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

DATA DOCUMENTATION INITIATIVE (DDI) IN THE REAL WORLD

A. Michelle Edwards[1] and Marie-Josée Bourgeois[2]

ABSTRACT

The Data Documentation Initiative (DDI) is an internationally developed standard used to develop metadata. The Data Liberation Initiative (DLI) along with partner universities, including the University of Guelph are working towards the goal of creating metadata for all Statistics Canada surveys available to the DLI community. Over one hundred tags were selected from the complete DDI codebook and applied to surveys in the DLI repository. The University of Guelph and the DLI team use the Nesstar Publisher product to develop metadata and the Nesstar Server to provide the data and metadata to our users. This presentation will review the tags that were selected, the process of building the metadata and demonstrate a complete survey now available to the University of Guelph community.

KEY WORDS: DDI, DLI,

1. INTRODUCTION

1.1 Academia

When the cost of Statistics Canada data increased in the 1980's, researchers, students and instructors at Canadian post secondary institutions made increased use of American, British and even Chinese data. This cheaper foreign data did not always reflect the Canadian situation, and there were often gaps in the data. Furthermore, many academic institutions had not provided the technical support scholars needed to handle complex data files.

To help buffer the increased costs of the 1986 Census, an ad hoc buying consortium was organized in 1989 by the Canadian Association of Public Data Users (CAPDU) and the Canadian Association of Research Libraries (CARL). The experience demonstrated the possibility of a successful consortial arrangement between Statistics Canada and Canadian academic institutions.

In 1993, a working group sponsored by the Social Sciences Federation of Canada (SSFC) came up with a plan that was acceptable to both Statistics Canada and the academic community. Statistics Canada and the Depository Services Program played key roles in this process. In February 1996 the Data Liberation Initiative received approval from Treasury Board for a five year pilot project and was included as part of the federal government's Science and Technology Strategy in March of that year. The project proved to be an overwhelming success and in April 2001 the DLI was made into a permanent program situated in the Library and Information Centre at Statistics Canada.

[1] A. Michelle Edwards, Academic Services, University of Guelph, Guelph, Ontario, N1G 5H8; Marie-Josée Bourgeois, Data Liberation Initiative, Statistics Canada, Ottawa, Ontario, K1A 0T6

1.2 Data Liberation Initiative

The Data Liberation Initiative (DLI) is an excellent example of a cost effective method for improving data resources for Canadian post secondary institutions. Prior to the start of the DLI program, Canadian universities and colleges had to purchase Statistics Canada data, file by file. With the advent of the DLI, participating post secondary institutions pay an annual subscription fee that allows their faculty and students unlimited access to numerous Statistics Canada public use microdata files, databases and geographic files. Academic researchers now have affordable and equitable access to the most current statistics and other data, which gives them powerful tools to use in their analysis of Canadian society.

The DLI has been at the forefront of the creation of a data culture in Canada. Data librarians and the beginnings of data centres have sprung up all over the place. These data librarians have trained one another and are working together to provide a service to the community. They're also working with professors to get students involved with social statistics.

The DLI represents a major application of Canada's information highway technology. It allows universities and colleges, for the first time, to offer a full range of data services to students and faculty alike. There is also growing evidence that the Initiative is making important contributions to Canadian teaching and research. Courses are being reconfigured to encourage students to use the data and grants have been won for proposals directly related to the availability of data through the DLI.

Researchers, who formerly had to depend mainly on public opinion polls as a source of Canadian data, can now supplement them with Statistics Canada public use microdata.

Today, there are 68 participating universities/colleges in the DLI community.

DLI has now over 250 surveys in their collection of aggregated files, Public Microdata Files, SPSS /SAS files and accompanied documentation in Word or PDF. These files are accessible to the academic institutions through a FTP site which resides on a DLI Statistics Canada server.

2. SURVEY DOCUMENTATION

2.1 Today – The Real World

Survey documentation often includes several parts, a codebook which describes survey question coding; a user's guide, which outlines how the survey was conducted, how weights were calculated and collections methods; and a datafile, often a flat ASCII file containing a series of numbers. To interpret the datafile, users need either a record layout file which outlines question/variable positions and the file contents, or a statistical package syntax file, such as SAS or SPSS.

When DLI first came into existence, academic institutions received raw data files (ASCII format), and an accompanying record layout file. Any accompanying survey documentation available, whether it is a codebook and/or a user's guide, was received as hard copy in a binder. Researchers looking to analyze any survey results needed to borrow a copy of the documentation from the academic institution's library. Since a file with numbers was not very useful in determining variables and codes needed for their analysis.

As the Internet developed and users became more computer savvy, survey files were made available to users via centralized web interfaces. Statistical package syntax files were also made available to help users work with the files in an environment they preferred. In many cases, users could now conduct preliminary analysis without borrowing hard copy documentation from the library; however, at some point during their research, the documentation would be required to review the survey methodology.

Today, many codebooks and user guides are available in electronic format. Users can now obtain required documentation electronically. However, users still need to access several files to obtain the information required for their research and analysis. A syntax file to read the datafile, a codebook to interpret the codes used in the datafile, and a user's guide to understand the survey methodology.

2.2 Tomorrow – The Perfect World

Access to survey data and documentation has moved from a mix of electronic and hard copy format to the case where all documentation is available in electronic format. However, to get a complete picture of the survey, users still need to access a number of different files. A codebook or electronic file that contained all survey documentation including the data file would be the next generation of survey access. Users could then obtain methodology details, code definitions, data and survey background information all in one file – a true one-stop shopping idea. Imagine if this file were searchable as well. Users could search for question / variable level information before gathering the data, cutting down on time required to do preliminary analysis.

The question then becomes, how do, as data providers, help make the transition from several electronic files to one file or codebook. A new initiative, the Data Documentation Initiative, may have the answer. An XML-based file composed of tags describing all documentation associated with a given survey.

3. DATA DOCUMENTATION INITIATIVE

3.1 Background

The Data Documentation Initiative (DDI) is an international effort to establish a standard for technical documentation describing social science data (http://www.icpsr.umich.edu/DDI/). The goals of the DDI are to facilitate the exchange and transport of documents by creating an XML-based standard, to assist in the preservation of documentation for datasets and to improve survey documentation by "retaining all the capabilities of the electronic codebook but greatly increasing the scope and rigor of the information contained in it" (http://www.icpsr.umich.edu/DDI/codebook/index.html).

The DDI metadata specification was initiated by the Inter-University Consortium for Political and Social Research and is now a project including an Alliance of 25 institutions in North America and Europe. The Alliance meets approximately twice a year and is comprised of several working groups.

3.2 The DDI Codebook

The DDI codebook is XML-based file composed of a number of tags and attributes within the tags. Each tag describes an aspect of the survey, for example the title of a survey would appear as: <titl>Survey of Consumer Finances</titl> or the distributor of the data would appear as: <distrbtr abbr="DLI" affiliation="Statistics Canada">Data Liberation Initiative</distrbtr>. The tags are <titl> and <distrbtr>, whereas the attributes would be metadata listed within the tag, such as abbr="DLI" and affiliation="Statistics Canada" shown inside the <distrbtr> distributor tag.

As an XML file, there are several options available to developers to make the information available for viewing to users. The file can be easily viewed in an Internet browser as a listing of tags and their contents. However, users prefer to view the XML file in a familiar format, HTML. Programs such as Saxon and Nesstar are just 2 examples which can be used to show XML files in an HTML format. Both the DLI group and the University of Guelph have chosen to use the Nesstar suite of products for both developing and viewing DDI codebooks, primarily because Nesstar was developed specifically for the DDI standard.

The DDI codebook is divided into 5 sections: Document description, Study description, File description, Variable description, and Other Documentation. Each section contains a number of tags and attributes to help describe the

contents. Within each section, the tags are deemed required (these map to the Dublin Core tags used by library cataloguers), recommended, or repeatable. Repeatable tags refer to tags that may be repeated to show that there are several instances of a given piece of metadata, variables for example. The <var> tag is used for each variable available in the survey and is therefore listed as a repeatable tag.

3.2.1 Document Description

The first section of the codebook describes the XML file or DDI-compliant codebook. There are 71 tags available, 4 of which are required.

This section consists of the bibliographic information which describes the DDI-compliant codebook file itself. Tags in this section include items such as: the title of the document, ID number, names the document creator, and where it is stored.

3.2.2 Study Description

The study description section deals with the study or survey itself. It contains information on who conducted the survey, who collected the data, how and when the data was collected, units of analysis, geographic coverage, an abstract of the study, keywords for searching, etc… This section has a total of 243 tags available, of which 26 are required.

3.2.3 File Description

The third section of the DDI-compliant codebook describes the datafile accompanying the study in question. Structure of the files, number of variables, number of observations, and format of the files are some of the tags included. A total of 34 tags are available and of which none are deemed required.

3.2.4 Variable Description

The variable description section is one of the most valuable sections of the codebook to the user. Each variable in the datafile is tagged with information such as variable name, variable values, literal question associated with the variable along with any pre- and post- question instructions. A frequency or mean value and range are also available for each variable. Users can search variable and question information prior to retrieving data for further analysis.

There are a total of 91 tags of which only 4 are required. The tags available in this section are repeatable which means they are available for each variable in a dataset.

3.2.5 Other Documentation

The last section of the DDI codebook allows codebook developers to link documentation created by survey authors to the data. User guides and survey codebooks are typical examples of documentation that are listed in this section. By adding these files to the DDI codebook, users now have access to one file which contains both the metadata and data for a given survey. There are 49 tags available and 8 tags are required and repeatable for each document listed in the codebook.

3.3 Selecting DDI tags

There are close to 500 tags available for the user to mark-up or create a DDI compliant codebook. How does one decide which tags should be included or not. At the 2000 IASSIST conference held in Amsterdam, a meeting of individuals working towards creating DDI codebooks suggested that creators start with the 30 required tags at that time. These tags included items such as title, producer, and date created. As work began at the University of Guelph, we realized that 30 tags were too few. Each dataset mounted through the University of Guelph's web retrieval system already had a 'readme' file associated with it. The metadata contained in these 'readme' files

exceeded the tags listed as the basic 30 tags. As a result, we started with the basic 30 tags and the additional tags required to include the 'readme' file metadata.

As work continued at University of Guelph and we began to tackle some of the more complicated surveys and surveys whose name changed over the years (Survey of Smoking Habits of Canadians, Labour Force Survey, Census files), the number of tags we felt were essential increased. The tags were added to facilitate user interpretation of the metadata. To this date, we are using approximately 150 tags and attributes to document the datasets in the University of Guelph collection.

4. DDI at the University of Guelph

At the University of Guelph, we have been creating DDI compliant codebooks for each dataset in our data holdings, an approximate total of 770 datasets. Our holdings comprise of DLI surveys, ICPSR surveys, and many datasets from international sources including the International Monetary Fund (IMF) and the United Nations (UN).

The TriUniversity Data Resources (TDR) is a joint project among University of Guelph, University of Waterloo and Wilfrid Laurier University. The TDR provides access to the datasets through a centralized webpage. At the beginning of the Winter 2006 semester (January 2006), all datasets held at UG have DDI compliant codebooks and are now available to our TDR partners using the Nesstar program. Users can now search across and within datasets for the information required for their research projects. To date, users have welcomed the Nesstar web interface and the increased ability to search all survey documentation.

5. DDI at DLI

Our goal is to have the entire DLI collection DDI compliant in both official languages for better and more consistent documentation of data files and long term access/preservation.

All English and French monthly and annual files of Labour Force survey (1976 to 2005 – 720 files) are completely done.

DLI has chosen to purchase the Nesstar Product in order to provide complete DDI files to the academic institutions.

Nesstar Publisher is an advanced data management program, DDI compliant and meets our needs. It consists of data and metadata conversion and editing tools. Ultimately, Nesstar Publisher can be used to publish data and the accompanying documentation to a catalogue on a Nesstar Server. From here these resources can be made available to the DLI community via Nesstar WebView.

Since metadata authoring can be time consuming, Nesstar Publisher's use of the DDI makes the process simpler, more efficient, and more flexible. Within Nesstar Publisher the DDI structure (or a subset thereof) may be used to import information, integrate data and metadata, and customize the process of producing metadata within the overarching framework of the DDI itself.

In this way, metadata can be produced according to established DDI templates, or can be customized according to the needs of individual users.

To create the metadata, we start with a SPSS portable file, apply template of tagsets (UG and DLI) and collect documentation from user guides, codebooks, questionnaires, the Statistics Canada Online Catalogue and the Integrated Metadatabase (IMDB). After completion of a dataset, it is publish on the NESSTAR server and will be made available to our users.

6. CONCLUSIONS

Over the past 10 years we have seen survey documentation progress from hard copy binders accompanied by raw datafiles and record layouts to today's DDI compliant codebook which holds all survey documentation, data and related documentation. Researchers can now use programs such as Nesstar as their one-stop shop to search for survey data and accompanying information.

As we create more and more DDI codebooks we are moving away from the "Real World" scenario and slowly reaching our "Perfect World" scenario. As data creators recognize the benefit of creating one file that contains all survey documentation, data providers will be able to provide DDI codebooks and data to their users efficiently and in a timelier manner.