WP. 20
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Dissemination and Communication
(12 to 14 September 2006, Washington D.C., United States of America)

Topic (iii) How to present metadata

KEEPING WWW.STATCAN.CA USERS IN THE METADATA LOOP

Invited Paper

Submitted by Statistics Canada, Canada[1]

I. INTRODUCTION

1. We know that our website users represent diverse constituencies, and that their information needs are equally diverse. While the content on our website continues to expand as new information becomes available, the challenge articulated by user feedback is to organize and link this growing information store in a way that helps users find the information they seek. Data are only useful as long as you can find them and understand them! Dumping all information in a heap in the user's lap is not helpful, and only leads to 'infobesity'. Statistical organizations must make an extra effort to organize and present our information so that users can find the most comprehensive information when they visit our websites.

2. When our website users find specific data, they often require other related information from Statistics Canada. They often need to know more about its availability, its applicability or its related context. Such information is often called 'metadata': some of it is found in 'meta-tags' encoded in web pages, but at Statistics Canada much of it can be viewed by users, and resides in various databases. However, our users require some, if not extensive, knowledge of these resources. The data in meta-tags helps search engines provide raw results, but the 'other' type of metadata provides the context users need to understand and make best use of the data. Users should be able to access our metadata easily and intuitively. To facilitate this, users have direct access to metadata via strategically placed links throughout our website.

3. This paper is not about metadata itself. Rather, I will attempt to demonstrate that users can access complex sets of metadata—the context around the data they want—without actively looking for them. We will look at Statistics Canada's approach to metadata and taxonomy, our user's needs and the solutions proposed on www.statcan.ca.

II. METADATA + TAXONOMY = ENRICHED CONTENT

4. The basic purpose of taxonomy is to define content in categories that can be organized and accessed efficiently by using a predefined set of common vocabulary. Our taxonomy is based on

[1] Prepared by Louis Boucher.

themes and sub-themes that organize our content into manageable subjects so they can be found by search or navigation. Such taxonomy enables us to relate data and publications from different sources; avoid confusion caused by synonyms; give the data a hierarchical structure; help users understand the relationships between data; and provide comprehensive navigation across our website.

5. Statistical metadata—the information about statistical data and processes—enhance search and understanding of data for users; for us, statistical metadata improve and automate survey processing and facilitate data harmonization. However, metadata is not an absolute concept. Individual pieces of information, mostly data, become metadata when they are put into a descriptive relationship with something else. Simply put, a database of publication descriptors (price, media, author, etc,) becomes the metadata of the publications catalogue itself. At Statistics Canada, we have two metadata sources.

6. The first one is a comprehensive online catalogue of all products and services offered. Its underlying ORACLE database, called the Corporate Database of Products and Services, or CDPS, contains up to 60 fields to describe each publication and service available. Once a day, the latest changes are uploaded to our website. The CDPS can be searched or browsed by users.

7. The second metadata source is a comprehensive description of concepts, definitions, subjects, variables, methodologies and quality indicators about our statistical programs. Known as the IMDB (Integrated Meta Data Base), it is the second generation of a database of records that pertain to each statistical program such as surveys, administrative data acquisitions or the census. The IMDB also covers derived programs such as national accounts. As in the CDPS, each record has a unique identification number and up to 120 fields of meta-information about the source program. Like the CDPS, the IMDB is updated regularly.

III. WHAT ABOUT META-TAGGING?

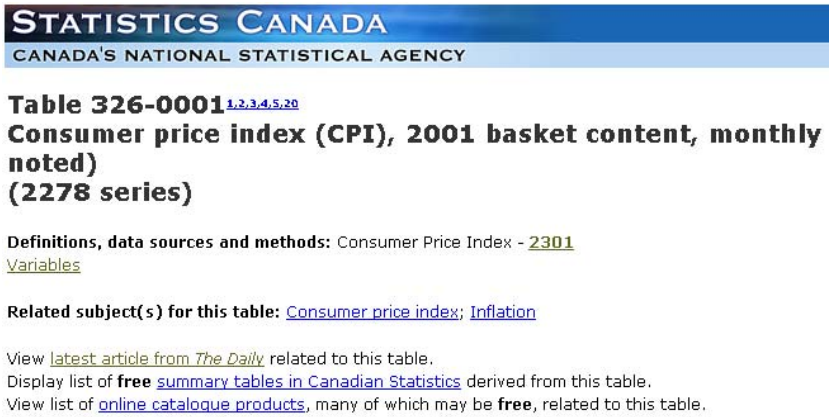
8. Metatags—those hidden attributes encoded in the header of web page—describe precisely the characteristics and content of any web document. In theory, it sounds obvious that every document should be properly tagged. In practice, whose responsibility is it to do such daunting work? It is somewhat complicated, technical and constraining after all. Should the author determine the proper metadata and encode it in the source document? Is this the responsibility of the 'webmaster' or the catalogue librarian? We are currently addressing this issue by designing a framework of roles, responsibilities and processes with regards to metadata, taxonomy and meta-tagging.

9. Using a standard taxonomy across the Agency, both in our meta-tags and in our statistical metadata, is essential for managing and retrieving our web-based content. There has been an effort recently within the organization to assign and maintain proper metadata as well as updating our taxonomy over time. With a site as large as www.statcan.ca, it is not enough to have either metadata or taxonomy. Metadata without controlled vocabulary results in authors using different words to describe the same topic (e.g., teachers versus professors versus educators). Using just taxonomy organizes data for management of content but not for retrieval of content. Used together, both concepts render some consistency of data across different databases and information sources, which simplifies its management and retrieval. To simplify its fundamental application, authors are provided with technical documentation on how to meta-tag their documents at the source. It is just a matter of time before all our website content is properly tagged.

IV. SHOULD USERS BE FAMILIAR WITH OUR OWN METADATA?

10. Simply put, they need to know as little as possible about it. We need to make those technicalities completely transparent to users. When users reach our website, they have an objective in mind, be it the latest population count or a specific study. They want this information from a trusted source. It is our responsibility to document and arrange our growing data holdings to enable most users to retrieve what they seek with the minimum effort and the least possible knowledge about our internal processes.

11. Assuming that the metadata and taxonomy framework discussed above is properly deployed a priori, the webmaster's task is then to develop a web interface that will assist users retrieve the information they seek. There are 'classical' proven approaches such as search engines, thematic lists and site maps that help users find their way around massive information holdings like those of statistical organizations. The approach we have chosen is to proactively present all contextual information to the users during their visit to our website. This enables us to make reference to our metadata in one comprehensive interface The following example illustrates this strategy.


STATISTICS CANADA

CANADA'S NATIONAL STATISTICAL AGENCY

Table 326-0001 1,2,3,4,5,20

Consumer price index (CPI), 2001 basket content, monthly noted)

(2278 series)

Definitions, data sources and methods: Consumer Price Index - 2301

Variables

Related subject(s) for this table: Consumer price index; Inflation

View latest article from *The Daily* related to this table.

Display list of **free** summary tables in Canadian Statistics derived from this table.

View list of online catalogue products, many of which may be **free**, related to this table.

12. In this scenario, the user has selected a detailed data table from our commercial CANSIM output database and is about to download it. (CANSIM, or Canadian Socio-economic Information Management System, offers users dynamically generated tables of much of our current and historical social and economic data.) At this stage, we propose that users consult supporting documentation related to this data table. All this documentation is available from different sources within the website; it can also be accessed in isolation from the different modules, assuming that the user is familiar with their original location. However, most users would not know they exist, nor would they be willing to spend more time looking for that documentation during their visit. This is where a little bit of proactive thinking and presentation creativity can transform rigorous metadata into intuitive documentation to users.

13. By clicking the *definitions, data sources and methods* link, users have access to the IMDB, the exhaustive metadata repository of the survey related to the selected data table. The *related subject for this table* link enables users to directly access our taxonomy and thesaurus. The *view list of online catalogue* link offers direct access to the CDPS, the products and services metadata. The two other links also lead to related information for this table. By presenting our metadata in a simple descriptive approach, we maximize the possibility that users will consult it. Users can access our various metadata sources without having to be familiar with metadata technical and conceptual framework. We call this 'metadata at work'.

14. Even though this seems simplistic, we had to create a small dedicated database called the Common Object Repository (COR) to support these links. One can imagine each link to and from each metadata source could be hard-coded in each source. However, this would be costly to maintain and difficult to keep current. What COR does is link every metadata source via the

unique survey identifier from IMDB. The assumption is simple: every data holding, publication and study has to be generated by a survey conducted by Statistics Canada. For the example illustrated above, the CANSIM table is related to one survey where the COR metadata for this survey directs the user toward the appropriate taxonomy and the catalogue items. This process is fully automated.

V. TOO MUCH INFORMATION LEADS TO "INFOBESITY"

15. Users' expectations are high and, like other websites, www.statcan.ca is constantly evolving to meet users' needs and expectations. This does not mean inundating them with more information than they need. Our task is to present results in an orderly manner, enabling users to then make their own choices. Some users prefer to launch a search and peruse the results list; others will browse the subject list. It is important to better understand users' behaviours in order to present our data holdings in a comprehensive and useful manner. The challenge is to find the right balance that suits most users.

16. We have an extensive market research program to assess the characteristics 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 works for them and for their clients.

17. One of the problems that both users and Statistics Canada staff had with the website was that the information was fragmented and voluminous. One solution we found is integrating the various elements of the site so that the puzzle is virtually complete. The starting point can be *The Daily*: every *Daily* release has a short summary on the www.statcan.ca home page that is linked to the full article in *The Daily*. The article contains links to the publication it describes; the publication in turn contains a full analysis of the data and a full set of charts and tables. *The Daily* article has its own tables and charts, and it also links to related metadata. Analysts who want to analyze the full dataset can follow the links to the relevant CANSIM tables given at the bottom of *The Daily* article or the survey metadata from IMDB. Our goal is to automate as many of these links as possible so that little manual intervention is required to create and maintain the relationships between the pieces of metadata. The following example illustrates such integration. Via *The Daily*, users have full access to our metadata and more.

Latest release from the Consumer Price Index

Friday, July 21, 2006
Released at 7:00 a.m. Eastern time in The Daily

PDF
Troubleshooting PDFs

June 2006

Previous release

Canadians paid 2.5% more for the goods and services in the Consumer Price Index (CPI) basket in June 2006 than they did a year earlier. This was slower than the 12-month change of 2.8% in May.

Available on CANSIM: tables 326-0001, 326-0002, 326-0009, 326-0012 and 326-0016 to 326-0018.

Definitions, data sources and methods: survey number 2301.

More information about the concepts and use of the CPI are also available online in *Your Guide to the Consumer Price Index* (62-557-XIB, free).

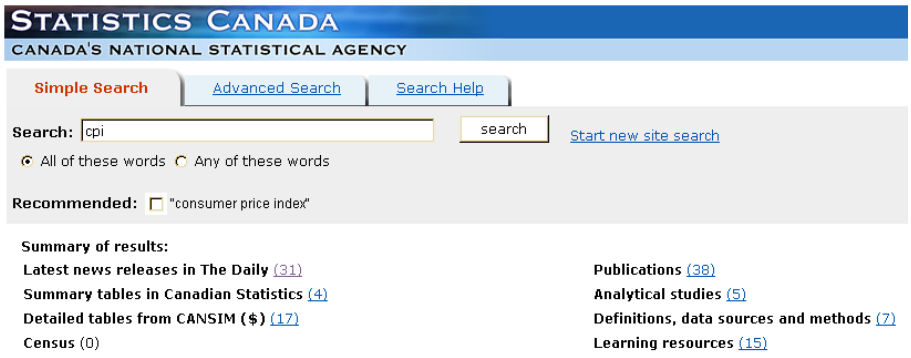
Available at 7 a.m. online under *Today's news releases from The Daily*, then *Latest Consumer Price Index*.

The June 2006 issue of the *Consumer Price Index*, Vol. 85, no. 6 (62-001-XIB, free) is now available from the *Our Products and Services* page of our website. A paper copy (62-001-XPB, $12/$111) is also available.

The July Consumer Price Index will be released on August 22.

For more information, or to enquire about the concepts, methods or data quality of this release, call Client Services Unit (toll-free 1-866-230-2248; 613-951-9606; fax 613-951-1539; *infounit@statcan.ca*), Prices Division.

18. For search, we used to present long list of results organized by date or ranking. This forced user to browse through extensive listings without any context. We came up with the idea of pre-arranging search results on a 'scorecard' as illustrated below.



19. The search results for "Consumer Price Index" lead to 158 results that would have been originally listed. Instead, we arranged the summary of results around metadata to facilitate selection by users. If users are looking for a particular study about the Consumer Price Index, it they will find it easily by clicking *Analytical studies*—in the right-hand column of the search results listing—and selecting it from the five items there. The process is the same for accessing the detailed CPI tables from CANSIM. This is another example of putting the metadata to work for user's benefit.

VI. CONCLUSION

20. In this paper, we have tried to demystify metadata by showing how a statistical organization like Statistics Canada can make it accessible to its users in a practical and transparent manner. There many other ways to do so. A better understanding of users' needs supported by a rigorous internal framework about metadata, taxonomy and meta-tagging will enable us to provide the most efficient interface for our users.

21. A live demonstration of www.statcan.ca will illustrate the approaches discussed in this paper.