STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE
EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint UNECE/Eurostat Work Session
on Statistical Metadata
(6 - 8 March 2002, Luxembourg)

Working Paper No. 14
English only

Topic (ii): Users and metadata, statistical information portals

THE STATISTICS CANADA PORTAL: EXPERIENCE AND LESSONS LEARNED

Submitted by Statistics Canada[1]

Invited paper

Abstract

Dissemination of official statistics via the Internet has entered a new stage. Effective Internet sites constantly evolve and change to meet client needs. 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 databases of text and numbers.

This paper reviews the current approaches of electronic data dissemination in general and the practical experiences in Statistics Canada and examines the driving forces behind the evolution of the structure and content of Statistics Canada's website

Keywords : Internet, database publishing, electronic dissemination, internet publishing, usability testing

I. INTRODUCTION

1. Statistics Canada had an online presence even before the advent of the Internet. A small group of clients had access to file transfer protocols (FTP) for large files, and *The Daily* (the agency's official release bulletin) was sent out through a text-based gopher service and listserves. When www.statcan.ca was launched in 1995—without any permanent funding—it was one of the very first Canadian government websites. Its major content was *The Daily*.

2. Six years later, the website has over 60 000 HTML pages and is one of the public's most important points of access to Statistics Canada, serving more than 15 000 clients each day. The website serves the general public interest by providing free data on Canada and Canadian society as well as information about the agency's products, services, surveys and statistical methods. It also provides an e-commerce gateway for paying clients to obtain selected products and services.

3. Statistics Canada's website was nominated a "Select site" by the editors of the Dow Jones Business Directory and was named the "Best Institutional Online Product" by the Canadian Online Products Awards. More than 10,000 other sites link to it. The Internet allows Statistics Canada to reach more people than ever before and to increase the effectiveness and efficiency of its publishing/ dissemination programme. 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

[1] Prepared by Louis Boucher.

through what is referred to as database publishing, i.e. the process of generating and updating the content of our Internet site from databases. In this regard, our existing CANSIM output database plays a crucial role.

4. However, client expectations increase and like other websites, www.statcan.ca is constantly evolving to meet its users' needs and expectations. The site's evolution has been driven by users of all kinds, including the National Statistical Council (the Chief Statistician of Canada's advisory body), our partners in the Data Liberation Initiative[2], the Depository Service Program[3] and the areas within Statistics Canada that produce the data and author the publications—and, most importantly, the general public.

5. We have an extensive market research program to assess the nature of our clients, their interactions with the site and their expectations. We do periodic market research using pop-up questionnaires on the entire site as well as specific segments of it. We conduct usability testing and observational research, and monitor e-mail and traffic logs. We work closely with the staff in our regional offices to make the website work for them and for the clients to whom they provide services.

6. 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 overview of our website development and our own experience with market research and user acceptance testing.

II. PART I. INTERNET AS A DISSEMINATION CHANNEL

7. 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.

8. Statistics Canada has always been concerned with delivering quality service in a cost-effective manner. The Internet provided us with the opportunity to curb the costs of our publishing program while at the same time enabling us to deliver our publications to clients in a timelier fashion.

9. 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.

10. 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, the costs of maintaining and updating individual HTML (HyperText Mark-up Language) pages manually

[2] The Data Liberation Initiative consists of a community of researchers in Canadian colleges and universities that access public use microdata files and macrodatabases through the Internet site.

[3] The Depository Service Program consists of a network of over 700 academic and public libraries across Canada and throughout the world that receive Government publications in order to ensure that they are accessible by all Canadians.

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

11. 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 impacting necessarily on the other.

12. 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 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 database publishing methods are used to make them accessible and to inter-link them on our Internet site.

The Daily

13. Statistics Canada releases new data and statistical products every day through *The Daily*. Because *The Daily* was the website's first content when it was launched, we began our Internet venture with a website that had to be updated at 8:30 a.m., without fail, every working day. *The Daily's* users—journalists, news agencies, policymakers, bankers, industrialists, consultants and the general public—needed to be assured that they would receive *The Daily* reliably and predictably. *The Daily* became the main feature of our Web site is. *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.

14. *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. 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. Database 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

15. CANSIM is Statistics Canada's online time series database and corporate data warehouse. Since 1973 and until 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 service:

16. 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

17. 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 700 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.

18. 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.

19. This update process of the Canadian Statistics tables 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.

20. 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 along with metadata.

Thousands of hits on Canada's thousands of communities

21. We needed an application that would dynamically generate tables of information on population characteristics, work, families and dwellings for each community across the country. This application became *Community Profiles* and was added to the site in November 1998. All a user had to do was type in the name of their hometown and seven tables of data would be generated for that city, town or community.

22. *Community Profiles* was an instant success, attracting roughly 20% of traffic on the site within two weeks. The number of visitors to the entire site jumped from roughly 4,000 to 7,000 visitors a day.

Catalogue and other Metadata

23. Statistics Canada maintains and publishes a comprehensive catalogue of all products and services. 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.

24. Comprehensive description of concepts, definitions, subjects, variables, methodologies and quality indicators about our statistical programs is also available. 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

25. 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 database publishing. But if one regards the total Internet site as a sort of structured "database" then each publication issue can be regarded as a "record" within the publication module which in turn is part of the overall Internet database. 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.

26. The issues related to database publishing and electronic dissemination are:

?? Electronic dissemination enables 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 centre of this process and great attention needs to be spent there.

?? Database publishing requires expert resources for the one time development of the necessary databases, 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 database publishing process needs to be evaluated for each case. On the other hand, once a database exists, new opportunities can be exploited which are not feasible without such a database.

?? Stringent data quality procedures have to be instituted to verify the accuracy of the information before it is entered into the database. This applies both to data and metadata. There has to be absolute confidence that the data in the database are "correct" and that automatic database 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 database, all future presentations extracted from the database 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 database publishing: if the database is down, nothing can be published.

?? The interface between extracting data from a database 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 database 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 database 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.

III. PART 2. THE USERS: A PROFILE OVER TIME

27. While the foundations of Statistics Canada's Internet presence were laid between 1995 and 1998, by 1999 it became evident that the site needed to be more than just a collection of documents placed on the Internet in order to meet the Agency's needs. Some benchmarking market research had been conducted in 1997 that profiled our clientele. We conducted a second round in 199 to update our users' profile, assess our clients' needs and behaviours and to find out how satisfied they were with the site.

28. The findings from these quantitative studies revealed that there had been a significant increase in the proportion of students visiting the site, up to 40% from 22% in 1997. The proportion of frequent users to the site had also increased in this two-year period by 10% to 34%. In fact, many respondents (29%) had bookmarked the site.

A satisfactory experience

29. Surfing Statistics Canada's website was overall an agreeable experience for our visitors. More than two-thirds of the respondents indicated that they were very satisfied or satisfied with the site. Most said that they would visit the site again (87%) or recommend it to others (82%). They rated the site well on the consistency of design, said the language was clear and easy to understand, and that there were no delays in using the site.

30. However, when we looked in more detail, we found that although we performed well in these areas, they were not necessarily the most important aspects of the site for our clients. Quadrant analysis revealed that the site actually had a low rating in a number of areas that were very important to our users, notably:

?? the range and scope of information was too narrow;
?? the information was not easy to find;
?? the search engine did not return effective results;
?? there was not enough free information; and
?? the home page did not direct users to the information that they needed.

31. Users had told us what they thought; now we had to make the improvements. An action plan was drawn up that addressed the recommendations from the research. The plan focused on content development i.e. increase of the range and scope of free information, searching, navigation and design.

Increasing public good content

32. The project to increase the range and scope of free information on the site was tackled on three fronts. We encouraged authors to put their national- and provincial-level data into *Canadian Statistics* module of the site. *Community Profiles* were expanded to include more than just census data. The number of births and deaths in each community was added along with data for over 130 health regions across the country. Eventually we hope to incorporate education and justice data in this module. There was also a move to make older versions of publications available on the site for free.

33. Of the three initiatives above, the idea of making older versions of publications available for free was perhaps the easiest to grasp and the hardest to implement. Author areas were very supportive of the initiative—they wanted their publications to be widely accessible—but there a was a need to establish a policy that determined how "old" a publication had to be before it could be free. We did not want the

publication sales to decline and we needed a standard that could be applied in all cases. There was also a technical limitation that is now being addressed.

Searching successfully

34. With upwards of 60 000 pages on the site, searching and retrieving accurate results is critical. When the market research was conducted in 1999, searching on the site was fragmented. *The Daily, Canadian Statistics* and the databases (the Online Catalogue, CANSIM, International Merchandise Trade database) were being searched using one search engine. *Community Profiles* used another. In fact, this application was running on a different server and could not be found using a site search. The remaining areas of the site (Education Resources, Concepts, definitions and methods, Are you in a Statistics Canada survey etc.) were searched using a free search engine that was a legacy from the site's early days.

35. One of the first challenges was to integrate searching for all free content on the site—including the Community Profiles. We also needed a search engine that could search metatags in the HTML code. A new search engine was installed on the site in February 2001 that did all this. However, we had to exclude CANSIM and the International Merchandise Trade databases which use separate search engines.

36. Though we have made great strides in improving the search features on our site, clients are still having difficulty finding information. In the summer of 2000, we conducted research on our users' information needs and their success rate. Only 42% of the clients found all or most of the information they sought.

37. Although this survey was conducted prior to the installation of the search engine, it revealed that the difficulty that clients had finding information on the site went beyond search engines and search results. There were three facets to this problem:

?? the information was gathered by Statistics Canada but was not available on the site
?? the data was not gathered by Statistics Canada
?? the data was on the site but could not be found using the client's terminology.

38. People will come to the Statistics Canada website to find out how to get a copy of their birth certificate or to get a weather forecast. We needed to develop a strategy to manage the expectations of our site visitors—this is the scope of the information that we do have and this is what we don't have.

39. We have added two modules on the site that outline the scope of information that we do have. A "First visit" section, which describes in narrative format the type of information on the site, and an A to Z index—essentially an alphabetical site map with links to the top three layers of information on the site.

40. We added links to information that is frequently sought on our site but is the responsibility of other arms of government. For example, we have added a link to the portion of the Government of Canada website where people can find out how to get a copy of their birth certificate—a service under the jurisdiction of each province.

Integrating content

41. One of the problems that both clients and Statistics Canada staff had with the site was that the information was fragmented. There would be a news release in *The Daily*, a publication under Products and services, data in Canadian Statistics and metadata in Statistical methods. This was frustrating for both authors and users because they often found only one small piece of the puzzle.

42. We are working on integrating the various elements of the site so that the puzzle is virtually complete: the user has only to find one piece to be linked to the rest. For example, a major release has a short summary on the home page that is linked to the full article in *The Daily*. *The Daily* article contains links to the publication it describes, which contains a full analysis of the data and a full set of charts and tables. *The Daily* article has its own tables and charts, but it also links to summary tables in Canadian

Statistics, which are linked to the metadata for the survey. Analysts who want to analyze the full data set can follow the links to the relevant CANSIM tables and matrices given at the bottom of *The Daily* article or the tables in Canadian Statistics. Our goal is to automate as many of these links as possible so that little manual intervention is required to create and maintain the inter-relationships.

Redesigning the site to meet users' needs and government standards

43. Two things drove us to redesign our site: the results of the market research and the requirement that we implement new Government of Canada standards for the Common Look and Feel for Internet. The standards were developed to ensure that information is consistently presented on the government's many departments' and agencies' websites.

44. The new Government of Canada menu bar changed our site's navigational paradigm, shifting the primary navigation tool from the sidebar to the top menu bar. It was a shift that enabled us to move some information up one or more levels—and respond much better to our users' needs.

45. Our first objective in redesigning our home page was to make it easier to locate information, so we modified it to include direct access to data and the latest releases without forcing the client to click through several levels of navigational pages. It was the users who suggested that we include an article or a feature from *The Daily*. Now brief, hyperlinked, summaries of the major releases in *The Daily* and a list of all the other releases and new publications appear on the home page. So in fact, Statistics Canada's website has a new home page every day.

46. Finally, we had a prototype home page that met with the agency's approval and conducted usability testing during which we watched the interactions of 10 site users (5 English, 5 French) as they completed a series of information retrievalexercises. We also consulted the staff in the Regional offices because they are on the front lines and often assist clients over the phone who are having difficulty finding information on the site. The feedback from both these exercises resulted in a few small changes to the prototype—and on April 19, 2001, we launched it.

IV. INTERNET CHALLENGES FOR STATISTICS CANADA

Writing for the Web

47. Statistics Canada's employees have been authoring paper publications for decades, but now that the Internet is becoming the agency's primary dissemination vehicle, they are having to learn how to write effectively for users who will no longer flip pages, but scan, scroll and click. It is not an easy change to make, because it requires all our authors, agency-wide, to think about and structure information in a completely different way. It is a major undertaking, but one which the agency is supporting through a course on Writing for the Web.

48. The Marketing and Information Services Branch began its "Writing for the Web" workshops in 1999. The course shows Statistics Canada authors how users interact with the Web and how to write for them while maximising the opportunities the Web offers and working within the limitations the Web imposes. So far, nine workshops have been conducted with a total of 94 participants.

49. The 94 authors that have taken the course so far are just the tip of the iceberg. We have a unit that provides consulting services to the other authors who still need help making their publications "web-friendly". In addition to this, we are finding that our authors are increasingly Internet users themselves and so are learning through their own experiences on the Web.

Online data collection

50. One of the biggest challenges for Statistics Canada in the next years is developing methods that allow clients to respond to our statistical surveys online. These applications must be intuitive and flexible enough to suit the "lowest-common denominator" of our respondents' computer systems, and ensure that

the confidentiality of the information provided to the Agency is secured through leading-edge encryption technologies. This is an area with a steep learning curve and that will be supported by user testing.

New technologies

51. Government departments and agencies will also have to be prepared to deliver information that is accessible in multiple electronic formats. Canadians are now using wireless devices and televisions to access online information. Providing content for these devices with as little re-working as possible (undoubtedly applying the "automate!" principle) will be a burgeoning area of activity in the coming years. The challenge of the Government On-Line strategy is ensuring that all development is done in consultation with clients to ensure that it responds to their needs for online service delivery.

V. CONCLUSION

52. 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.

53. 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.

54. 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.

55. We built our site on our own assumptions about what users wanted, and we modified it on the basis of their reactions—in other words, when they complained, we fixed it. We believed we would never get it right the first time, so we might as well put the draft on line. What we were actually doing was user-testing live. Our early market research taught us a hard lesson: we cannot trust our assumptions about what users want and our predictions about how they will interact with our website.

56. Now we test at the earliest stages of our projects. We listen to users' recommendations, build prototypes, conduct usability testing or observational research, re-build the prototypes and consult with our users again. We sit with our staff in the regional offices to find out what kind of problems they face on the website every day as they help their clients. It all helps us get it closer to right before we go online, and helps us get better as we go along.

57. We have to remember that pre-launch user-testing still only gets it almost right. In the final analysis, our users are our greatest resource and their feedback through e-mail and formal market research is invaluable.

Acknowledgement

I would like to thank Pamela Best, former manager of the Internet Content Development Team in Statistics Canada for her leadership and commitment in making Statistics Canada's web site such a success. She has also produced many high quality discussion papers on website development and usability testing. Her work has greatly influenced this paper.