

Distr.
GENERAL

CES/SEM.47/25
21 March 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 IV: Ways of making statistical information systems more responsive to users

**IMPROVING ACCESS TO STATISTICAL INFORMATION AT OECD
IN RESPONSE TO USERS' REQUIREMENTS**

Invited paper

Submitted by the Organization for Economic Cooperation and Development¹

Abstract:

The rapid and easy access to a comprehensive set of statistical information is a major challenge for many organizations, particularly if the management of certain data collections is decentralized in specialized subject matter areas. This paper focuses on work carried out by the OECD Secretariat to render its statistical data and metadata accessible to different constituencies. It briefly describes the context in which OECD's analytical/statistical computing and communication requirements should be viewed and then outlines the background and vision for the development of a corporate statistical information system. It also provides details on user requirements and implementation plans. A key success factor for the horizontal integration of different sets of statistical data and metadata is the availability of a central catalogue, which contains comprehensive information about the various data sources. Modern information and communication technologies (ICT) play an essential role to offer a corporate view into statistical information held in a decentralized environment. The concepts outlined in this paper have been developed in close collaboration with the Statistics Directorate and through a specialized internal task force² which includes major users of statistical information at the OECD.

Keywords: statistical information management, statistical data-/metadata repositories, data-warehouse, information dissemination, statistical portal services, statistical reference data.

I. INTRODUCTION

1. The OECD is recognized throughout the world for its reliable and comprehensive analytical and statistical work. The gathering and harmonization of international data in a multidisciplinary environment are key to international comparison and policy making. Statistical, analytical and policy work is essential

¹ Prepared by Peter Lübker, Head of Systems Development and Support (peter.lubkert@oecd.org).

² The OECD Analytical Statistical Task Force (ASTF) - see also acknowledgements

to meet the current as well as medium and long-term business challenges of the OECD. It enables OECD committees to carry out a wide range of assignments in today's global economic and social landscape. International trade, economic forecasting and sustainable development are only a few examples of issues which heavily rely on this area of activity, and which illustrate the diversity of themes that need to be addressed.

2. Statistical computing activities at the OECD encompass the collection of national statistics, their validation and harmonization, the subsequent derivation of internationally-comparable indicators, economic policy analysis, data modelling, and statistical publishing in both paper and electronic forms. To carry out this work, OECD economists, analysts, and statisticians require a comprehensive set of tools ranging from the standard spreadsheet package over more specialized data manipulation and management systems to a variety of econometric and modelling software packages.

3. Traditionally, decisions on the organization of statistical processes have been taken putting priority on the requirement of individual statistical units. This has resulted in making individual statistical activities very efficient at the expense of the co-operation between analysts from different subject matter areas for cross subject studies. Presenting statistical information in a corporate way while being managed in a decentralized structure is a challenge that modern ICT³ can help to address. The organizational structure of statistical activities is described in the paper *Metadata requirements for the Integration of a Decentralised Statistical Systems*⁴

4. A vision for an OECD wide statistical environment has been developed based on a survey of user needs. The main element of this vision consists in providing a unified view of OECD statistics through a common data browser, involving a repository of reference series and common search and query tools.

II. REQUIREMENTS

5. The OECD is a knowledge-based organization with a multi-faceted working environment and a wide spectrum of information management and dissemination needs. Many of its activities rely heavily on the availability of specialized ICT tools, which need to communicate well among each other.

6. Collections of statistical information are used internally for analytical work, econometric modelling and forecasting, and subsequently in publications. In addition, the online access to published OECD statistical information has become a priority in the past few years. Offering easy online access to OECD statistics is one of the Organization's strategic management and communications objectives, as well as a medium-term information technology direction.

7. The OECD computing environment has to integrate a variety of software tools, and facilitate the sharing of information among applications residing on PCs, workstations, and networked servers. In this context of particular importance is the availability of a user-friendly data management software environment, unifying the Organization's various databases and meeting the diverse requirements of both database managers and end-users. The increasing importance and availability of metadata compound this requirement. New methods for the collection of statistical data must be implemented to accommodate evolving international reporting standards and take advantage of new transmission media.

8. Equally important is the availability of analytical software for econometrics, statistical analysis, macro-economic modelling, and graphics which, in turn, integrate with standardized procedures for the preparation of compound documents (integrating text, tables and graphics) for internal and committee use, and for external distribution.

³ Information and Communications Technology

⁴ Paper by Gérard Salou presented at the ISIS 2002 seminar.

9. Last, but not least it is essential to constantly increase the speed, quality and flexibility of the Organisation's publication production and printing processes and to enhance facilities and procedures used for electronic data capture and data dissemination.

10. The rapid evolution of ICT technologies make it feasible to envisage innovative, cost-effective solutions to these computing needs, while at the same time providing better access to user-friendly, commercial software tools and database management and manipulation facilities. Also, the Analytical/Statistical systems architecture must evolve in harmony with overall ICT developments and leverage existing investments.

11. Opportunities exist for increasing the efficiency of information production and management, and for improving the quality and effectiveness of presentation to targeted audiences, both internal and external. The requirements for providing access to OECD statistical information are as follows:

From the users' point of view

- More integrated analytical/statistical working environment;
- Easier to use data manipulation software and easier access to data;
- Simpler data collection and improved metadata handling;
- Availability of modern analytical tools/packages;
- Easier, more automated, higher-quality graphics production;
- Extended interactive graphical data analysis facilities;
- Shorter publishing cycles;
- More attractive work (IT) environment for outside professionals (new recruits, consultants);
- Flexibility to meet future end user needs, shifts in technology;
- Less risk of duplication and improved data quality through better metadata.

From ICT service point of view

- Opportunities to further standardize and streamline the overall ICT infrastructure
- Adoption of emerging industry standards and international norms
- Increased application scalability
- Portability and inter-operability of analytical and statistical information systems across a range of different operating systems and manufacturers' equipment and a range of equipment of different computing capacity.

Data

12. There exist two main types of data objects to be stored and managed at the OECD in support of analytical/statistical applications. The most common type is time series, i.e. a vector of numeric observations ordered by time, associated with various textual attributes such as the time series name and description (e.g. agricultural production for Switzerland, total at current prices, in million Swiss francs), the frequency of observations (e.g. monthly), starting period, etc. Today time series objects are generally managed centrally in a common SQL-based systems environment and with FAME⁵, a time-series manipulation and graphics tool.

13. At the OECD, there also exist many sets of data values, which are homogeneous enough to be structured as multidimensional data arrays. Typically, one dimension of such arrays is time and another one country (i.e. the list of OECD Member and partner countries), while other variable dimensions are more specific to the application. For example, for foreign trade data by commodities, other dimensions

⁵ Forecasting, Analysis, and Modelling Environment

include the list of trading partner countries, a normalized list of products (SITC, HS), value or quantity, and the type of flow (imports, exports). A cell of the array, i.e. an observation, then corresponds to a unique combination of one element in each dimension. One obvious advantage of multidimensional arrays over individual time series is that data indexing comes as a natural by-product of the multidimensional structure definitions. Another advantage is the ability to process at once logical groups of many time-series and, thus, simplify group operations such as data retrieval or aggregation. However, when time series data sets are volatile or heterogeneous, it may prove more efficient to treat time series individually. Commercial software offerings using the concept of multidimensional data objects became available in the 1990's. At that time the OECD chose IRI Express (today ORACLE Express) as their standard product for the manipulation of this kind of data.

Metadata

14. Information about data objects, i.e. metadata, also needs to be maintained along with data objects. Reference item lists, such as nomenclatures and/or dimensions of multidimensional data arrays (e.g. list of OECD Member countries) are obvious examples of metadata objects. These lists often need to be hierarchical (e.g., countries within geographic zones) and sharable amongst database users, so as to facilitate structure harmonization across databases. Metadata also include descriptive information, documentation and notes at different levels (e.g. at the application database, time series or even individual observation level) to describe important characteristics of data object contents. In the context of application development, metadata are a challenge and tool at the same time, meaning that metadata always has to "travel" with (be attached to) the data, but also has to help locating the data.

Catalogues

15. The development of a central catalogue of all officially available application databases and their respective contents is very important. It should include data sources, committees, user directorates, etc. - and about application databases - including associated data collection methods, database system implementation, database usage, access control rights, database output products (publications, electronic data products), etc.

16. Obviously, not all application data objects used need to appear in the catalogue. A distinction thus needs to be made between catalogued application databases (i.e. clean, official, shared groups of data objects), on one hand and non-catalogued databases (i.e. unofficial working data-sets) on the other. For instance, incoming data from Member countries will, in general, transit first through non-catalogued databases, to be validated and checked for consistency, before becoming catalogued and thus part of the central data environment.

User Communities

17. When deciding on how best to provide access to a cross-section of statistical data and metadata in support of multidisciplinary projects, it is particularly important to differentiate user communities. Two major categories can be distinguished - internal and external users. In a recent study on improving access to OECD statistics for internal users at the OECD were categorized by:

- Statistical Assistant
- Database manager
- Economist
- Visiting Expert
- Management
- IT Specialist

18. The study highlighted the importance of the following needs for internal users of OECD statistics:
- Locating OECD 'Reference' series (most popular series - i.e. Exchange rates, CPI, GDP, Unemployment)
 - Locating data by theme from different sources
 - Global keyword search across directorate for specific series or across an entire catalogue
 - Making your data more visible/accessible to others
 - Clarifying/highlighting areas of duplication
 - Monitoring data access
19. Some of the key needs identified for internal target audiences extend clearly also to external users (notably the first three in the list above) and have to become part of related data dissemination strategies.

III. PUTTING IT TOGETHER

20. The successful implementation of a strategy to develop and keep up-to-date an ICT infrastructure in support of analytical /statistical processing is a complex and multi-faceted challenge. Obviously information and user requirements come first, but elements like ICT strategy and a management structure are also important.

Governance

21. A proper management framework with executive-level support is critical for any large ICT investment. Objectives, costs and expected benefits have to be spelled out from the beginning. A clear vision statement underpinning the business relevance is essential. Further technical directions have to be adopted and coordinated. The OECD Statistical Policy Group (SPG) is a senior management group chaired by the Chief Statistician⁶, which takes the lead in the co-ordination of organization-wide statistical strategies. A Board of Directors for Computer and Communications Strategies provides global guidance for ICT investments and agrees on overall strategic directions. An inter-directorate Task Force on new technologies for Analytical/ Statistical systems (ASTF) carries out the more detailed analysis of technical strategies and implementation plans. The ASTF coordinates its efforts very closely with the SPG and regularly reports to the Board of Directors. Last, but not least efforts are undertaken to get input and feedback from the final clients/users through thematic seminars and targeted surveys.

The Role of ICT

22. As developed in the previous sections, ICT plays a role in many domains related to analytical/statistical information processing. The following list provides an overview of different development domains.

Development Domains

- Data Collection and Validation
- Corporate Data Environment
- Metadata Management
- Short-Term Statistics
- Sectorial Statistics
- Multidimensional Data Analysis
- Time-series Management

⁶ Enrico Giovannini

- Software for Economic Analysis
- Graphics Production
- Desktop Tools and Front-Ends
- Statistical Documents & Publications (Database Publishing)
- Online OECD Statistics
- Electronic Data Dissemination / Data Products for Sales
- Application-Specific Requirements
- A/S Computing Infrastructure (Hardware and Operating Systems)

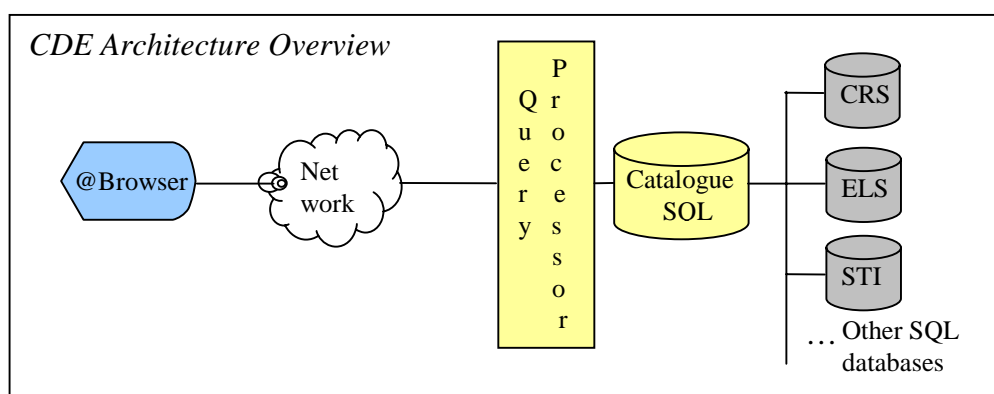
23. An essential responsibility of the OECD Information Technology and Network Service, which cuts across the complete range of business areas, is to elaborate a coherent overall architecture throughout the above development domains in line with the ICT strategy.

24. Major bricks in the software infrastructure used across OECD analytical/statistical areas today are detailed in the above-mentioned paper by Gérard Salou⁴. It is important to mention that the bulk of processing requirements relates to “programmed” data manipulation. Users have a need to automate recurrent activities, such as the reception of statistics from Member countries and the preparation of statistical publications. This requirement - bulk data processing - is reflected by the availability of ORACLE Express and FAME, both of which feature a sophisticated, specialized fourth generation programming language.

Initiatives to Date to Develop Statistical Services Based in Internet Technology

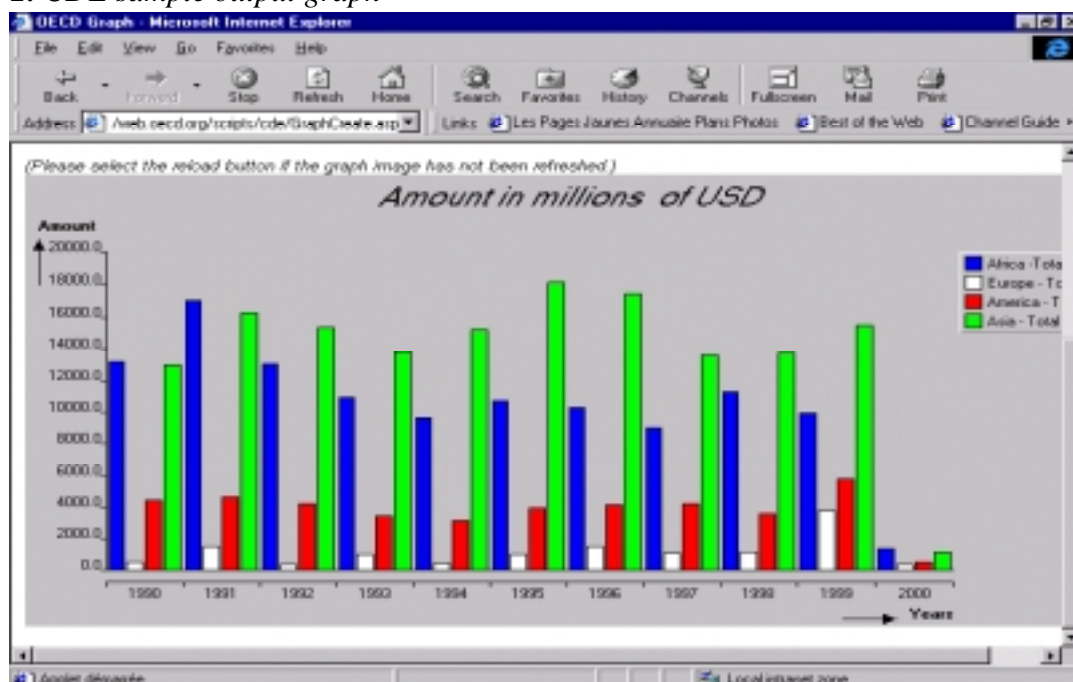
25. In light of the above needs and strategic directions, the OECD has been working on innovative ways to offer new services for the management and dissemination of statistical data and metadata. First steps were taken in 1996 to offer access to specific statistical data sets, in particular the OECD Creditor Reporting System (CRS), which is characterized by a relatively high update frequency. A system based on standard relational database technology was devised to offer Intra- and Internet access to dynamically updated data from standard Web browsers. From a technical point of view, this work then evolved into the creation of a “Corporate Data Environment” (CDE), which allows accessing statistical indicators held in different locations from a single entry point using a data catalogue. The visualization and export of selections, however, remains limited to a single source at a time. Figure 1 illustrates the architecture of this data-warehouse. All databases included in this initiative are SQL Server databases.

Figure 1.



26. Access to metadata is ensured at both the catalogue and the individual database or data set level. Through its query processor, the CDE features mechanisms to easily find selected statistical data and metadata, carry out comparisons, and compile outputs in various formats. Figure 2 shows a sample output graph from the CDE.

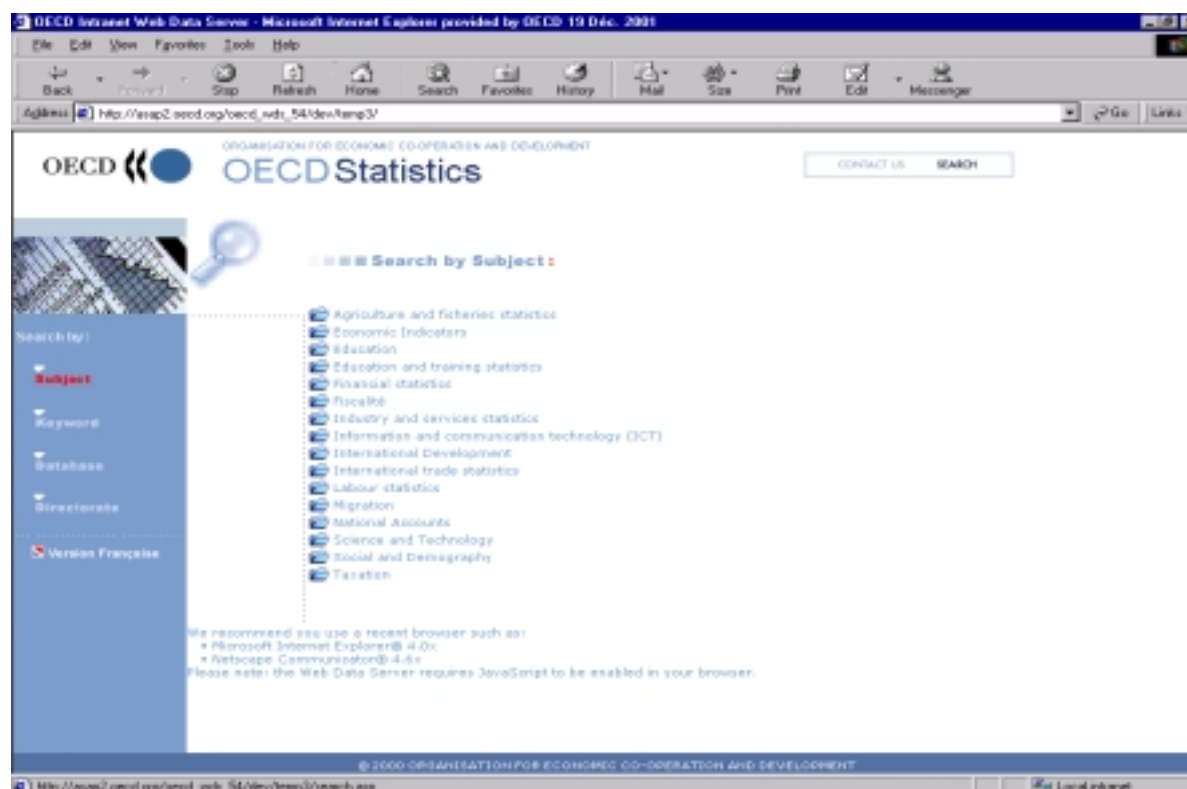
Figure 2. CDE sample output graph



27. More recently, efforts have concentrated on improving the on-line access to a broader range of official OECD statistics independently of their source database environment. The interactive access to these sometimes rather large data sets has complemented traditional publishing channels/media - see Figure 3. An important consideration in this context was to ensure that the respective data sets are compiled only once for use on different media. This not only allows a reduction of the amount of work and time to market, but also diminishes the risk of inconsistency. Altogether OECD Statistics online cover 14 themes (or subject areas) from Agriculture to Taxation.

28. Following the standardization of OECD Electronic Data Products (EDPs) on a single dissemination software – Beyond 20/20 – the OECD started to work with the editor of this software⁷ on the functional design for an access mechanism to Beyond 20/20 data files using Internet technology. As with the CDE, all data is stored in a common format. In the first phase of this collaboration, efforts concentrated on the definition of functional and ergonomic aspects. A further standardisation of nomenclatures used in the different data sources was required before releasing a consolidated set of official OECD data on the Beyond 20/20 Web Data Server (WDS). In the context of this work, the multidisciplinary character and the decentralised structure of the OECD very well illustrates the importance for adequate systems support of metadata requirements.

⁷ Beyond 20/20 Inc. - © Beyond20/20, Web Data Server (WDS)

Figure 3. OECD statistics On-line today

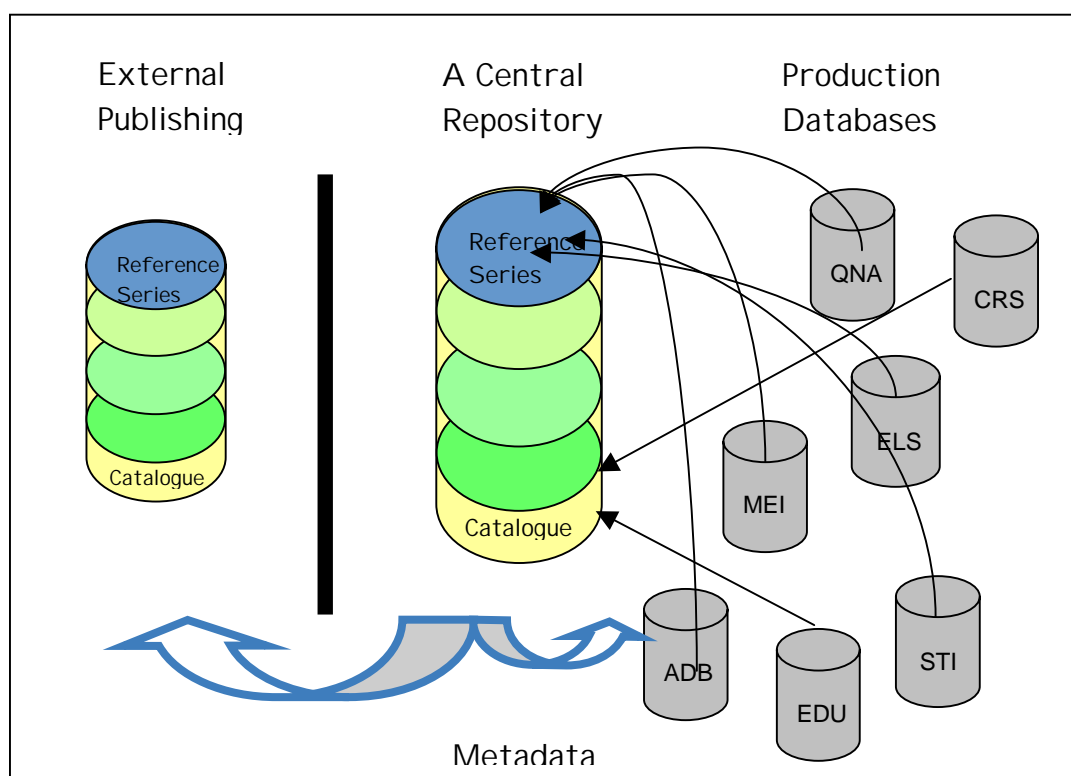
The Common Browser Concept

29. Based on the results of the requirements study outlined in section 3, and in line with the OECD Chief Statistician's "New Strategy for the OECD Statistical System", technical solutions for implementing a Common Browser are being studied. Looking at the overall functional concept illustrated in Figure 4 it becomes clear that there are many commonalities with the CDE architecture. An important difference from the technical viewpoint is that relational database management system vendors have begun to integrate OLAP (online analytical processing) functionality into their engines. This OLAP functionality is particularly useful for the management of multi-dimensional data objects. This tendency is complemented by the steep increase in speed and of power of modern processors as well IO systems and should allow a tighter integration of different data sources into a common repository. The possibility to build a "broker function" between different types of database-engines and data manipulation tools (SQL Server, ORACLE Express, FAME, etc.) providing "real-time" access to various data sources (production databases) also has been considered. It, however, appears to be more suitable to follow a replication approach as was done for the CDE for a number of reasons, and in particular the following:

- A clear differentiation between working data sets containing temporary or preliminary raw data and checked/validated data - better security;
- Established and traceable data update responsibilities ;
- Less direct exposure to change (application & technology) in OECD production environments - simpler maintainability;
- An opportunity for harmonisation of nomenclatures and taxonomy;
- Better scalability and robustness for future evolution.

30. Nevertheless, two important considerations/conditions have to be taken into account:
- (i) The system must be able to provide the possibility to instantaneously reflect updates made in the individual production (source) databases - users will view the common database as a virtual database.
 - (ii) Database owners require simple tools (or automated mechanisms) to update data and metadata in the central repository or data warehouse.

Figure 4. Common Browser Functional Concept



Design and Implementation Considerations

31. The following design and implementation considerations for a Common Browser have been identified:
- Access to all Directorate statistical databases to be from a single browser interface available on the Intranet;
 - Ability to search across databases to locate data by theme;
 - Database-level metadata (objectives, person responsible, scope, sources, last update, etc.);
 - Development of a central data catalogue and dictionary;
 - Ability to select and combine time series from different data databases (“Shopping Basket” feature);
 - Ease of access to frequently-accessed time series, such as GDP, CPI, Population statistics; these series are referred to as “OECD Reference Series”;
 - Means to save selected data in a format suitable for export to analysis software of the user’s choice (Excel, E-Views, FAME,..);
 - Ability to apply FAME’s (or other vendors’) time series analysis and charting functionality to any

time series available from the browser;

- Adoption of a scalable approach: databases could be added on a database by database basis;
- Leverage existing design concepts and browser developments (CDE, STD Browsers, OecdFAME Wizard, WDS...);

Expected Benefits

- **Accessibility:** Access to all OECD databases using a common interface: “Common Browser” feature.
- **Simplicity:** Minimize the need to install and learn several applications.
- **Efficiency:** Rapid location of data for a single theme from many sources.
- **Visibility:** Increased visibility for the most frequently used statistics across the OECD: Reference series feature.
- **Quality:** Opportunity and incentive to improve the quality of OECD statistical data, especially metadata
- **Functionality:** Analysis and charting functionality
- **Reusability:** Save selected series in a format suitable for use by other software (FAME, Excel, E-views etc.).
- **Integration:** Integrate data from different locations into the same presentation: “Virtual database” feature.
- **Manageability:** Creation of a uniform platform enabling the monitoring of data usage.
- **Integrity:** Identification of areas of data duplication.

32. In summary, the “Common Browser initiative” aims to provide end-users with means to search, view, retrieve, and re-use data from a range of relational and specialized database sources via a central data repository based on input parameters and on Catalogue information. Common tools will allow owners of individual data collections to facilitate the updates of data and metadata to the data warehouse. The client will use interactive panels (wizards), based on a data catalogue and a dictionary/glossary and the interface will also offer a set of pre-defined queries for frequently accessed information (“reference series”), as well as the possibility to store queries and to easily export results to other packages. Very importantly, and in addition to the interactive mode of access, programmable interfaces (through APIs⁸) with data manipulation and programming languages (e.g., FAME, Express) are essential. The seamless updating of the common warehouse will involve ETT/ETL⁹ tools and some specific developments.

IV. CONCLUSION

33. The easy and efficient access to key statistical data and metadata in support of international policy analysis and decision making is of strategic importance to the OECD. The innovative use of ICT is key to meeting this challenge. The development strategies described in this paper are intended to assist in the development of statistical information management and services architecture. The objective is to federate “vertical” sets of data and metadata into comprehensive repository through a crosscutting catalogue, a glossary of statistical terms, and an overarching database of OECD statistical activities.

34. Other benefits cover savings in staff time (training, data management), the identification of areas of data duplication, and improving the quality of OECD statistical data, especially metadata. Finally, the implementation of common statistical data warehouse may offer new opportunities in the area of statistical publishing beyond providing timely online access to data and metadata.

⁸ Application Programming Interface.

⁹ Extraction Transformation Transport/Extraction Transformation Language.

Acknowledgements

The preparation of this paper would not have been possible without the valuable input of many colleagues most of which are also longstanding members of the Analytical/Statistical Task Force. Particular thanks go to: Ayse Bertrand, Susan Cartwright, Michel David, Jens Dossé, Eric Espinasse, Trevor Fletcher, Michael Franklin, Jean-Louis Grolleau, Andreas Lindner, Pascal Marianna, Casper Meyer, Douglas Paterson, Gérard Salou, Cengiz Tarhan, and Colin Webb.