Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

DELIVERING THE METADATA: AN RDC EXPERIENCE

Irene Wong[1]

ABSTRACT

The metadata associated with microdata production of major Statistics Canada population-based surveys are often voluminous. To reduce administrative burden while maximizing the potential use of the metadata, a harmonized process based on standards for storing, capturing and disseminating survey metadata is required. A systematic approach both in terms of metadata content and method of dissemination of microdata files across all surveys is desirable. This paper summarizes the challenges in creating the metadata of General Social Survey cycle 17 based on the Data Documentation Initiative (DDI) standard, the comprehensiveness of DDI elements in capturing the metadata, and how useful this format is subsequently to Research Data Centre (RDC) staff and researchers.

KEY WORDS: Metadata standard, Metadata dissemination, DDI

1. INTRODUCTION

The creation of a questionnaire, descriptions of the survey procedures, system parameters of a computer assisted interview, definitions of the data variables, characteristics of the collected statistics etc., are all part of the metadata of a survey data file. Metadata creation begins when a data gap is identified and someone decides to collect the missing information using a survey. The flow of creation will continue even after the questionnaire responses become meaningful statistics.

The metadata associated with microdata production of major Statistics Canada population-based surveys are often voluminous. The decision on what metadata to keep and to whom they will be distributed is often made by the individual survey managers. In practice, the contents and the distribution of those metadata may vary even within a national statistical office like Statistics Canada, and the decisions on metadata capture and dissemination depends on a good understanding of different clients' needs, but to establish what users need can be difficult.

The social science community has long recognised the importance of better communication between the data users and producers. The Data Document Initiative (DDI) is a recent international effort to improve communication in terms of better capture, structure, presentation and delivery of metadata, and to establish a standard for social science metadata that is recognized and agreed upon by the social science community. This community includes institutions in North America and Europe representing many of the largest data producers and data archives.

2. PROVIDING CLIENT ORIENTATED SERVICES

Statistics Canada, as the national statistical office of Canada, has been consulting with end users over the years and providing services, either for free or on a user fee basis, and is guided by several principles to address the needs of the client (Podehl, 2004). Survey documentation must be informative about the data collection and its design besides ensuring the data quality and mutual consistency of data outputs from different collections.

[1] Irene Wong, Research Data Centre, Statistics Canada, Rutherford Library, University of Alberta, Edmonton, Canada, T6G 2J4. irene.wong@statcan.ca

Statistics Canada's Integrated Meta Data Base (IMDB) has been working with international statistical agencies to establish definitional metadata standards, and has created the Framework for Subject Matter Standards (Mechanda et al., 2003) to help survey managers capture their surveys' metadata consistently within the bureau. IMDB also provides the public with survey descriptions and summaries that can be accessed on the Internet for free, but to avoid overloading the web site with the too many details that may potentially confuse some users, only the most popular information are posted. Metadata on data characteristics and variables descriptions are usually provided in separate documents, which are formatted and disseminated at the discretion of each data collection manager.

Each survey collection program usually has a call-centre supported by survey methodologists and subject matter experts to handle inquiries surrounding the clarification of their data collections, including questions on data quality, control and measurement of coverage, sampling, non-response errors, sample weighting strategies, etc., as well as to provide general advisory services of data access and purchase. However, frequently asked questions are not always restricted to a single survey. Some of most frequently asked questions from the clients (ranging from students to senior researchers) are data discovery related - "Which survey do I need to answer this research question?" and "In which survey can I find those variables?".

2.1 Limitation in data discovery

For a researcher, finding the data collection(s) best suited to address his/her research questions is essential in any research project. For Statistics Canada to provide comprehensiveness and completeness of metadata for any data product is very important to researchers, but for data discovery, having expedient access to the metadata from many surveys is just as critical.

With easy access to the Internet, researchers are increasingly relying on on-line searching on the Statistics Canada web site for data files. Despite high quality of our data and sometimes voluminous related documents, microdata users find that the tools for navigating metadata beyond IMDB for the population-based surveys are inadequate and lack built-in data extraction systems to support any form of preliminary analysis. Survey searches using keywords or themes may not necessarily lead searchers anywhere, as many demographic and household characteristic variables are common to almost all household surveys and many non-sample databases. If an off-the-mark keyword is used, potential useable surveys may not all be identified by the search engine. Even if documents are found, very often information and data extraction require different application software than might be available to potential users.

Search and navigation issues seem to be acknowledged by both the data clients and Statistics Canada, as noted in the concise report by Podehl in 2004 on *Service improvements in Statistics Canada*[2]. From the results of the *Statistics Canada Website Survey* in 2002, combined with metrics on visits, it was found that students represented the largest category of visitors (about 30%), with the "economists and social scientists" group second with about 10%. *Population and demography* remained the most popular topic of interest, and the search function remained the most popular entry to Statistics Canada's site. It also reported that, while search and navigation functions had improved, they were still rated as the area requiring further improvements. For example, there should be hyperlinks to analysis of the data,and, conversely analysis should be linked to the data behind the analysis.

2.2 In the context of the Research Data Centre Program

Today new technology has led to more sophisticated users with higher expectations regarding data quality, coverage and access. To help strengthen Canada's social policy research using Statistics Canada microdata, the Research Data Centre (RDC) program was formed in 2000 in partnership with the Social Science Humanities Research Council and university consortia across Canada. The program is to improve access to the rich data resources of Statistics Canada in Canada's postsecondary institutions, supporting research and teaching in the social statistics field.

---

[2] Podehl, W. Martin (2004), "Service improvements in Statistics Canada", *Statistical Journal of the United Nations*, ECE 21 (2004) 1-6, , IOS Press. Page 4-5.

The centres essentially are secure computer labs set up in university campuses across Canada operated by one to two Statistics Canada staff, who act as data support staff for local researchers and as local operational managers of their own centres. Each centre acts as a repository of Statistics Canada population-based and household-based restricted (confidential) microdata files, accessible to researchers who satisfy the security provisions with the approved research projects.

Microdata clients come to RDCs as well as other Statistics Canada contact points for data advisory and data discovery consultation services. Their dependence on Statistics Canada staff becomes evident in the context of RDC because they will not have access to the microdata prior to submitting their research proposals.

Since the RDC is only the repository of survey data collected by survey divisions, metadata dissemination decisions lay with different survey production units; and each of the units may decide to disseminate their collections using different metadata structures and formats. Helping researchers in data discovery really becomes a burden when the number of complex surveys housed in the RDC has gone from four core longitudinal surveys to over fifty surveys in less than five years, and the number of researchers who want to learn to use those surveys and demand RDC services has also noticeably increased. Even more of a challenge is keeping track of multiple versions of the metadata for different survey cycles or multiple revisions within a survey. Helping researchers in data discovery has become an operational issue for RDC staff, when their other duties have also expanded and they are no longer able to keep up with all of the survey documents.

There is a need to decrease the users' dependence on RDC staff in their data search. There is a need for something that will enable a new data user to gain a full understanding of the data collection with minimal consultations from the producers. One solution is to improve the way metadata is organized and delivered, and have survey units better coordinate their survey documents to allow researchers to do most of their data discovery easily using the Internet.

3. The RDC DDI METADATA PROJECT

Ideally, a harmonized storing, sorting, capturing and disseminating metadata process based on standards will allow survey divisions to maintain and update their metadata easily, to minimize the work of capturing metadata from the disseminator, and to facilitate searching across surveys quickly by end users. As discussed by Colledge (1999), integration provides an effective mechanism for addressing the challenge faced by statistical agencies. The solutions seems to be simple, and the general approach does not seem hard to follow, but the review and rationalization of all the information about the collections, their data items, the classifications and more, is time-consuming and conscientious work. This may not be a welcome responsibility in addition to the regular activities of the data production staff.

3.1 Project outline

Since Statistics Canada's IMDB follows the ISO/IEC 11179 model, there is equivalency of terms used among Statistics Canada surveys, and equivalency of quality guidelines[3] and assurance framework[4]. A good degree of coherence exists among the metadata of different data collections within the same subject area, what is really missing is consistency in how metadata are organized, formatted, structured and how much detail is provided. Given that RDC major clients are social science researchers using population-based micro survey data, a pilot project was proposed in 2004 to present the metadata of an analytical file housed by the RDC to comply with the Data Documentation Initiative (DDI)[5] standard.

The objective of the project was to identify the survey document information available to a RDC researcher, to assess what was needed to use DDI as the metadata structure, to create a DDI product and evaluate the use of DDI

[3] "Statistics Canada quality guidelines", Statistics Canada (2003), Catalogue number 12-539-XIE, fourth edition.
[4] "Statistics Canada's Quality Assurance Framework", Statistics Canada (2002), Catalogue number 12-586-XIE2002001.
[5] Details about DDI specification, its alliance and other information about the initiative can be found on-line, among many, information may be found in a Canadian site at http://www.icpsr.umich.edu/DDI/index.html.

compliant metadata in the RDC. Thanks to the support of the General Social Surveys (GSS) program, the GSS cycle 17 (Social Engagement) analytical microdata file and its documentation were provided for the project.

3.2 Summary of the creation of DDI compliant metadata product for the GSS Cycle 17

The University of Alberta RDC had taken on the task to identify and gather the survey information source, and to create the GSS 17 metadata compliant with DDI version 2 using *NESSTAR*[6] publisher.

DDI version 2 consists of five major sections and approximately 300 sub-elements called tags. *NESSTAR* publisher closely follows the DDI standard. Four of the sections are purely for metadata, and one section contains the microdata. The major sections are namely:

1) **Document Description** is essentially the "header" and citation information about the makeup of DDI products which may contain more than one survey and survey cycle;
2) **Study Description** describes the data collection at a broad level and includes information on geographic and temporal scope as well as methodological information;
3) **Files Description,** that is, description of the physical data file(s) in terms of record and variable counts, logical record length, etc.
4) **Data element (Variables) Description** presents detailed information on each data item including elements such as question text, variable label, category labels and values, etc;
5) **Other Related Materials** that is, documents or other information related to the collection, that is left for the publisher's discretion to include.

Most of the known DDI documentation activities in Canada were based on "post-event" information harvesting and it was no exception for this project. "Post-event" harvesting basically means that metadata is gathered from survey documents and information that had been disseminated for end users in one form or another. In other words, there had been no simultaneous capture of metadata during the survey design stage, interviewing stage, data editing stage or at any other stage of survey data production. Most input materials of this DDI project were recycled documents from IMDB's on-line survey description, and the GSS dissemination unit's microdata accompanied user guide, questionnaire, data dictionary, and data record layout. They provide good deal of information concerning the methodology used of the data collection and creation of the variables. Those articles were in the format of text (*Microsoft Word*) and image (*Adobe PDF*); the microdata file came in the form of a flat file[7].

Checking for completeness of the survey documents was the first step of creating the DDI metadata. For this particular survey, some modification and adjustments were needed for the data element and variable descriptions. These included the re-creation of the variable labels, definition of missing data values for each variable, verification of the character vs. numeric attribute of the variables and to define each of their measures as being ordinal, nominal or scale. Other value-added activities included grouping variables by theme and selecting summary statistics according to the characteristic of each variable. These modifications and value-added activities were necessary in order to create meaningful summary statistics, which would be included in the variable description section in the DDI metadata for the benefit of end users better understanding the coding and values distribution of the variables. For example, the statistical mean and standard deviation were included for continuous variables while frequencies and proportions were included for categorical variables.

After organizing the data documentation, the information was manually entered into the *NESSTAR* publisher application corresponding to different DDI elements using its interface, and it automatically transcribed those

---

[6] Nesstar Ltd. is a wholly owned subsidiary of the UK Data Archive and the Norwegian Social Science Data Services. License for this project is provided by Nesstar Americas Inc., a subsidiary of Sharbot.com Inc., formed in 2003 through a relationship agreement with Nesstar Ltd. of Colchester, Essex, U.K. to deliver and support the Nesstar suite of products to the North American marketplace.

[7] A flat file is usually used in conjunction with the data record layout file; it usually contains a continuous long string of mixture of integers and/or characters and the data record layout identifies the position of the beginning and the end of each variable as well as each data record on the string.

statements into XML[8] code. One could also modifiy directly the XML programming code to customize the output to fit particular needs such as data suppression. If there is small revision to metadata content or arrangement, changes made to the XML code document would be passed to any output generated without the clerical work usually associated with revising image type metadata documents.

Two comprehensive versions of the metadata product were created. The first one, which contains the confidential microdata and its metadata, was completed and demonstrated for an internal audience in Ottawa in October, 2004. The second version of the product, which retained the metadata of the analytical file but contained only the Public Used microdata, was created for demonstration in May 2005.

3.3 *NESSTAR/DDI* compliant metadata product assessment

Once the data documentation was entered in *NESSTAR/DDI*, basic testing on the utilization of the product in response to text searches of the questionnaire was encouraging. The single page display of variable descriptions, including questionnaire text and the statistical summary tables, made browsing for variables much easier, as there was no need to make concordance between variable name and the questionnaire using the data dictionary.

The DDI decision to adapt the use of XML is good. Using XML as the publishing language allows users to access the metadata without constraint by any particular statistical or reading application software, as long as users are equipped with any kind of web browser. Users no longer need an expertise in any particular statistical software or the survey to search for needed variables and to carry out simple data analysis. It is particularly useful when a researcher is in the stage of doing data discovery to be familiar with the data collections.

After the user decided on the survey using the web browser, sections of the survey description could be printed or download. Statistical summary tables could also be printed or exported to other applications such as Excel. The microdata or/and record layout and data definition could be converted to the statistical software format of the user's choice, such as SAS, SPSS and STATA, etc. Then the files could be downloaded for complex statistical analyses. Conditions could be set by producers to allow the user to browse the variables and the variables' statistical summary but to stop the actual download of the microdata file.

The potential uses of *NESSTAR/DDI* should exceed the creation of original copies of the metadata. However, the current version of the product was designed following the traditional code book format and geared toward handling the single path questionnaires often used in the "paper and pencil' type interview. Household surveys conducted by Statistics Canada commonly use multi-path questionnaires, which make use of filter questions to direct respondents to different sets of sub-questions in order to avoid having them answer irrelevant questions. Computer Assisted Interviewing (CAI) or Computer Assisted Telephone Interviewing (CATI) are usually used to conduct this type of multi-path questionnaire. All skip instructions are programmed in advance. During the interview, the computer automatically does the respondent error checks and directs the interviewers to the next set of questions based on the answers provided in the filter questions. There is no such computer assistance in the current version of *NESSTAR/DDI*.

Clients browsing a survey with hundreds of variables may not necessarily start browsing at the beginning of a questionnaire. In order to make users more aware of the complex nature of the questionnaire, the sample universe of each question was repeatedly defined in each question's associated variable descriptions. The filter and skip instructions were also noted in the questionnaire text included in the description of those variables. It was up to the users to include all the valid responses from all possible sub-questions with no automatic error check by the computer. The only benefit to users would be to make keyword searching easier compared to the original disseminated products, i.e. the printed questionnaire image and data dictionary.

3.4 Limitation of the assessment

[8] XLM was adopted by the World Web Consortium in 1998; a common language used for World Web publishing.

The major draw back of this pilot project is that it is difficult to draw any conclusions about the full adaptation of DDI by all or most population-based surveys in Statistics Canada.

Other limitations are that the assessment of the GSS 17 DDI compliant metadata product was mostly done from the standpoint of end users and limited to a few RDC staff and users. The products were produced retrospectively in the RDC instead of by the survey unit producing it as part of its dissemination.

Although there are claims that searching among multiple surveys or survey cycles in *NESSTAR/DDI* documents is easy and comprehensive, this study used only a single survey. Therefore no conclusions can be made about whether or not it is better at facilitating searches across surveys quickly by end users.

From a production point of view, applying DDI to a single survey's metadata conducted in isolation from others is straightforward, but to harmonize many surveys at different levels of dissemination and details can be complicated. For example, agreement on definitions, keywords and themes. It is conceivable that the feasibility of adapting DDI into Statistics Canada mainstream data production is more involved. A cost-utility evaluation at the survey unit level is recommended as well as a user utilization assessment using multiple surveys and involving more end users.

4. FUTURE WORK AND CONCLUSION

A consistent and unified view of metadata of different surveys will better help clients in selecting the most suitable data file for their research and identify variables they need among all available. Each step of information seeking depends on the availability of metadata.

Metadata are generated automatically throughout the process of data collection and beyond. There is a strong need to determine whether they are captured, and whether they are captured in a form solely for data collection agency staff or for end users as well. Statistics Canada has spent many resources and computerized each step for efficient collection and processing of survey data. The same attention should be placed on the dissemination, so that materials can be reprocessed and reused to avoid incrementally storing and reproducing the same information in different collections and in different formats.

Researchers' proposals using multiple datasets in a single RDC research project are getting more common and comparing data from one subject matter area with that of another will not be rare. In upholding the high regard from our Statistics Canada data users and our sponsors for our ability to provide high quality data, it is important that we adopt standards for metadata capture and dissemination of the population-based surveys that emphasizes the consistency of concepts, definition and measurement units, and the coherence of output data among subject matters areas.

If there are future attempts to assess the feasibility of Statistics Canada to adopt DDI metadata standards, the production units represent by far the most knowledgeable people regarding the associated costs in every aspect of the process and dissemination of the collection.

If DDI is to be adopted by Statistics Canada, the production should be undertaken or overseen by the production units, because relating the survey information to the DDI metadata with no conceptual errors requires in-depth knowledge of the survey design and the processing of the collected data. Furthermore, instead of retrospective metadata capture, producers should also consider possibly integrating the capturing, processing and dissemination systems. Computerizing the metadata capture should reduce the likelihood of human error and make the capture more timely and efficient. Since DDI documents are standard in using XML, it should be possible to integrate DDI into the existing Statistics Canada survey processing system, the *Blaise System*.

Data end users and data producers are not always the same group of people. It is important to have data producers becoming more active partners and to bring their perspectives into social science metadata specification and

development with the help of groups like the *DDI Alliance*[9], in order to build a metadata standard that provides better understanding and sharing of knowledge between data producers and users. In doing so, data producers should encourage DDI to evolve from the current version of the traditional codebook approach to the survey life-cycle approach (Colledge et al, 1996). Simultaneous capture of metadata at all stages of the survey process from questionnaire design, via data capture to data editing, instead of solely at data dissemination, should reduce the chance of knowledge loss in the event of staff turnover prior to data dissemination.

Data production agencies such as Statistics Canada should see DDI as a welcome tool for the survey units, as its mark-up extends down to the variable level and provides a standard uniform structure and content for variables, which not only helps in providing guidelines in metadata capture: it is also a means to help organize standards for harmonization of variables within, and collaboration across, subject matter areas.

REFERENCES

Colledge, M.J. (1999), "Statistical integration through metadata management", *International Statistical Review*, 67 (1999), 1, pp. 79-98.

Colledge, M., F. Wensing and E. Brinkely (1996), "Integrating metadata with survey development in a CAI environment", Australian Bureau of Statistics, Proceedings of U.S. Census Bureau 1996 Annual Research Conference.

Kunzler, U. (2002), "Electronic data reporting (EDR), metadata, standards and the European statistical system (ESS)", *Statistical Journal of the United nations*, ECE 19 (2002), pp. 119-130.

Mechanda, K., P. Johanis and M. Webber (2003), "Conceptual model for the definitional metadata of a statistical agency", *Proceedings of Statistics Canada Symposium 2003*, Statistics Canada, Catalogue no. 11-522-XIE.

Podehl, W.M. (2004), "Service improvements in Statistics Canada", *Statistical Journal of the United nations*, ECE 21 (2004) 1-6, IOS Press.

Richter, W. (1996), "The ABS Information Warehouse - Present and Future", *Proceedings of the Conference on Output Database*. Voorburg, November 1996. 11-18 JE Smith Statistics Netherlands.

[9] The DDI Alliance is a self-sustaining membership organization whose members have a voice in the development of the DDI specification describing social science data. Membership is open to educational, commercial, or governmental organizations. For more information, contact DDI Secretariat by email at secretariat@ddialliane.org.