**Federal Committee on Statistical Methodology
(FCSM) Research Conference
Washington DC, November 14-16,2001**

**Session: Uses of advanced technologies for data collection, processing and dissemination, part 2**

# *Laying the foundation for electronic dissemination : Statistics Canada's output database, CANSIM II*

Prepared by:

Louis Boucher

Statistics Canada

Ottawa, ON, K1A 0T6, Canada
louis.boucher@statcan.ca

# Laying the foundation for electronic dissemination : Statistics Canada’s output database, CANSIM II

**Louis Boucher**
**Statistics Canada**

Abstract

Dissemination of official statistics via the Internet has entered a new stage. Creating individual HTML pages or PDF documents is expensive and requires manual quality control which is difficult to sustain on a continuing basis.  The next development is to create HTML pages on the fly from information organized and maintained  in data bases of text and numbers.

This paper will review the current approaches of electronic data dissemination in general and the practical experiences in Statistics Canada so far with its output database CANSIM II interfaced to its web site.

Keywords : database publishing, electronic dissemination, output database, internet publishing, data warehousing

## INTRODUCTION

In 1995 Statistics Canada embraced the Internet as a strategic dissemination channel.  Growth in access has been steadily growing in the last 12 months.  Currently, on average, more than 15,000 clients consult our site daily.

The Internet allows Statistics Canada to reach more people than ever before and to increase the effectiveness and efficiency of its publishing/dissemination program.  In particular, we are trying to balance the resource requirements for maintaining publishing in conventional media (e.g. paper publications) with the development and operation of electronic dissemination through the Internet.

From the start, we tried to avoid manual creation and maintenance of information on our Internet site and to use, as much as possible, automation through what is referred to as data base publishing, i.e. the process of generating and updating the content of our Internet site from data bases.  In this regard, our existing CANSIM output database plays a crucial role.  We are well on the way to transforming the existing CANSIM into the corporate data warehouse (CANSIM II) for all published and publishable statistics.

This background paper has two parts.  The first part  will outline our general approach on database publishing on our website.  The second part will provide an update of our CANSIM II development and the existing and planned processes to use CANSIM II as the data warehouse source for the various publishing/dissemination products and services.

## DATABASE PUBLISHING

***Internet as a dissemination channel***

Since 1995, the Internet has emerged as an important dissemination vehicle for national statistical offices (NSOs). 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 for the future. Similar to other organizations at that time, the first web sites were conceived on the notion of telling visitors about the particular NSO. Client feedback quickly changed the orientation to become a statistical information site, providing official statistics in a variety of formats to a variety of clientele.

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 much more in depth information than would be feasible on paper;
- the opportunity to publish information much more in context by providing hyperlinks to related information such as detailed data tables, 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.

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 now has over 60,000 pages on its web site www.statcan.ca ) the costs of maintaining and updating individual HTML (HyperText Mark-up Language) pages manually become significant. Methods have to be employed through which such pages are created and/or updated in some dynamic and automated form from an organized set of information. This is referred to as data base publishing.

The main concept of data base 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 impacting necessarily on the other.

Statistics Canada has embraced the concept of data base publishing as a fundamental design concept of its Internet service. Information on our site is grouped into categories called “information modules” with each module representing a particular set of pages or documents of the same nature. In the following, we describe some of these modules in more detail and indicate how data base publishing methods are used to make them accessible and to inter-link them on our Internet site.

***The Daily***

A popular feature of our Web site is *The Daily*. *The Daily* is the vehicle for first (official) release of statistical data and publications produced by Statistics Canada, provides highlights of newly released data with source information for more detailed inquiries. *The Daily* references (as hyper links) the publication titles with their catalogue numbers and the table numbers of the time series in CANSIM (see below) which contain more data as well as metadata details released at the same time as the announcement in *The Daily*.

Each issue of *The Daily* is added to a repository of all past issues. This growing set of individual issues functions as a database in the sense that keyword searches can be executed against all past issues. Data base publishing in the case of *The Daily* means creating a structured document each day (text, tables, graphs, hyper links) from which all disseminated versions are derived, and adding the most recent issue as a new "record" to a repository for future access.

### ***CANSIM***

CANSIM is Statistics Canada's online time series database. Since 1973 and until 1996, CANSIM data were made available to the public only through commercial online data base 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 data base has become the source for two types of service:

Using an interface programmed with CGI (Common Graphical Interface) scripts for input specifications and HTML pages for output presentation, clients search the CANSIM directory metadata, 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 for analytical experts. The innovation here is the ease of use and instant response via the Internet and the paperless payment method through e-commerce.

### ***CANADIAN STATISTICS***

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, The State. In 1995, *Canadian Statistics* was launched with about 100 tables. The current number is 400 and growing. *Canadian Statistics* is one of the most popular features of our Internet site. 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 became quickly 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 can be obtained from CANSIM. Each morning at 8:30 am precisely, an automated clock initiated process retrieves the latest data points from the CANSIM database, updates the tables, and posts them on the Internet site.

This update process of the *Canadian Statistics* tables is an excellent example of data base 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.

In creating the *Canadian Statistics* tables we took advantage of the intrinsic feature of Internet to offer hyper links from each table to more detailed information. For example, the specific time series in the CANSIM data from which the table was derived is linked as well as the publication itself where the analysis context can be found.

### *Catalogue and other Metadata*

Statistics Canada maintains and publishes two Meta databases, which are available on our Internet site.

- The first Meta database is a comprehensive catalogue of all products and services offered by Statistics Canada. A record in the **Online Catalogue** pertains to a specific product or service and uses fields to describe it in detail (e.g. catalogue number, author, abstract, subject key words, price, contact, etc). The Online Catalogue of about 6,000 records in each official language (English and French) is maintained in an ORACLE DBMS on an internal file server and is updated continuously. Once a day, the latest changes to this database are uploaded to our external Internet site and stored as HTML web pages (one page representing one record). The Online Catalogue can be searched by keywords directly by clients looking for information. As well, hyperlinks to the Online Catalogue exist in other information modules on our Internet site, e.g. *The Daily*, *Canadian Statistics*, CANSIM. Online Catalogue records also linked to the process for ordering a product for electronic (i.e. downloading from our site) or physical delivery.

- The second Meta database 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)**. It is now 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 which 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. In 1999, the existing content of SDDS/IMDB has been made available on our Internet site for access through hyperlinks from CANSIM and from the Online Catalogue. Clients can now check the source of particular time series which they have selected for access and downloading.

### *Downloadable Publications*

Similar to other NSOs, Statistics Canada has started to convert publications from paper-only distribution to electronic distribution in the form of Internet downloadable documents in HTML and

PDF (Portable Document Format, specifically Adobe/Acrobat). This in itself cannot be classified as data base publishing. But if one regards the total Internet site as a sort of structured "data base" then each publication issue can be regarded as a "record" within the publication module which in turn is part of the overall Internet data base. Similar to *The Daily* module of all past issues, this "publications module" can be searched by keywords and hyperlinks can be used to link publications module records to records in other modules on our Internet site.

## CANSIM II : THE OUTPUT DATABASE

### *Background*

In the mid 1990's, Statistics Canada decided to replace the CANSIM (Canadian Socio-economic Information Management System) database. CANSIM was conceived in the late 1960's and its success is attributable to the foresight of the creators and the discipline of its maintainers over the years. Housed in Statistics Canada's mainframe, CANSIM organizes common sets of macro data derived from various surveys into *Time Series* which in turn are grouped into *Matrices*.

Comprising over 700,000 time series (8,000 matrices), the CANSIM system has evolved from being accessible to economists and analysts through commercial online services in the 1970's to being referenced by people across the world through the Internet ( www.statcan.ca/cansim ) Each day, when Statistics Canada announces the release of a set of information in *The Daily*, the data corresponding to the release are simultaneously accessible on CANSIM.

The underlying data base software for CANSIM is being 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 data warehouse for all macro data available on our Internet site as *the* source for direct data access as well as data base publishing with increased scope.

The content of most paper publications is tables. As all publishable data will be stored in CANSIM II and as most 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 on *Retail Trade* has already been re-engineered to be updated automatically each month as soon as the latest estimates have been added to CANSIM. Once produced, these publications are then available in multiple formats: HTML, PDF, and paper.

Sales of standard publications have started to decrease. On the other hand, there is a growing demand for custom information services. Data base publishing can be used to create a custom publication which presents tables from a variety of statistical source programs (surveys) with CANSIM II being the source for all the data for those tables. Since we already have an electronic commerce interface on our site, the associated costs can be charged to, and paid by, the client conveniently.

CANSIM II is at the heart of database publishing, and in the future could offer several advantages, such as decreased publication costs and increased timeliness for release of publications. By having data for release resident in only one location, the potential exists for increased data accuracy. Hyperlinks to sources of metadata should improve interpretability and thus relevance of Statistics Canada's information to the end-user.

### *Lessons Learned*

A database model and a prototype database were developed in the first 2 years of the project development. The results gave merit to the idea that a multi-dimensional, relational database could be constructed to house and display data that followed harmonized and standardized concepts/definitions. The prototype however pointed out the substantial amount of effort (human intellectual) it took to restructure data and data labels into a rigid set of rules (arrays and cubes) prescribed by a relatively complex data model.

At the end of a two-year development period, a number of concerns remained unanswered, among them:

- accommodation of non-standardized/harmonized data,
- continued access to CANSIM data in the manner users were accustomed to
- the conversion of all feeding systems,
- the conversion of all output systems,
- the long term sustainability of the database by CANSIM operations staff,
- slow system response times,
- the time and effort (thus investment) it would take to have a full production system ready and operational
- the need to have two parallel systems in operation (and synchronized) during this time.

A considerable amount of effort had been placed on the database design and on structuring and inputting data into the database. Output facilities were not fully developed and as a result, subject matter areas could not see the multi-dimensionality of their output (meaning their input into CANSIM II). The inability to view something tangible led to a degree of frustration on the part of subject matter areas and the perception of a lack of progress with the project.

### *Building on the success of CANSIM*

We went back to the drawing board and, based on the experience gained, started with a simpler database design as an evolution from the existing CANSIM. In order to show and test the acceptance of multi-dimensionality from CANSIM, a relatively simple Paradox database was developed in which existing CANSIM time series could be presented as multi-dimensional tables.

With a handful of multi-dimensional tables, the Paradox prototype application was placed on Statistics Canada's ***Intranet*** web site for all to evaluate multi-dimensionality. The same application was later presented to a small group of external CANSIM clients over the ***Internet***. The feedback from within Statistics Canada and from the external group was extremely positive.

CANSIM operations staff started comparing the effort it took to structure and reference data using this database and the constraints of the CANSIM II prototype, and it was not long after, that a complete review of the future direction of the project was undertaken. Following a review of the findings, we decided that a more "evolutionary" approach was required to develop a final operational CANSIM II base quickly and to maintain it efficiently in the long term. The revised objectives were to:

- make CANSIM II operational as soon as possible
- build a relational database structure allowing for CANSIM data to be transferred to CANSIM II “en masse”, with the building of multi-dimensional structure and enhanced labeling afterwards

(this meant that the database had to contain both CANSIM and CANSIM II formats),

- encourage standardization and harmonization of data concepts/definitions, but not to make them a pre-condition to housing data on CANSIM II,
- design the database to allow for data that did not follow harmonized/standardized concepts/definitions.

### *CANSIM II: Development*

With the lessons learned from the two generations of prototypes, the development of the current CANSIM II relational database using commercial RDBMS software was started in February 1999. The development was divided into two major phases.

The first phase is the reconstruction of CANSIM as a relational database with all the restrictions, nuances, features, and idiosyncrasies built-in over the past 30 years. The first phase will culminate in the transfer of all CANSIM data (as is) into the new relational database, with continued ability to load, add, delete, modify, and update using current formats and tools.

The second phase adopts the multi-dimensional concepts developed in the Paradox prototype database described earlier. While phase 1 was underway, the CANSIM team has used the Paradox database to build multi-dimensional structure and enhance labels. The team has managed to rebuild 100% of the active CANSIM time series (approx. 400,000) as multi-dimensional tables. The amount of interest in the availability of CANSIM II started to grow internally and areas that had little or no interest became eager to have their data accessible through the multi-dimensional tables once they saw the possibilities offered to others.

A fully operational CANSIM II was made available internally as early as the spring of 2000 while our external users were introduced to it in the spring 2001. The transition to CANSIM II is progressing more rapidly than expected, feedback from users indicate that the userfriendliness and the extensive data tables available are the prime factors for them to let go the original CANSIM and move forward.

Obviously this document presents the two phases in a summarized manner. Several technical documents outline each phase and its components in fine detail and can be made available.

### *Introducing CANSIM II to the market place*

Statistics Canada has defined the goal for CANSIM II (as its output database) as *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.* Its exploitation will no doubt continue in an evolutionary fashion in response to emerging client needs and in recognition of new opportunities in linking pre-thought and pre-defined summaries (analytical text, summary tables) to the underlying details in the macro data warehouse.

The technical work of rebuilding CANSIM into CANSIM II is close to completion although changes will be made for quite some time in response to actual experiences, performance, and client feedback. Efforts now concentrate on pricing and marketing. While it is unclear what the actual demand will be for access to data as multi-dimensional tables in addition to individual time series, there is no question that our current pricing formula needs to be supplemented. Today, we charge $3 per time series over the Internet. The traditional time series access will continue in CANSIM II, but a simple multi-dimensional table could contain several hundred time-series. Applying the existing price formula would make the table unaffordable and effectively inaccessible. We are in the process of investigating a number of pricing options ranging from subscriptions of blocks of information (e.g. Labour, Manufacturing, Health, Economic Indicators, etc.) to packaging CANSIM II data with priced publications available electronically. The pricing options have to be acceptable to our secondary distributors with whom we have long standing partnerships.

We are not sure how much new value the multi-dimensionality of CANSIM II represents to our existing secondary distributors. Their clients may be quite content to continue with access to individual time series. Or, demand may increase substantially as new information becomes available in CANSIM II which was never stored in CANSIM because it was not of time series nature.

With the addition of new data from subject matter areas, CANSIM II will definitely be a richer source of data for our clients. Internally, CANSIM II will be the macro data warehouse of Statistics Canada providing economical and efficient database publishing possibilities for a variety of products and services. Externally, CANSIM II provides online access in new formats, new data depths, and with much better supporting Meta data.

## ISSUES

- Data base publishing requires expert resources for the one time development of the necessary data bases, systems and procedures. For an occasional or less frequent publishing program it may be simpler and cheaper to use software tools to create manually HTML pages from word processing texts or data in spreadsheets. HTML conversion tools have become easy to use. The trade-off between such a manual process and the automated data base publishing process needs to be evaluated for each case. On the other hand, once a data base exists, new opportunities can be exploited which are not feasible without such a data base.

- Stringent data quality procedures have to be instituted to verify the accuracy of the information before it is entered into the data base. This applies both to data and metadata. There has to be absolute confidence that the data in the data base are “correct” and that automatic data base publishing can proceed without further manual verification of data quality. We have had several experiences where our Internet visitors pointed out to us real or perceived inconsistencies in our *Canadian Statistics* tables generated automatically from CANSIM. On the positive side, once such errors have been found and corrected in the data base, all future presentations extracted from the data base will be correct. (In widely distributed paper publications such errors could not be corrected.)

- As paper publishing is more and more supplanted by Internet information services, the uptime of the Internet server becomes critical. If it is down, nobody has access to the information. This becomes even more critical with data base publishing: if the data base is down, nothing can be published.

- The interface between extracting data from a data base and their final presentation on clients’ screens has to be based on robust, standard interfaces so that any change in the Internet presentation technology does not require a change in the data base access interface. Statistics Canada has good experience with SGML in this regard. As much as possible, we build such interfaces using SGML as the interim format for information transfer from the data base layer to the presentation layer.

- The current speed of technological changes is phenomenal. Constantly, new Internet access and presentation features are offered, particularly in the form of plug-ins. Of course, one should take advantage of such generally accessible features. On the other hand, many clients may not have the necessary client platform or the technical skills to deal with complicated downloads etc. Thus a balance needs to be struck between forward looking design and conservative assumptions of the skills and infrastructure on clients premises.
- Electronic dissemination enable NSOs to present richer and more comprehensive data to it users like never before. The challenge then is to ensure that the users, from the “first visitor” to the experienced analyst has the proper tools to find and retrieve the needed data in an effective manner. Search engines and browsers are at the center of this process and great attention needs to be spent there.

## CONCLUSION

Internet has already changed fundamentally the way NSOs disseminate official statistics. Internet offers opportunities to reach more clients with more information in a more timely way and also to reduce the costs of the total dissemination process in the long run. The lower costs can only be achieved by automating as many steps as possible within the chain of producing statistics from collected survey data and putting them into the hands of the clients.

In this chain, a data warehouse of published or publishable statistics (macro data) will play a pivotal role as a central staging area: survey and other statistical programs deposit their estimates into this data warehouse; the various dissemination processes retrieve data from the warehouse to be disseminated in a variety of formats and distribution channels, foremost Internet in the future.

The data warehouse must accommodate both the actual estimates as numeric values as well as all labeling, explanations, quality indicators, methodological notes etc. associated with the statistics. Such a data warehouse can then be the primary source for publishing automatically in electronic form on Internet in a variety of packages and formats.