



## Economic and Social Council

Distr.  
GENERAL

CES/AC.71/1999/6  
11 November 1998

ORIGINAL: ENGLISH  
ENGLISH AND FRENCH ONLY

---

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

### CONFERENCE OF EUROPEAN STATISTICIANS

Meeting on the Management of Statistical Information Technology  
(Geneva, Switzerland, 15-17 February 1999)

Topic (i): The impact of Internet on the statistical production and dissemination process

#### DATA BASE PUBLISHING ON THE INTERNET

Submitted by Statistics Canada<sup>1</sup>

#### I. INTERNET AS A DISSEMINATION CHANNEL

1. 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 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.

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

- one location (the NSO Internet site) where the variety of information

---

<sup>1</sup> Prepared by Martin Podehl.

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 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.

3. By now it is quite clear that electronic information services via 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.

4. 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 now over 60,000 pages on its web site) the costs of maintaining and updating individual HTML (Hyper Text Markup Language) pages become significant. Industry figures state that currently on average 4 to 5 hours are required to manually create and then maintain one HTML page. 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.

5. 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 data base, 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.

6. 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 bins" with each bin representing a particular set of pages or documents of the same nature. Examples of our more popular bins 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 data base; Trade: our detailed data base of monthly commodity exports and imports; downloadable publications: electronic versions of our official publications; IPS (Information on Products and Services): our catalogue of all products and services. Some of these bins are actual data bases in the sense of a DBMS (Data Base Management Systems). Others are an organized set of documents/pages. In the following, some of these bins will be described in more detail and how data base publishing methods are used to make them accessible and to inter-link them on our Internet site.

## II. THE DAILY

7. The most popular feature of the Statistics Canada 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 enquiries. It contains weekly and monthly schedules of upcoming major news releases and announces new non-print products and new services. *The Daily* is released every working day at 8:30 am. It is written for the media but is also of great interest to analysts in government and industry. *The Daily* is a highly structured and thoroughly edited document. Major statistics are summarized under the rubric AMajor Releases@ with highlights, statistical graphs and summary tables. Other statistics are announced in short paragraphs of a few lines. As well, *The Daily* references (as hyper links) the publication titles with their catalogue numbers and the matrix numbers of the time series in CANSIM (see below) which contain more data details just published by each statistical program.

8. Each issue of *The Daily* is added to a repository of all past issues. This growing set of individual issues functions as a data base in the sense that keyword searches can be executed against all past issues. As well, links from other pages on our Internet site can reference specific issues of *The Daily*. On the technical side: *The Daily* is produced every day as a fully encoded SGML (Standard Generalized Mark-up Language) document that is then rendered into a variety of dissemination and presentation formats. These formats include: HTML documents on the Web; print versions; e-mail messages to 1,500 subscribers of a *The Daily* listserver; a voice synthesized dial-up service for the visually impaired; an ASCII text file which some secondary distributors download to their own information distribution networks. 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.

## III. TIME SERIES DATA BASE CANSIM

9. CANSIM is Statistics Canada's online time series data base. All major socio-economic statistics are stored in great detail in CANSIM as time series with varying frequencies and length of series, some starting in 1914 (e.g.

Consumer Price Index, monthly). The data base 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. 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.

10. 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 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. Recently a dynamic graphic display option was added as well. 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 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, The State. In 1995, *Canadian Statistics* was launched with about 100 tables. The current number is 300 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 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 data base. 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 data base, 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 data described on the *International Monetary Fund's Dissemination Standards Bulletin Board (DSBB)*.

11. 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 data base.
- The data are released in a timely manner and are always current.

12. The *Canadian Statistics* tables have become the second most popular feature of our web site after *The Daily*. In creating these tables we took advantage of the intrinsic feature of Internet to offer links (hyper links) from each table to more detailed information, for example the specific CANSIM time series in the CANSIM data base in case that a client wishes to access the complete historical time series from which the table was derived.

#### IV. CANSIM AS THE DISSEMINATION DATA WAREHOUSE

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

14. **CANSIM II:** 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. (Exceptions to this are the data from the Housing and Population Census and the existing Export/Import commodity data base which continue having their own data base systems for the time being).

15. **Multi-dimensional table browsers:** These software tools have recently become available (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 data base while still preserving the easily absorbed presentation of statistics in flat tables or as a set of time series.

16. **Table creation software:** The content of most paper publications is tables. As all these data will be stored in CANSIM II and as most of these publications will be re-engineered to become electronic publications on Internet, there is the opportunity here to generate publication tables automatically from CANSIM II (as well as from CANSIM I already now). For example, the monthly publication on *New Motor Vehicle Sales* would be updated automatically as soon as the latest estimates have been added to CANSIM. Then at 8:30 am on the day of release of new motor vehicle sales in *The Daily*, the associated fully composed table publication would be available on the Internet containing the most recent details. The required software for this function is being developed in Statistics Canada based on SGML as the

structured language for marking up tables in HTML for Internet display.

17. **Custom publishing services:** Our recent experience is that sales of standard publications are falling steadily and that there is a growing demand for custom services. Data base publishing could 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 have already an electronic commerce interface on our site, the associated costs can be charged to, and paid by, the client conveniently.

#### V. CATALOGUE AND OTHER META DATA

18. Statistics Canada maintains and publishes two meta data bases which are and will be available on our Internet site.

19. The first meta data base is a comprehensive catalogue of all products and services offered by Statistics Canada. A record in the **IPS (Information on Products and Services)** pertains to a specific product or service and uses up to 60 fields to describe it in detail (e.g. catalogue number, author, abstract, subject key words, price, contact, etc). This data base with about 6,000 records in each official language (English and French) is maintained in an ORACLE DBMS on an internal file server. It is being updated continuously. Once a day, the latest changes to this data base are uploaded to our external Internet site and stored as HTML web pages (one page representing one record). Currently this set of HTML pages represents the IPS data base on our Internet site which can be searched directly by clients looking for information and which is also accessible through links from other information bins on our Internet site, e.g. *The Daily, Canadian Statistics*, CANSIM. IPS records also link to the process for ordering a product for electronic (i.e. downloading from our site) or physical delivery. In the future, we plan to store the actual IPS records in an SGML enabled data base on our Internet site and to generate individual HTML pages on the fly when an IPS record has been requested either through search or through hyper-link. The IPS catalogue has become the corner stone of our Internet site. Records in IPS are hyperlinked to and from other information bins. For example, *The Daily* lists the catalogue numbers of all publications and products most recently released as hyperlinks to IPS. IPS records display all the pertinent information about each publication or product and in turn hyperlink to functions such as downloading an electronic version or ordering a paper publication. All these hyperlinks are generated automatically as new information is added to our site.

20. The second meta data base 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 Data Base)**. 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 the fall of 1998, SDDS/IMDB will have been made available on our Internet site for direct access (browsing or word search) and through hyperlinks to and from the various other information bins using the SDDS number. For example, time series in CANSIM are referencing already the statistical program which is the source for a set of time series. Hyperlinks from the CANSIM data directory to the SDDS/IMDB records will allow clients to check the source of particular time series which they have selected for access and downloading.

21. Statistics Canada has a long standing policy of providing indications of data quality and underlying methodology for all published statistics. For paper publications this typically takes the form of a separate chapter within each publication issue or sometimes of a separate publication altogether. As the Internet has become the means of disseminating an increasing proportion of our data, we needed to develop a practical approach for satisfying this policy requirement for Internet disseminated statistics as well. Such quality indicator and underlying methodology information is to be added to each IMDB record. But such information, as mentioned above exists already as text in many publications. So it has been decided to create, at least in the short run, distinct documents on the Internet from the quality chapters of all publications which will be hyperlinked to the various "statistics" bin on our Internet site. For example, the release text in *The Daily* for a specific survey may have a concluding phrase such as "for further information on data quality, concepts and methodology click here" and a hyperlink will bring up the related SDDS/IMDB record and/or the document containing the quality indication information. Similar links will be provided from the *Canadian Statistics* and CANSIM as well as IPS catalogue records.

#### **VI. DOWNLOADABLE PUBLICATIONS**

22. 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 bin which in turn is part of the overall Internet data base. Similar to the *The Daily* bin of all past issues, this "publications bin" can be searched by keywords and hyperlinks can be used to link publications bin records to records in other bins on our Internet site.

#### **VII. ISSUES**

23. 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.

24. 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 meta data. 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.)

25. 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.

26. 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. [1][2]

27. 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 (e.g. 16 bit vs. 32 bit micro computers) 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.

#### VIII. CONCLUSION

28. Internet has started to change 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.

29. 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.

30. 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.

#### **REFERENCES**

[1] F.E. Hutton, W.M. Podehl: Statistical Systems to Support Analysis in Statistics Canada. Proceedings of the 49th Session of the International Statistical Institute, Florence, Italy, 1993

[2] M. Podehl, M Parisian: Effective Presentation of Statistics in Electronic Dissemination. Proceedings of the 51st Session of the International Statistical Institute, Istanbul, Turkey, 1997