risk of human intervention and, thus, of mistakes;
- reducing time-to-publish;

- giving the Organisation's statistical publications and electronic products a common "look and feel";
- reducing the number of different software tools and corresponding support effort involved, and;
- minimiZing the time that statisticians spend dealing with dissemination and formatting issues.

34. PubStat contains, for a given publication, information on the structure of the tables, the statistical data and metadata to be extracted from OECD.Stat, where these data are to appear in a statistical table, the "publication metadata" containing headings and labels, and the templates implementing the graphical presentation of the tables.

35. PubStat provides the database manager with:

- an interface for managing publication layout instructions;
- a facility for storing these instructions for later re-use;
- an interface for generating an XML output file combining statistical data and metadata with publication layout instructions.

36. PubStat, in conjunction with OECD.Stat, will notably make it possible to streamline the production of the Organisation's existing "horizontal" statistical publications, where statistics are drawn from several separately-managed datasets (e.g., "OECD in Figures"), and make it feasible to produce new ones (e.g., "OECD Statistical Yearbook").

Traditional publications

37. When a statistical publication is to be produced, PubStat invokes an OECD.Stat Web to generate an XML file combining relevant statistical data and metadata with information on the publication table structures. This XML file is exported for processing by a batch composition engine. This composition engine generates a PDF file, based on formatting information contained in a template. The PDF file is then sent to the database manager for validation. When the database manager is satisfied (perhaps after one or more iterations), the validated PDF file is sent for printing on the Organisation's computer-to-plate equipment, and also made available through OECD's online services.

Electronic statistical products

38. Greater visibility and availability of OECD statistics are key elements of the Organisation's Statistics Strategy. The OECD presently provides external users with access to its statistical resources via OLIS, the OECD Web site (Statistics Portal), SourceOECD and CD-ROMs. However, it is apparent that the present user interfaces are not sufficiently easy to use for outside users or users who are not already familiar with the structures of the parent production databases. The OECD is seeking to remedy these problems.

39. Opportunities are therefore being reviewed for developing a state-of-the-art presentation tool for the electronic dissemination of OECD's statistical outputs. The objectives are to provide a more user-friendly and functionally rich statistics browser for locating and retrieving OECD statistical data and metadata, enhance the potential for generating cost recovery from access to statistical databases, and minimize resources required to create files for electronic dissemination.

II.E Workflow – automating statistical and publications processes

40. A workflow system is being developed to automate statistical and publication processes wherever possible, and track the steps involved. This workflow system will be used to initiate and monitor the transmission of data from one layer to another – notably from the Production Layer (StatWorks, other production systems, MetaStore) to the Storage Layer (OECD.Stat), from the Storage Layer to the Dissemination Layer (PubStat), and from PubStat to the various channels used for dissemination of the Organisation's statistical products: PDF files for traditional publications, specialized files for online

dissemination (Internet, OLIS, SourceOECD), and specialized files for dissemination via CD-ROMs. Dataset managers can elect to receive notification by e-mail of the results of each process.

41. Development of the workflow system and its innovative features will be reported more fully in a future paper.

III. CO-OPERATIVE DEVELOPMENT WITH OTHER INTERNATIONAL ORGANIZATIONS

42. The OECD seeks, wherever feasible, to exchange experience and "best practices" with other international organizations, national statistical office and central banks, in order to make the best possible use of information technologies to facilitate the exchange of statistical data and metadata. Indeed, a number of software tools and standards for management and exchange of statistical information are being developed cooperatively with other international organizations. The OECD, for example, putting great emphasis on the SDMX[3] initiative sponsored jointly with six other international organizations, aiming at finding common solutions for exchange and sharing of statistical data and metadata[4].

43. Furthermore, OECD has in the *NAWWE*[5] project piloted a new method of collecting and sharing national accounts data from the web sites of member countries, using a common XML format aligned with the SDMX initiative. This effort is intended to lead to coordinated data collection with other international organizations.

44. Following agreement to share responsibilities for collection of annual foreign trade data, the OECD and UN have additionally agreed to work jointly to establish a common system for managing annual foreign trade data, using an SQL-based data model designed by the UN Statistics Division (the *ComTrade* project). Development effort for the data collection, validation, processing and management software is being shared by the OECD and the UN. Responsibilities for collecting and validating data will also be shared, with data periodically replicated from one site to the other.

45. The Organisation has also taken a number of other initiatives for cooperative development of software tools and standards for management and exchange of statistical information, notably: a Technical Collaboration Project with the IMF on Web Services, and development of StatWorks with the UNESCO Statistical Institute.

[3] Statistical Data and Metadata eXchange, see www.sdmx.org
[4] Described in another paper for this meeting: *Practical Experience Towards Implementing SDMX at OECD*
[5] National Accounts World Wide Exchange

IV. CONCLUDING REMARKS

46. While the construction of the new information system has already advanced very well and many elements, as well as their interaction and the overall structure, are in place, the full potential will only be realised gradually over the coming years. The following road map indicates the present position and the milestones expected in the short to medium term:

Present situation (March 2004)	OECD.Stat StatWorks MetaStore PubStat	In production, 5 major databases loaded v./ 1.0 v./ 1.0 One pilot publication (Government Debt)
July 2004	PubStat Work flow	v./1.0 launched v./1.0 launched
End 2004	OECD.Stat PubStat	Populating: 17 databases loaded 4 publications produced
2005	OECD.Stat	Another +20 databases

47. The problems which the OECD seeks to solve by building its new Statistical Information System seem to be general and exist in many statistical environments world wide. The system is modular and the interfaces are built on well-established standards, making the exchange of components relatively easy. Therefore, it is expected that the components would be reusable in national and international organizations (especially for managing decentralized statistical systems). Likewise, the OECD is looking for best practices among peers in order to integrate such practices and as far as possible avoid duplication of work.

48. With this aim, OECD has recently set up an international expert group –- the OECD Expert Group on Statistical and Metadata Exchange -- to foster dialogue with national and international partners on strategic issues related to the development and the practical implementation of new procedures for statistical data exchange.

- - - - -