NATIONS UNIES | ОБЪЕДИНЕННЫЕ НАЦИИ | UNITED NATIONS

COMMISSION ECONOMIQUE POUR L'EUROPE | ЗКОНОМИЧЕСКАЯ КОМИССИЯ ДЛЯ ЕВРОПЫ | ECONOMIC COMMISSION FOR EUROPE

SEMINAIRE | СЕМИНАР | SEMINAR

STATISTICAL COMMISSION AND
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN
STATISTICIANS

Distr.
GENERAL

CES/SEM.43/3
8 February 2000

ENGLISH ONLY

Seminar on integrated statistical information systems and related matters (ISIS 2000)
(Riga, Latvia, 29-31 May 2000)

Topic I: Data warehousing and the development and use of statistical databases in a network environment

PLACE OF OUTPUT DATABASE IN THE SURVEY PROCESSING LOOP

Invited paper

Submitted by Statistics Canada[1]

I. INTRODUCTION

1. National Statistical Offices (NSOs) were early adopters of computing technology to automate survey processing. The first phase (1960s-70s) involved the processing of survey files once the data had been transcribed from paper questionnaires into electronic files. During this phase, database technology was used to provide flexibility in system implementation and operations. The second phase (1980s) introduced the use of microcomputers to capture survey data at source through CAPI, CATI, etc. This allowed extending the benefits of computer processing to the input side of survey programmes. We are now in the third phase where Internet allows publishing the results of survey processing (the statistics) directly from databases onto Internet and to deliver the information electronically to clients’ screens. Although publishing/dissemination is still in a transition period from mostly paper-based to mostly electronic-based delivery channels, the trend towards having an Internet enabled society in the future is clearly visible.

2. As measured by Statistics Canada, 23% of Canadian households had access to Internet from home in 1998, up from 16% in 1997. 36% of households (any one member) had access to Internet, mostly in their work place in July 1999. Traffic to Statistics Canada’s web site is increasing by more than 50% annually. Our most recent market research has shown that more than 95% of our business clients have access to Internet from work.

3. Databases play a crucial role in this closing of the loop of electronically collecting, processing and publishing statistics. Different NSOs have adopted different strategies for integrating databases in this task. Statistics Canada follows an evolutionary path where existing databases will continue to exist for individual survey processing systems while corporate output databases will contain

[1] Prepared by Martin Podehl.

the output statistics and their metadata from all surveys. The corporate output database, as a web based data warehouse, feeds the various publishing/dissemination products, services and access and distribution channels.

4. This paper outlines the current development of the Statistics Canada output database and its placement within the survey-processing loop from data collection to dissemination. As well, some of the conceptual and practical issues of developing and operating an output database are discussed.

II. INTERNET

5. Since 1995, the Internet has emerged as an important dissemination vehicle for National Statistical Offices. While there has been some time lag between different NSOs adopting Internet for dissemination purposes, by now Internet is seen as the principal dissemination channel of the future. Similar to other organisations at that time, the first web sites were conceived on the notion of *telling* visitors about the particular NSO. Client feed back quickly changed the orientation to become a statistical information site, providing official statistics in a variety of formats to a variety of clientele.

6. The advantages of Internet as a dissemination channel have become obvious:

- one location (the NSO Internet site) where the variety of information published and released by an NSO can be accessed regardless of time and distance;
- timely release of the latest information with instant access by clients;
- opportunity to publish information in much more depth than would be feasible on paper;
- the opportunity to publish information much more in context by providing hyperlinks to related information such as details, explanatory notes, previously published information, quality indicators, underlying methodology, etc;
- cost avoidance in physical distribution compared to paper publications where each additional copy incurs costs for printing, order processing, shipping, billing, etc; on Internet, the marginal costs for having an additional client access an existing piece of information is close to zero for both the client and the NSO.

7. By now it is quite clear that electronic information services via the Internet, or its future variations, will become ubiquitous in society. The question is not whether, but when. The speed of transition depends on many factors, all of which seem to be addressed in a constant change mode: adoption of micro computers in homes; access to Internet at work, school or home; increasing communication bandwidth; costs of Internet connections; user friendliness of access, navigation and display.

8. In addition to using the Internet for dissemination, some NSOs have also started to collect survey questionnaires via the Internet. There are currently some barriers for widespread use of Internet for data collection purposes (access to Internet by respondents, security and privacy concerns, etc.), but clearly data collection will be another aspect of the use of Internet. Statistics Canada is considering offering citizens the option of filling out and returning their 2001 Census questionnaire on the Internet. The benefits of having the respondents fill out the questionnaire and sending it to the NSO are obvious. This paper will not discuss this aspect. However, it is part of the *closing of the loop*.

III. DATABASE PUBLISHING

9. Internet allows to disseminate information cheaper and more effectively than through conventional paper publishing methods. While it is true that, in contrast to paper publications, the marginal costs of informing an additional client through Internet is very low if not close to zero, there are significant costs in operating an Internet site and in developing and updating content for it. In particular, as the content grows (e.g. Statistics Canada has over 60,000 pages on its web site) the costs of maintaining and updating individual HTML (HyperText Mark-up Language) pages

become significant. Methods have to be employed through which such pages are created and/or updated in some dynamic and automated form using an organized set of information as the source. This is referred to as database publishing.

10. The main concept of database publishing is to separate the maintenance of the underlying information from the representation of its contents as HTML pages. This has two advantages:

- As new information is added to the database, new or updated HTML pages can be generated automatically without any manual intervention and coding.
- By separating the two functions, improvements can be made to either of the two functions without necessarily impacting on the other.

11. Statistics Canada has embraced the concept of database publishing as a fundamental design concept of its Internet service. Information on our site is grouped into information modules with each module representing a particular set of pages or documents of the same nature. Examples of our more popular modules are:

- *The Daily:* our daily news release;
- *Canadian Statistics*: a set of statistical tables about Canada;
- CANSIM (CANadian Socio-economic Information Management): our time series database;
- Trade: our detailed database of monthly commodity exports and imports;
- Downloadable publications: electronic versions of our official publications;
- Online Catalogue: our catalogue of all products and services.

12. Some of these modules are actual databases in the sense of a DBMS (Database Management Systems). Others are an organized set of documents/pages.

IV. TIME SERIES DATABASE CANSIM

13. CANSIM is Statistics Canada’s online time series database. All major socio-economic statistics are stored in great detail in CANSIM as time series with varying frequencies and length of series, some starting as far back as 1914 (e.g. Consumer Price Index, monthly). The database is updated daily and the latest data points are released at the same time as summary information is released in The Daily. Currently, CANSIM contains about 700,000 time series. Between 1973 and 1996, CANSIM data were made available to the public only through commercial online database services (e.g. Reuters, Wefa, Datastream, etc) under license with Statistics Canada. In 1996, Statistics Canada added its own commercial online dissemination service by interfacing a copy of CANSIM to its Internet site. This daily updated database has become the source for two types of services:

- **Direct online access to time series:** Using an interface programmed with CGI (Common Graphical Interface) scripts for input specifications and HTML pages for output presentation, clients search the CANSIM directory meta data, select the time series of interest, specify the retrieval parameters, pay the specific retrieval fee (unit pricing based on number of time series requested) with credit cards via an electronic commerce service (operated by an Internet service provider and a bank), and receive the time series in the desired format displayed on the screen and for downloading to their micro computer in a variety of formats. This interface, in a sense, offers the traditional online service to analytical experts. The innovation here is the ease of use and instant response via the Internet and paperless payment method through e-commerce.
- **Updating statistical tables on the Internet:** Like many other NSOs, Statistics Canada started to publish on its web site a statistical overview of Canada, Canadians, and its institutions in a set of summary tables referred to as *Canadian Statistics*. These tables are grouped under four major themes: The Economy, The Land, The People, and The State. In 1995, *Canadian Statistics* was launched with about 100 tables. The current number is 350 and growing. Each table presents a certain subject, and its display has been optimized for the screen, i.e. scrolling is avoided where possible. The initial set of tables was created manually and kept

up-to-date manually. It quickly became obvious, that manual maintenance could not be sustained given the limited resources allocated. As most of the statistics are maintained in CANSIM, we hit upon the idea to update the *Canadian Statistics* tables automatically from the Internet interfaced copy of the CANSIM database. Software templates were developed for all tables where the data could be obtained from CANSIM. Each morning at 8:30 am precisely a clock initiated, automated process retrieves the latest data points from the CANSIM database, updates the tables, and posts them on the Internet site. The same process is also being used for the *Economic and Financial Data* table which is updated daily and corresponds to the content and format prescribed by the *International Monetary Fund's Dissemination Standards Bulletin Board (DSBB)*.

14. This update process of the *Canadian Statistics* tables on Internet is an excellent example of database publishing. It has the following benefits:

- No human intervention is required to keep the tables up-to-date.
- The layout of all tables remains consistent.
- The integrity of the figures is ensured as they are retrieved from the verified and authorized database.
- The data are released in a timely manner and are always current.

V. CANSIM AS THE OUTPUT DATABASE (DATA WAREHOUSE)

15. Encouraged by the success of using the existing CANSIM database as a source of data for electronic publishing, we are pursuing several developments to strengthen the role of CANSIM in this regard.

CANSIM II

16. The underlying database software for CANSIM (vintage 1969) has been redeveloped (using RDBMS software) to accommodate multi-dimensional tables, not just individual time series. This project is referred to as CANSIM II. CANSIM II will become the (output) data warehouse for all macro data and will be available on our Intranet and Internet sites. CANSIM II will be *the* source for direct data access as well as database publishing with increased scope. (Exceptions to this are the data from the Census of Housing and Population and the existing Export/Import commodity database which continue to have their own database systems for the time being).

Multi-dimensional table browsers

17. These software tools (sometimes referred to as OLAP tools) have become available recently (e.g. Beyond 20/20 from the company Ivation Inc.). They allow flexible and convenient browsing of multi-dimensional tables (cubes) as two-dimensional presentations on the screen. They allow powerful access to the flexibly structured database while still preserving the easily understood presentation of statistics in flat tables or as a set of time series. In addition to considering and using commercial table viewers, we developed our own table viewer as a baseline service. This software is based on SGML as the structured language for marking-up tables in HTML for Internet display.

Table creation software

18. The content of most paper publications is tables. As all these data will be stored in CANSIM II and as most of our publications will be re-engineered to become electronic publications on Internet, there is the opportunity here to generate publication tables automatically from CANSIM II. For example, the monthly publication *Retail Trade* is created automatically as soon as the latest estimates have been added to CANSIM. At 8:30 a.m. on the day of release of Retail *Trade* in *The Daily*, the fully composed table publication with all the latest details can be made available on the Internet for electronic access, viewing, and downloading.

Custom publishing services

19. Our recent experience is that sales of standard publications are steadily decreasing and that there is a growing demand for custom services. Database publishing can be used to create a custom publication containing data in tables from a variety of statistical source programs (surveys) with CANSIM II being the source for all the data. Since we already have an electronic commerce interface on our site, the associated costs can be charged to, and paid by, the client conveniently.

Metadata and the output database

20. Metadata (or Metainformation) must be part of the output database either very closely linked or at least loosely coupled. Two types of metadata are distinguished:

- Documentation and labels required so that a user can navigate the database and obtain the labels associated with the individual data points as well as the descriptions of the multi-dimensional tables in which the data points are organized.
- Information needed to understand and interpret the actual statistics.

21. With regard to the latter, information needed to understand statistical data falls under three broad headings:

(a) the concepts and classifications that underlie the data;
(b) the methodology used to collect and compile the data; and
(c) measures of accuracy of the data.

22. Essentially these three headings cover respectively: what has been measured; how it was measured; and how well it was measured. Users should be able to obtain easily the information what has been measured (to assess its relevance to their needs), how it was measured (to allow appropriate analytic methods to be used), and how well it was measured (to have confidence in the results).

23. Statistics Canada’s metadatabase is a comprehensive description of concepts, definitions, subjects, variables, methodologies and quality indicators about our statistical programs. This base was initiated in 1981 as the **SDDS (Statistical Data Documentation System).** SDDS is being enlarged and improved to become the **IMDB (Integrated Meta Database).** A record in this base pertains to a statistical source program such as a survey, administrative data acquisition program, or census. It also covers derived statistical programs, e.g. the various National Accounts programs that produce statistics from primary or secondary data sources. Each record has a unique identification number (referred to as the “SDDS number”) and up to 120 fields in which the various Meta information about the source program are stored.

24. The first version of the IMDB has been made available on our Internet site and linked to CANSIM. Each table in CANSIM (and CANSIM II) is hyper-linked to the documentation of the statistical program which is the source of that table. Thus clients can check with one click of the mouse the source of a particular time series or multi-dimensional table which they have selected for access and downloading.

VI. ISSUES

25. In the following, some of the issues in building and operating an output database are discussed along with the strategies which Statistics Canada is pursuing.

Scope of output database

26. Statistics Canada has defined the goal for its output database as to become *the single electronic window on all publicly available statistics to support all elements of our dissemination program, from printed publications to data access on the Internet*. This means that all statistics (estimates), which are derived from micro data files or analytical value added programs, should be stored in the output database. This goal is shared by most other NSOs that have built, or are in the process of building, a corporate output database. However, there is the question to what extent of detail estimates should be tabulated and stored in the database. While machine resources for the tabulating step are cheap and storage costs for the final macro data are low, relatively expensive human resources are required to specify the tabulations and the resulting tables with their documentation. As well, storing data in too much detail may clutter the database and obscure the more relevant and more broadly requested data. In other words, a trade-off has to be made between creating pre-defined tables for the database and offering custom tabulation services from the various source micro databases according to the specific and unique requirements of clients. Statistics Canada learned this lesson with the CD-ROM products for the 1996 Census. So much detailed breakdown was included that clients complained about the difficulties they had with finding data relevant to their needs.

Data model

27. Both time series and multi-dimensional tables need to be supported. In the most general case, time is just another dimension of a multi-dimensional table, however with special properties. Output databases have existed since the 70’s as time series databases for economic and other analysis. Existing clients wish to be able to access individual time series from a variety of survey sources to be loaded into their value added analysis packages. Our clients requested as much in our initial market research. Consequently, the *Consumer Price Index for Canada for all-items*, for example, can be accessed in the existing CANSIM and in CANSIM II by specifying its permanently assigned so-called data bank number *P100000.* In CANSIM II this time series can also be selected in *Table 1234* as the cross-section of *Canada* in the *geography* dimension, *all items* in the *item* dimension, and *seasonally adjusted* in the *type of estimate* dimension.

28. Multi-dimensional tables have been adopted by most NSOs as the basic entity in their output databases, but their scope may be different. Some NSOs have designed systems with very large multi-dimensional tables which allow “slicing and dicing” by common dimensions such as geography or standard classifications. This requires the standardisation of all concepts and labels, and the co-ordination of production and of joining of estimates from many different sources (surveys) along common axes. We tried this approach in our first design of CANSIM II but after two years came to the conclusion that this was too ambitious given the time and resources allocated to the development of CANSIM II. CANSIM II is now based on multi-dimensional tables where each table

- has typically only one source (e.g. a survey),
- is relatively dense, i.e. most cross-sections of dimensions have actual estimates (as opposed to holes because the particular combinations are not available or do not make sense).

29. The consequence for users is that they need to consult several tables if they are looking for estimates from a variety of sources.

Harmonization and standard definitions

30. Ideally all estimates stored in the output database, regardless of their source, adhere to standard classifications and concepts harmonized across sources. However, the statistical programmes, which are the sources for the estimates for the output database, most often have been initiated and developed at different times and their underlying concepts reflect the evolution of such concepts. For the output database the question is then whether to store estimates as they are produced from the various sources or to insist on having them adhere to standard definitions across the complete output database. Different strategies in this regard are being pursued by NSOs. But even those NSOs with strict criteria for standardization and harmonization of data stored in the output database have to allow for exceptions or to make compromises. The historical concepts on which surveys are based cannot be

changed quickly. Similar to the experience with the data model, Statistics Canada initially pursued harmonization within CANSIM II but concluded that populating the output database would take far too long in relation to the goal to establish as soon as possible a comprehensive output database as the central depository of all publishable statistics. We are now entering data as they can be provided by subject matter from their existing survey programs. Data will become standardized and harmonized in step with such efforts initiated by subject matter. No doubt that demands from external clients will provide additional impetus to do so.

Integration with processing systems

31. The actual implementation of an output database can be positioned between two extremes:

- The output database is a standalone system, completely uncoupled from the individual survey/source processing systems. A standard data format is agreed to in which each survey program passes the estimates and their documentation to the output database. Updates are passed as files which are applied against the output database.
- The output database is one of the many components of an integrated, corporate survey processing system in which all micro data from all sources are stored according to common documentation and system standards. Generalized software is then used to process the individual surveys from data capture to tabulation within this micro/macro data warehouse environment.

32. Both approaches have their pros and cons. Which path to choose depends on the specific circumstances and culture in an NSO. For example, Statistics New Zealand opted for the latter approach when it decided to decommission the mainframe computer with its legacy software. SNZ seized the opportunity to create a comprehensive, unified processing systems for most of its surveys. Micro data files are stored in a relation database that also houses the output database. Statistics Canada has adopted the former approach based on its desire to create a comprehensive output database as soon as possible with minimum interruption to the existing survey processing systems.

Internal vs. external database

33. An output database has two user groups:

- Within an NSO, the database allows access to data across sources for analysis and for the creation of pre-packaged products and services.
- Externally, the database is accessible online by clients and can be the source for secondary distribution by other public or private sector organizations under license to the NSO.

34. Ideally one database should fulfil both requirements. But practical considerations of security and operations may demand two versions, an internal and an external output database.

- Some NSOs have two networks to absolutely protect the confidentiality of the micro data stored within the internal computing/communication network. The internal network is not connected to the outside world. An external network is designed to allow public online access to information prepared for public consumption.
- The internal output database may contain more data than will be made accessible to the public in the external database. As well, estimates can be stored in the internal database prior to their official release time, and quality control is enhanced by comparing new estimates with previous estimates or estimates from other sources easily accessible in the database. Quality control of data before they are made final in the output database is a major issue. Any mistake would be promulgated immediately through the various dissemination channels which rely on the output database.

- Typically the internal database is updated from the various source programs. In turn the external database is updated from the internal database.

35. In the case of CANSIM II, we have two databases with the same architecture and software, one operating on the Intranet, the other on the Internet.

Database architecture and access software

36. The concept of Online Analytical Processing (OLAP) for access to and browsing of multi-dimensional, summarized business data was developed some time ago. It is only in recent years that access and visualization tools, which adhere to the OLAP concepts, have started to be offered on the market. Much is happening currently with regard to developing such tools for the Internet. However, these tools are only visualization tools. They expect that a database exist in some form to which then these software tools can be interfaced. Some OLAP tools also offer a database as part of their package. However, implementing an output database based on such an integrated tool carries the risk of being dependent on the particular vendor and being restricted in the choice of new or additional visualization tools. Statistics Canada decided to separate the database from the access and visualization tools in the design of CANSIM II. The database uses commercial RDBMS software. The access software, so far, is based on evolving from the traditional time series view to the multi-dimensional view. For the multi-dimensional view we will offer our own basic table viewer, but we will interface to popular OLAP tools as requested by our clients and/or where we see economic benefit. As said above, much is happening on the Internet currently and we expect more OLAP tools to be offered on the market in the future. We then simply add, as an option, the output format required for the OLAP tool.

37. Today computing and networks are one. We decided from the beginning to base the design, development, and operation of CANSIM II on Web based technology and to adhere to run-of-the-mill www standards (for example, so far we have avoided Java applets as most corporate firewalls will not let such software through). This allows us to take advantage of new features as developed on and for the net, be it browser features, or visualization packages. In particular, the hyperlink feature of the Web allows us to link CANSIM II to and from a variety of complementary information sources, such as general metadata, standard classifications (future), glossaries, and quality statements.

Free vs. for fee services

38. The ultimate goal of the output database is to be accessible directly in online mode (that means Internet these days) to external clients, in addition to being the source of data for all other dissemination and publishing programs. In the past, online time series databases were offered as value added, priced services. The Internet is changing this approach:

- Electronic dissemination/publishing is becoming the pre-dominant and default mode while paper based publishing will become a value-added service for most products (if not soon then in years to come).
- The marginal costs are near zero to service additional clients through the Internet. Much information is now offered for free on Internet, which hitherto were only available through priced products/services.

39. Some NSOs have decided to make access to their output database on Internet a free service. Others, such as Statistics Canada, will use the output database to increase the amount of free information as preformatted summary tables (as extracts from the output database) but will continue charging for access to the details available in the output database.

Government Online

40. The specific position of an NSO to pricing is also guided by the general policies of its Government to the use of Internet for improving communications with its citizens. The future may see more changes in this respect. The Canadian Government has initiated a major program called Government Online (GOL) with the target of making all its information and services accessible on the Internet by the end of 2004, mostly for free, but some for a fee with e-commerce as a payment option. Presently, many individual, departmental Internet sites exist which provide information to Canadian citizens and organisations, often including much statistical information obtained from Statistics Canada. We are actively pursuing with these partners arrangements where such information could be provided and updated from the output database (with full credit to Statistics Canada as the source). The next step is to *cluster* services where information and services from various departments are accessible under one entry point (so-called single window). Cluster themes are to address specific citizen (and organisation) needs such as Consumer information, Export information, etc. Much discussion and testing is going on currently. Again, statistical information from Statistics Canada is a highly welcome part of cluster services.