

International Conference on Statistics
5 – 7 September 2007
St Kitts and Nevis

The Statistical Data Warehouse - the Key to the Gold Mine of Statistics¹

Lars Thygesen
Head of Statistics Information Management and Support Division
OECD, Statistics Directorate
Tel +33145248402
Mobile +33616050468
lars.thygesen@oecd.org

1. A national statistical system – for whom ?

A national statistical system exists in every country of the world because all good decisions must be based on facts. Historically speaking, the statistical system has originally been seen simply as a government service provided to the government itself, being responsible for policy decisions, legislation and follow-up on results of programmes. However, increasingly other players are seen as legitimate users of and partners in the system: The business community, interest groups, citizens as voters and taxpayers who have a right to understand what government is doing for them, or who need statistics as a basis for their own decisions, and finally the international community.

Therefore, the demands to the statistical system are quite diverse. Many parts of it should be seen as “general purpose”, able to respond to unforeseen data needs in future.

2. Centralised vs. decentralised statistics production

An important feature of a statistical system, national or international, is its degree of centralised or decentralised organisation. Does each ministry, each sector or local authority perform or control the statistics production within its domain, or is there a central authority, a national statistical organisation (NSO), that controls or performs the statistical activities.

In most countries and international organisations the situation is somewhere in between the extremes. The central statistical organisation is responsible for a larger or smaller proportion of the statistics, while decentralised authorities are responsible for each their part. The competencies of the central body in regards to coordination and standard-setting may vary.

Both organisation principles have some advantages and disadvantages. In the centralised model statistics producers may be lacking sufficient knowledge and sensitivity to needs of users and decision makers. Some of the disadvantages of the decentralised model are related to the degree of independence that can be expected from statistical services integrated with policy making organisation². But others relate to the coherence and integration of the databases that are generated by diverse statistics producers. This problem relates partly to efficiency of the production; several actors may need overlapping information and they tend to collect and process it independently, thus duplicating or multiplying work. But more seriously, dispersion of statistical activities may threaten several dimensions of statistical quality, mainly coherence, usability and findability.

¹ Paper prepared for the conference *Statistics and Policymaking in Small Economies: Developing Effective Statistical Systems*. Basseterre, St. Kitts and Nevis, 5-7 September 2007.

² See UN Fundamental Principles of Official Statistics, <http://unstats.un.org/unsd/dnss/fundprinciples.aspx>

Even when statistics are mainly concentrated in one organisation, experience shows that the quality problems mentioned here may be difficult avoid. Within a NSO (or an international organisation), production tends to be organised in so-called stove-pipes³ or independent production lines, each controlling their own process from collection through data editing and estimation to dissemination.

Such an organisation hampers the use of data across several domains. The stove-pipes often have their individual and different dissemination systems or channels, using different methods of retrieval and searching data, they tend to use different delimitations of populations, their own variants of concepts and classifications, making it difficult or impossible to sensibly make joint use of the data, e.g. calculating ratios between different statistics. Each database is only conceived as aimed at one specific use. Ad hoc use for unforeseen purposes is not supported.

One solution may be to develop a statistical data warehouse, where all data, stemming from quite diverse systems or stove-pipes are integrated into one coherent system. This can unveil the information gold mine that has been hidden in the many databases. However, it is important to see this not only as a technical exercise but rather as an organisation project, implying a – perhaps gradual – adoption of common standards and rules for contents.

3. *The OECD Statistical Information System*

3.1 Decentralised statistics production

The mission of the OECD is to compare and evaluate national policies of member countries, sometimes also of non-member countries, in different areas, such as economic policy, education, health, etc. The main idea is to compare the policies themselves and try to identify differences in outcomes – difficult though the causal effects may be to isolate.

In this endeavour, OECD is promoting informed decision making – or rather informed analysis of benefits and deficiencies of different policies. The comparisons have to be based on facts. Consequently, the process generates large amounts of statistics, mainly collected from statistical and other authorities in member countries. Work is organised in a number of domain-oriented committees with representatives of the governments of member countries, each committee being responsible for analysis and data collection in its area. So data has traditionally been controlled and managed in a decentralised fashion, the main point being that the data collected fit the purpose of the analysis carried out by the individual committee.

While this organisation of data management has had the advantage of nearness between data needs and data control, it does not sufficiently take into account the importance of coherence across domains. Can education policies be reasonably evaluated without also considering effects on labour market? And if not, should the same or similar data be collected and managed by different parts of the organisation?

This raises several issues. From an efficiency point of view, it is important to avoid asking member countries for the same or similar information for several purposes; so OECD should as far as possible reuse what is already collected. But there are also quality considerations. Most importantly, consistency is an aspect of quality advocating that the same measures and measurements should be used for the same

³ See for instance discussion in Sundgren, Thygesen, Ward (2007): A model for structuring of statistical data and metadata to be shared between diverse national and international statistical systems,

<http://www.oecd.org/dataoecd/5/58/38541998.doc>

concepts in different areas. This becomes particularly important when the statistics and analyses are to be evaluated and understood by others than members of one specific committee.

Although OECD is essentially a forum where governments meet and discuss, increasingly OECD advocates spreading of knowledge to larger communities of users. If we strive for reforms of societies endorsed by the whole of societies, the statistical indicators used for evaluating and monitoring progress must be known and be transformed into knowledge of citizens. And when statistics are being made available to researchers and societies at large, it may become very difficult to explain the differences between different domains, unless there is a cohesion and consistency between them.

3.2 Integrating the statistics – the data warehouse

OECD has an immense statistical information bank, virtually all of which is uniquely held by the Organisation. Provided they are properly organised and managed, the data can be combined and used over and over again, for innumerable outputs. Until now, this potential has been underutilised, because each database was developed and managed independently using a wide variety of sometimes incompatible technologies. To activate this potential, the Organisation has developed a single system, the OECD Statistical Information System.

The concept of the Statistical Information System encompasses the whole production process from data collection to dissemination, including the management of statistical metadata (the explanatory text that accompanies the numbers). But the system is not just a technical solution. It also encompasses an ongoing, quality assurance process to ensure that data and metadata are harmonized across themes and core reference data (such a GDP) is re-used across all themes.

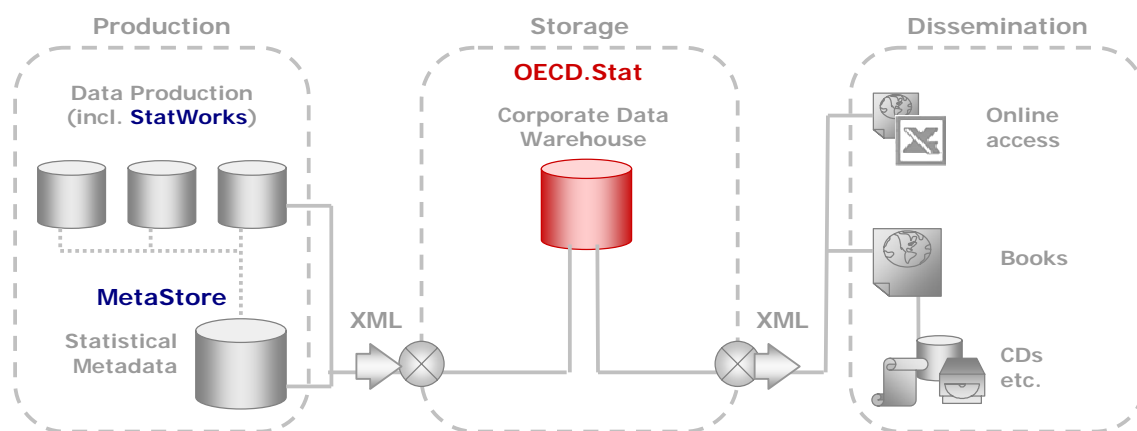


Fig. 1: A schematic diagram showing the three main parts of the Statistical Information System, centred around the data warehouse OECD.Stat

The main idea is to move away from the “stove-pipe” production model where each product is produced in its own complete “production line” from start to end, independently of all others, towards a common environment where different types of statistics are produced, stored and disseminated using the same tools.

The new browser for the common data warehouse OECD.Stat, developed in-house, offers a range of state-of-the-art features and functions. Data can be found using full text search, browsing by themes and by accessing stored queries. Metadata can be shown in an advanced manner, attached to the appropriate level of detail of the statistics. A feature gives the possibility of accessing and combining multiple datasets in one query.

The benefits and potential of the Statistical Information System can be summed up as follows:

- *Quality*
 - Harmonisation of concepts across themes, thus opening the way for cross-cutting (horizontal) analysis
 - Improved management and presentation of statistical metadata, now attached to data points at any level of detail inside a dataset rather than being in a separate document
 - Better coherence of data and metadata across datasets
 - Application of International Standards now possible across all databases (e.g. SDMX)
- *User friendliness*
 - OECD.Stat is a true one-stop database, with the potential to offer a single access method for all OECD statistics
 - Possibility to combine data across themes
 - Possibility to develop alternative outputs for different audiences
 - Easy access in very few clicks
 - Common look and feel of all databases
- *Internal Efficiency*
 - Common tools reduce the need for training on multiple systems and improves mobility of statisticians across OECD departments
 - OECD analysts need only learn how to use one system to access all databases
 - ITN need only maintain and support one system, not many

The OECD.Stat system has been favourably evaluated by other organisations and is already being shared with IMF, who has adopted the system. A Memorandum of Understanding stipulates how the two organisations will work together on further development of the system.

3.3 Recent developments in OECD.Stat

Content

Data content in OECD.Stat continues to gradually increase, and it now covers almost all domains, although in many areas the data do not include all existing statistics. A few domains still have very weak coverage (e.g. Environment), mainly because of lack of resources to carry out the migration. But it is fair to say that OECD.Stat is becoming a real data warehouse, the place where all statistics can be found.

More complete metadata are being added to better describe the data. The MetaStore management system helps data managers organise their metadata in accordance with corporate guidelines.

Functionality

The Web Browser is a web based tool that lets internal and external users interactively find data in the data warehouse. Functionality is being continuously developed in response to expressed user needs. A user satisfaction survey for the web browser has been carried out, giving input to future developments.

Examples of new functionality of the web browser are:

- A new schema for statistical metadata has been introduced in the summer 2007, showing the full structure of the statistical metadata.
- Pivoting functions are being improved.
- A number of additions to the functionality have been recommended by the OECD's international Task Force on Dissemination (see below) and will be made.
- Development of dynamic graphics to be created on the fly from OECD.Stat is underway.

Access can be made directly to the database, whereby users can see only the data they are entitled to access. In addition to this, all statistical metadata are being put out and made searchable on the Internet in a structured way to enhance the discoverability via Google and other search engines. Once you have found the metadata and want to access the data, links lead to the publicly available data as well as to SourceOECD. . This has already been launched in 2006 for the database Main Economic Indicators⁴. More metadata will be put on the Internet in the coming months.

An SDMX⁵ web service gives internal and external users the possibility to query the database and get SDMX outputs that can be directly imported in the recipients' own systems, provided these are compatible with the SDMX standards. This web service is used on a regular basis for data sharing and is being enhanced,

Subroutines named "DotstatGet" can be called from FAME and (soon) from SAS, allowing users to integrate specific data directly into their own analysis systems.

A new Excel add-in, the ".Stat Populator", is a user-friendly tool that allows extracting data structures, data and metadata from the central data warehouse OECD.Stat directly into Excel. Based on the engine of a similar tool for Fame databases, this add-in provides access to all information through the means of an Excel function. User-friendly and simple syntax makes it easy for users to browse the data warehouse content and construct queries. As it can connect to the external data warehouse web services, the add-in can also be used by users outside the OECD.

Access to the data warehouse

Internal access to the database from the Secretariat has been open since 2004. Internal users may use the web browser to navigate and find out what is there, but they often find it more expedient to use the other tools to integrate the data directly into their analysis systems.

External access to the data warehouse has been gradually developing. OECD's Publishing Policy distinguishes three communities of users. All information is made freely available to governments, mainly through a secure extranet called OLIS. Non-government users can get access to the full scale of information, including statistics, by paying a subscription fee to a service named SourceOECD. And through the OECD web site, all other members of the public can get access to "basic data", i.e. all OECD statistics of a general interest.

After starting giving access to OECD.Stat to government officials through the extranet in 2005, OECD started in mid 2006 giving public access to limited sets of data, in accordance with the Publishing Policy, through the Internet. As can be seen from Figure 2, usage is growing steadily and now averages close to 300,000 data views per month.

⁴ *Optimising Data Accessibility via Reference Metadata Management Principles*, Russell Penlington, OECD, Q2006 Cardiff April 2006. <http://www.statistics.gov.uk/events/q2006/downloads/WedSessions1-7.pdf>

⁵ SDMX is a recently agreed international standard for the exchange of statistical data between International Organisations and National Statistical Offices.

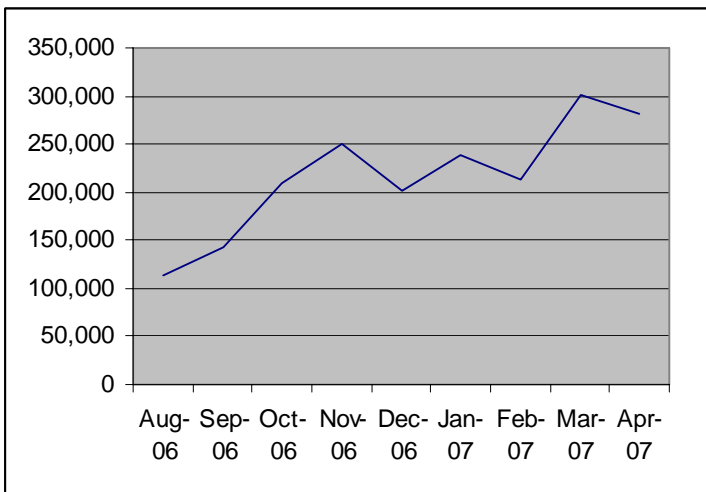


Figure 2: Monthly number of data views from OECD.Stat, August 2006 – April 2007

4. Content guidelines

OECD has adopted a set of guidelines for the contents of the statistics in the data warehouse, see references. It lays down the structure of the warehouse, based on a hierarchical thematic structure and gives guidelines as to how to delimit and structure datasets in order to make them as accessible and understandable as possible.

Very important are rules about common dimensions. These are dimensions which exist in many different domains, such as country, time and frequency, age. Dataset managers must use such common dimensions in their datasets in order to improve coherence and to allow integration of data from two or more datasets into one database query. Common dimensions, as well as the thematic structure, are maintained centrally.

5. Metadata management

The Achilles heel of many statistical systems is metadata. While a book can be provided with explanations and footnotes to explain concepts, limitations etc., it is not easy, on an ongoing basis, to ensure that data in large databases are always carrying the necessary metadata. The metadata may be scattered in many media or may even exist only in the head of the statistician responsible for the statistics. In this way, data can only be properly utilised after consultation with the relevant expert. In the Internet age with large numbers of users who are normally far away from the statistical organisation, this is quite untenable.

OECD has set up, as part of its Statistical Information System, a metadata management system called MetaStore. It gives the facilities to link all necessary metadata to any data element: A dataset (or database), a dimension, a dimension member, a time series an observation. It also facilitates the proper management of good metadata and interlinks directly with other elements of the Statistical Information System.

In addition to and linked to MetaStore, OECD has adopted a set of metadata principles, containing guidelines for the kinds of metadata that should always be provided, see references below. The guidelines present a standard for classification of the metadata in 42 subcategories or items, based on the SDMX work on standardising statistical concept. All metadata must be put and presented under one of these items. The

principles also state how the metadata should be presented in the web applications accessing the data warehouse.

One innovative feature in the metadata management system is reflected in Guideline 4, which states: “Where possible, preference should be towards the use of existing metadata – no new metadata elements are created until the proposer has first determined that no appropriate metadata element currently exists. The metadata management environment will issue warnings in such cases. All metadata must be created only once, for efficiency reasons and also in order to avoid the insertion of duplicated and/or inconsistent metadata into MetaStore.” The purpose is to promote standardisation and coherence across domains, allowing one manager to be responsible for metadata about the same concept reused in other databases.

6. Lessons learnt – Do’s and don’ts

From OECD’s work with the data warehouse, as well as from the experience of other statistical organisations engaged in similar endeavours, the following lessons may be drawn:

1. **Engage users of the information**, e.g. government analysts and policy makers (or in OECD, substantive policy Directorates and governments). Make sure they think a data warehouse will help them. Listen to their demands
2. **Study best practice** and steal as much as possible. Don’t “build you own” unless there are very distinct reasons for it. There is a great temptation in any organisation to build your own. Remember that you are not alone in the world, many organisations are solving similar problems as yours.
3. **Top management engagement** is essential – without it, no data warehouse will ever work. Everyone in the organisation must know that it is the corporate policy to unite all data under this umbrella. Steps in this direction shall be seen as successful and worthy of praise and honour - and *vice versa*. Engagement must be persistent through a long period of time.
4. The management and the project leaders must invest in **spreading support and understanding** among all involved⁶. There will be a strong incentive to defend old ways and systems – they will always contain some special features, to which their owners and often some users are devoted.
5. A certain **reallocation of resources** will be necessary from subject matter units to the central data warehouse unit – making acceptance more difficult. This is a problem that must be specifically addressed.
6. The data warehouse system must **offer tempting tools**, making it clear to database managers that they have something to gain for themselves – it is not only for the future, unknown users. Such benefits could be: an easier way to manage good metadata; an easier way to produce publications or other output media considered necessary by database managers; an easier way to fulfil mandatory data requests, e.g. from international organisations.
7. **Offer central assistance** to data managers who migrate to the common system. Lack of resources in the organisation may be a perceived or real obstacle.
8. **Some element of obligation** may be necessary from the side of the management. Even though much can be done to persuade all database managers that it is beneficial to adopt the common model,

⁶ “there is nothing more difficult to take in hand, more perilous to conduct, or more uncertain in its success, than to take the lead in the introduction of a new order of things. Because the innovator has for enemies all those who have done well under the old conditions, and lukewarm defenders in those who may do well under the new. This coolness arises partly from fear of the opponents, who have the laws on their side, and partly from the incredulity of men, who do not readily believe in new things until they have had a long experience of them. Thus it happens that whenever those who are hostile have the opportunity to attack they do it like partisans, whilst the others defend lukewarmly, in such wise that the prince is endangered along with them.” The Prince, Nicolo Machiavelli, Florence 1515. Chapter VI: Concerning New Principalities Which Are Acquired By One's Own Arms And Ability (here quoted from <http://www.constitution.org/mac/prince06.htm>)

experience shows that there will always be at least a few managers in the organisation who will never be convinced. For this reason, some “Stalinist” measures have to be taken, e.g. through a corporate performance management system or by way of mandatory contracts of results.

9. Set up **governance rules** defining responsibilities of different parties, e.g. management, database managers, data warehouse coordinators, IT department, Publishing.
10. Set up **data content standards**, so that everyone knows what is expected. Standards should build, as far as possible, on best practice for content modelling or conceptual modelling.
11. Set up **metadata standards** and tools to assist them. Experience world-wide shows that getting sufficient quality metadata, linked to the proper data, is a big challenge. No organisation has so far succeeded completely.
12. Start migration with the **managers having a positive attitude**, making success visible and desirable to copy. This could also be labeled: Start by picking low hanging fruits, show quick results.
13. Make sure **database managers keep ownership**. Make use statistics available on a per database basis, and make number of accesses a success parameter for the owner divisions or directorates.
14. It is **hard and continued work**. It is not enough to start it up. Maintaining and improving the service is an ongoing task – for ever.

References

1. The Statistical Information System is briefly described in http://www.oecd.org/document/44/0,2340,en_2825_293564_33869292_1_1_1_1,00.html
2. A view of a limited version (publicly available datasets) of OECD.Stat can be seen on <http://stats.oecd.org/wbos/> Government officials can view the full database on their OLIS account (using logon) from <https://www.oecd.int/olis>
3. The SDMX web service of OECD.Stat, allowing users to download data (and later metadata) in SDMX-ML format, is available at http://stats.oecd.org/OECDStatWS_SDMX/test_sdmx.aspx
4. OECD’s metadata principles are found in <http://www.oecd.org/dataoecd/26/33/33869551.pdf>
5. Guidelines for OECD.Stat Contents, Version 1, 6 September 2006 (internal document, available on <http://www.oecd.org/dataoecd/22/3/39137688.pdf>)