

NATIONS UNIES

COMMISSION ECONOMIQUE  
POUR L'EUROPE

ОБЪЕДИНЕННЫЕ НАЦИИ

ЭКОНОМИЧЕСКАЯ КОМИССИЯ  
ДЛЯ ЕВРОПЫ

UNITED NATIONS

ECONOMIC COMMISSION  
FOR EUROPE

---

SEMINAIRE

СЕМИНАР

SEMINAR

---

STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN  
STATISTICIANS



CES/SEM.35/SV/3  
12 July 1996

Original: ENGLISH

Seminar on Official Statistics - Past and Future  
(Lisbon, Portugal, 25-27 September 1996)

**SESSION 5: OUR LEGACY TO FUTURE GENERATIONS**

**METADATA SYSTEMS TO TURN NUMBERS INTO INFORMATION**

Report submitted by Australian Bureau of Statistics 1/

**INTRODUCTION - the strategic context, new pressures and old problems**

1. In common with other statistical agencies, the Australian Bureau of Statistics (ABS) is finding that more and more users are accessing data in electronic form. Our clients are demanding and are willing to pay for, "over the counter" statistical databases, online databases and other electronic dissemination services. They want to integrate these products and services with their decision making processes and decision support systems, and they require immediate answers to any questions they have about the concepts, definitions and methodologies underlying the statistics they obtain from these products and services.

2. It would seem that users value statistical information more highly than they have in the past. This may be due to the fact that information technology now enables clients to use statistics more

---

1/ Prepared by Warren Richter and John Cornish.

GE.96-

effectively. However, it is also clear that increased competition and the increasing complexity of business and policy problems facing organisations (although this may be a perception rather than a fact because people now have more information about the issues) is leading to demands for more and

more statistics, particularly for more detail than has been quality assured in the past. The increased pace of change in most aspects of business, government and society also leads to demands for more timely statistics and more timely delivery of these statistics.

3. These changes are driving agencies to invest resources towards the development of statistical and information systems infrastructure that will provide a means of delivering statistics to users in their terms and in the form they require.

4. However, while new technologies can help us increase the quantity of statistics and the level of service we provide, there are old, persistent gaps in the quality of our statistics. Agencies have a duty of care to provide extensive explanatory material to current and of course, future generations of users. Some progress has been made by the ABS in standardising and explaining differences in the statistical concepts used across its range of products and over time (i.e. with the aim of optimising statistical reliability). The quality of metadata contained in our printed products (particularly the explanatory power of annotations to tables) could be better but at least they are available with the data. The metadata in our electronic products is however, sparse, generally of lower quality and difficult to access as it is usually not linked directly to the underlying data. Clearly, presenting statistical information in electronic form presents special challenges.

5. Given that an increasing proportion of our current users access statistics electronically, how are we to meet these challenges and at the same time, fulfil our duty of care?

#### Essential attributes of statistical data and metadata

6. The Scandinavians have contributed a great deal to the development of conceptual frameworks for statistical metadata. Professor Bo Sundgren from Statistics Sweden has undertaken two consultancies at the ABS in the last few years, and much of what follows is derived from his work and that of Professor Svein Nordbotten, although there are some new elements which reflect ABS experience with the development of the ABS' information warehouse (also known as the ABSDB).

7. If agencies are to achieve the fundamental objectives of ensuring their statistics are widely and correctly used, then their statistical information systems (Professor Sundgren would call these "statistical metainformation systems") should ensure that from the user's point of view, statistical data and metadata have certain information attributes such as visibility and accessibility. Although other information attributes such as timeliness and accuracy have always been essential, advances in information technology are now making it possible to achieve other ambitions -- and to address the problem of satisfying our duty of care "electronically". In this context, it is suggested that statistical

data and metadata should possess six "new" information attributes. They are summarised below together with a brief description of the approach the ABS is taking to achieve them.

Data and metadata should be:	Approach
visible	- Datasets should be described with adequate metadata in an electronic data catalogue which will act as the directory to all ABS information of potential interest to users. The catalogue supports a variety of search mechanisms capable of searching at the lowest level of detail (e.g. find every dataset using the term "retailing").
accessible	- data and metadata is loaded to an information warehouse (driven by the catalogue) with a single data structure. The electronic catalogue is created from the metadata and provides one interface to all the data. In turn, the warehouse minimises the volume and complexity of communications between output data sources and output systems and products. One set of dissemination facilities can be used for most data and metadata.
relatable	- by using standards wherever possible and linked metadata, warehouse facilities allow data to be readily related across common domains (e.g. industry, geography).
reliable	- consistency of data and metadata across output products, regardless of dissemination media, will be ensured by using one repository as the source for dissemination.
understandable	- data in the warehouse is linked to and able to be presented with, the concepts, sources and methods underlying the data.
media-independent	- data and metadata can be produced in both print and electronic form using common facilities.

#### Metadata for users

8. Users of statistics have widely varying needs and some will

always require more detailed explanations of concepts, sources and methods etc than others, but in our view, the following three sets of metadata constitute the minimum metadata needed to fulfil the duty of care:

- definitional metadata: metadata relating to statistical units/populations, classifications, data items, standard questions and question modules for collection instruments, and statistical terminology;
- procedural metadata: metadata relating to the procedures by which data are collected and processed; and
- operational metadata: metadata arising from and summarising the results of implementation of procedures, e.g. estimates of sampling error, response rates, imputation rates, time series knowledge.

9. Ideally, these metadata should be readily available to all users in the same form as the data to which they refer: i.e. if the data is in the form of an online time series service, the metadata should be accessible through the same interface; if the data is held in a "stand-alone" database product such as the ABS' Integrated Regional Database on CD, then the metadata should be an integral part of the database.

10. In addition, other forms of metadata such as key features and commentary on the results of surveys may be appropriate. Commentary on breaks and turning points in time series should also be available with and in, the same form as the data, as should annotations to statistical tables, including annotations at the cell level.

11. In its publications, the ABS provides sufficient metadata for most users but users needing more detail about complex areas such as national accounts and balance of payments can obtain separate concepts, sources and methods publications.

12. There are also practical limitations on the amount of detail that can be packaged within some electronic products such as online statistical services and databases and it may be necessary to provide a "metadata service" accessible to users. Some progress has been made with a new concepts, sources and methods products on CD (which has introduced some innovative methods for explaining the basis of the national accounts) but at this stage only a small proportion of the total metadata is covered.

#### Producing metadata/data and metadata management

13. In an ideal environment, metadata would be produced and captured

only once and as a by-product of day-to-day statistical operations. This "ideal" environment would have metadata management facilities which make it easier for statisticians to do the right thing rather than the wrong thing (Statistics Netherlands have a nice catch-phrase - "tools rather than rules"). However, this is easier said than done, particularly during the transition from subject area-based systems to corporate systems, and we are therefore putting in place a fairly rigorous set of data management policies. An important feature of the policy implementation strategy is the identification of control points in the statistical process and the specification of actions and requirements around these points (particularly the identification of actions associated with specific components of the new corporate metadata management systems).

14. For example, it is ABS policy:

- to maintain a corporate repository, readily accessible to all staff, providing storage, retrieval, and updating facilities for the operational metadata used in the conduct of collections.
- that operational and procedural metadata are stored in the corporate repository wherever it is practical and efficient to do so.
- that, before final outputs from the collection are disseminated, the collection, processing and analysis procedures, the output data items and datasets are fully defined, documented and approved.
- to maintain a corporate repository, readily accessible to all staff, providing storage, retrieval, presentation and updating facilities for output datasets comprising data and the corresponding metadata.
- that the metadata describing all macro or micro output datasets are stored in the corporate repository.
- that all macrodatasets produced for publication or other planned dissemination activities are stored in the corporate repository.
- that all publication or planned dissemination activity based on macrodata must take place by drawing the macrodata (and the metadata) from the corporate repository.
- that all microdata from which dissemination activity is undertaken is either stored in the corporate repository, or, if stored elsewhere, must be specified in terms of metadata stored in the corporate repository.

Metadata management systems

15. Four corporate metadata systems support these objectives, and the information warehouse provides other facilities for managing annotations and analytical comments as follows:

- Collection Management System. This supports the entry, access and update of procedural and operational metadata relating to collections.
- Data Items Management System. This enables creation of and access to detailed data item definitions which exist independently of datasets and to which individual dataset data items may be linked.
- Classifications Management System. This provides the mechanism whereby users can access, and link together as appropriate, codes, descriptions, levels, and hierarchies of a classification that has been previously loaded. All major standard classification metadata are copied from the ABS Classification System which predates the warehouse and will eventually be replaced by it. Other (mostly non-standard) classification metadata are loaded from a variety of sources in conjunction with specific datasets.
- Publication Assembly System. This provides facilities for formatting publication tables produced by the warehouse and linking them to metadata, key features and contact officer information etc within an electronic manuscript which is then used to produce a paper publication. Similar systems are needed for producing electronic products in accordance with standards and with appropriate metadata.
- Annotations facilities. The warehouse stores annotations together with the statistical data and can deliver these annotations to dissemination software such as spreadsheets (and eventually to all ABS dissemination vehicles) and to other facilities such as the Publication Assembly System.

**Servicing future generations -- the risk of technological change**

16. Apart from improving data and metadata practices and systems, and of course the service to our current and future users; implementation of the warehouse/single repository approach has the capacity to position agencies to respond quickly and cheaply to technological change. For example, agencies with multiple output systems on different software and hardware platforms will eventually face the need to migrate or at least interface these systems to new dissemination and storage systems such as

modern database systems, online analytical processing software, and new dissemination facilities such as the Internet. Although initially ABS output systems have to be converted/linked to the warehouse, from then on all output can be migrated or interfaced to a new system with one rather than several system changes. In our view, this is extremely important as the pace of change in information technology means that it is not at all clear "which way to jump" when providing new electronic services, when archiving and preserving data and metadata for the future, and when developing new statistical processing systems. If multiple systems are involved, the risk of leaving some data and metadata behind is greater (where it may be lost to future generations) and the cost of making the wrong choice can be substantial.

17. It has been suggested that one way of reducing this risk is to adopt standards. However, it is not always clear which standards to adopt. An example of this is the relative failure (in Australia and elsewhere) of the Government Open Systems Interconnection Protocols and Open Systems initiatives of the early nineties. These policies were "intended to achieve greater interoperability, scalability and portability through a commitment to use international standards, but the technology advanced too rapidly for the standards process to accommodate it and the potential benefits of these policies were never fully realised". 1/

18. In these circumstances, an appropriate strategy is to adopt an architecture for information technology that is based on accepted standards (where they are well accepted) and which provides as much independence as possible from proprietary hardware platforms, operating systems, networks, and database management systems. The other critical element is to be well positioned to move quickly to take advantage of new technologies and opportunities to service users by having data and metadata available in one place and in one format.

## **CONCLUSION**

19. Advances in information technology enable statistical agencies to be more innovative and to provide a better service to their users. In using new technology, agencies have a duty of care to support informed decision making by current and future users by providing ready access to metadata.

21. Metadata systems provide the means to capture and manage metadata in order to turn numbers into information but on the basis of the ABS' experience they will be difficult and expensive to put in place and the question we have to answer is "is it worth it?" The following quote from Professor Sundgren's latest report on the project may provide the answer:

"Realistically, for a long time still, the data and metadata providers from individual subject matter areas will feel that they have to do more work for the Warehouse system, providing the Warehouse with high-quality data and metadata, than they will actually get back from the system in the form of useful services. They will be able to make productive use of data, metadata, and experiences originating from other areas when they design, develop and operate their own surveys, and they will be able to benefit from generalised functions - like publishing - that are associated with the (Warehouse) system. But it will take time before the data and metadata management of survey design, development and production systems are so well integrated with one another and with the (Warehouse) infrastructure that they will feel that they recover more benefits from this infrastructure than they have put in to maintain it. My impression is that the subject matter people have accepted this, and they understand that the net effort that they make is indeed very well justified when the benefits of the Warehouse to the organisation as a whole and, most importantly, to the external users/customers of ABS services are taken into account."

---

1/ Framework and Strategies for Information Technology in the Commonwealth of Australia. Office of Government Information Technology, Canberra Australia 1995. pp25.