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

DISCOVERING MICRODATA VARIABLES: COMPARING DDI COMPLIANT DOCUMENTATION TO AN ISO/IEC 11179 METADATA REGISTRY

Tim Dunstan[1], Charles Humphrey[2]

1. INTRODUCTION

Documenting microdata from Statistics Canada surveys has always been an important task, whether the data are from a confidential master file or a public use file.[3] Both of these file types consist of a collection of variables that must be thoroughly and precisely described so that the data can be understood and used in analysis. Secondary uses of microdata are highly dependent on information describing the variables in full detail, including the record layout on the physical storage medium, the coding scheme making up each variable's response set, the origin of each variable in the original survey instrument, the treatment of non-responses and the distribution of responses for each variable. Contextual information, such as instructions about the use of a weight variable to adjust for the survey design or guidelines about sampling variability, also provides critical instruction in analyzing a survey's data. Collectively, all of this information is integral to the overall documentation story and provides a larger picture of the data and the circumstances in which it was created. In the data archiving community, discussions are now occurring about the use of a life-cycle model to capture data documentation, where information is gleaned throughout all stages of a survey. This leads to the following pertinent question: what information should go into data documentation?

For the most part, the frameworks used to define the accepted elements of data documentation have been based on convention or best practice. Very recently however, a couple of international projects have redefined the content and structure of data documentation through the development of two metadata standards. This paper compares how well these two standards, ISO/IEC 11179 and DDI, support the documentation of variables in microdata files. Using an overall model of the information most commonly sought by researchers performing secondary data analysis, we assessed how well each standard captured and described this information.

2. CONNECTIONS BETWEEN MICRODATA DANS TECHNOLOGY

Metadata are information organized systematically to describe other information through definitions, illustrations, classifications, conceptualizations and other ways in which relationships among objects can be expressed. Metadata within the field of information technology are computer processable to facilitate access to the information it describes or to generate new ways of viewing and understanding information. Metadata in this context are dependent both on existing computing technology and the intellectual frameworks used to organize information.

For example, CANSIM I records were structured in fields of fixed length and organized for sequential processing, which fit the dominant mode of computing in the 1960's and 1970's when information was stored on magnetic tapes and processed sequentially by mainframe computers. One impact of forcing CANSIM metadata into fixed fields was the proliferation of abbreviations used to squeeze as much information as possible into fields of a fixed size. This also resulted in many instances with multiple abbreviations for the same concept. A longer or shorter

[1] Statistics Canada
[2] University of Alberta
[3] The master file contains data about survey participants in enough detail that individual respondents could be identified. These data are not available to the public. On the other hand, public use microdata have been anonymized to minimize the likelihood of disclosing any particular respondent. The Statistics Canada Data Release Committee must approve the level of anonymization, however, before a microdata file will be made public.

abbreviation would depend upon the amount of space remaining in the field. Clearly, the computing technology of the era had a big impact on the substance of the metadata of that time.

The development of CANSIM II was conducted in the 1990's when the dominant form of information technology consisted of a distributed network using magnetic disk storage with direct file access to information. A relational database design replaced the structure of CANSIM I records freeing attributes from fixed record locations with limited space for information. This new structure also enabled the linking of information in one field with another field and eliminated a lot of redundant information over records. This increased the substantive control over concepts and their abbreviations.

The CANSIM example demonstrates how metadata have been shaped by the technology of the day. The two standards for metadata in our study similarly reflect the computing environments in which emerged. Their power is derived from today's technology rooted in distributed Internet computing. ISO/IEC 11179 is a schema that works well with distributed relational databases; DDI employs a text-encoded mark-up language built upon Web technology. While both metadata standards may hold identical information, implementing the standards leads to different processing streams. This leads to one question unanswered by this study: do advantages exist one way or the other in organizing the metadata for microdata in a relational database structure or through Web-based metadata management?

3. MICRODATA VERSUS AGGREGATE DATA: METADATA DIFFERENCES

One other background issue needing some discussion at this point is the distinction between microdata and aggregate data and how the metadata between these data structures differs. As mentioned above, microdata consist of information gathered at the observational level and contain details about specific individuals or members of the unit of observation. Microdata are often referred to as raw data because they generally have not been processed and remain at the level at which the information was initially observed.

Aggregate data, on the other hand, have been processed to a level higher than the one at which the data were observed. In other words, aggregate data consist of summarized microdata. Such aggregation typically occurs along geography, time or some social construct. For example census microdata are commonly summarized along standard census geographies, such as, census divisions, census subdivisions, census metropolitan areas, census tracts, etc. The exercise of aggregating microdata in itself requires precise definitions of the variable over which the summaries are to be calculated as well as the definitions of the variables being summarized. Such definitions are integral to the metadata describing the aggregate data file.

Why introduce the differences between aggregate data and microdata? Of the two metadata standards examined in this study, it may be that one is better suited to microdata while the other better serves aggregate data. Because aggregate data and microdata have different characteristics, it is possible that different metadata standards capture this uniqueness more adequately. The records in a microdata file are at the level of observation of the original survey, while aggregate data records are not (e.g., a record may summary information about many individuals within a census division). By comparison, the metadata for microdata must include more detailed information about the unprocessed data. In comparison, the metadata for aggregate data must include information about the concepts used to perform the aggregation. Thus, there are qualitative differences between the content that should be described in the metadata of these two types of data structure. While this study does not explore these differences, this distinction led us to a focus on just one of the two data structures, namely, microdata.

4. THE PURPOSE OF THE STUDY AND ITS METHODOLOGY

The purpose of our study was to compare the ISO/IEC 11179 and DDI metadata standards using a model of the information requirements of a researcher conducting a secondary analysis of a microdata file. Elements or attributes in each metadata standard were identified that could hold the information being sought by the researcher if the documentation was fully implemented using the metadata standard. This study did not attempt strictly to map one standard onto the other. The comparison was based instead on applying the standard to fulfill the information needs of the end user.

The rationale for this approach is derived from the above discussion about the different characteristics of microdata and aggregate data. Instead of letting the metadata standards determine the information domain for microdata documentation, we chose to use an external model, that is, the information requirements of a researcher conducting a secondary analysis. Using the information elements from this external model, we then attempted to locate similar elements in each metadata standard.

Another reason we chose the end-user perspective was that the metadata domain, as defined by the microdata producer, may not adequately cover all the information requirements of the end user. In the case of the Statistics Canada example, the data producer is also the creator of the metadata, which weights the design of the metadata to the producer's outlook rather than those of the end-user of the data. Taking the perspective of the end user is like looking backwards into the telescope of data documentation.

The perspective of the end user does not necessarily represent the most comprehensive model of data documentation any more than the outlook of the producer of microdata files. The information that should be included in data documentation is an ongoing research issue. As mentioned earlier, the data archiving community is now investigating a life-cycle model to address this matter. A life-cycle approach introduces process into the documentation model and identifies meaningful information to document throughout the stages of a survey.

We also chose the end-user perspective to control for the effects of technology in comparing the two metadata standards. We were not concerned whether a relational database structure was better than a text-encoded tag language. Given both of these metadata structures, we wanted to know how well they handled the information needs of the end-user.

We focused specifically on microdata because of the level of detail associated with these data. The nature of microdata and aggregate data, which was discussed above, may be just different enough to make one standard more useful for microdata than the other. We also wanted to examine how well ISO/IEC 11179 documents data at the variable-level. The Integrated Metadata Database (IMDB) in Statistics Canada already employs the ISO/IEC 11179 standard and has been implemented with time series data, which is a form of aggregate data. There was a desire to see how far this standard could be employed to describe variables at the microdata level in the IMDB. Finally, our study had a very pragmatic objective to see how easily metadata about variables in one of the two standards could be translated to the other standard.

Figure 1
Tasks Requiring Information in the Data Discovery Stage

Find variables according to common themes or subject headings.
Find data using a unit of observation appropriate for a desired analysis
Browse short title descriptions of variables of interest.
Read about the full description of a variable.
Identify a variable's level of measurement and response set.
Examine the range of values for a variable within a sample.
Examine the labels of variables consisting of categorical measurement.
Examine the distribution of responses for a variable.
Determine the type and rate of non-responses to a variable.
Determine who in the sample provided information for this variable.
Determine the need for sampling weights
Assess the quality of the data and determine the use of imputation and proxy responses.
Establish the context of the information within the questionnaire.
Determine the source for the information in this variable.

The methodology we used in comparing these two metadata standards consisted of two steps. First, the various stages in conducting a secondary analysis were outlined and the information requirements of each stage were identified. The elements in each standard that would contain this information were then identified. Secondly, a set of variables from the General Social Survey Cycle 17 (GSS 17) was chosen to represent the variety of variables found in microdata files. Metadata was then generated in each standard for these variables if possible. Finally, the completeness of representing the metadata for these variables was compared.

5. THE INFORMATION NEEDS OF THE END USER

The information needed to conduct secondary data analysis was categorized under three main stages: Data Discovery, Data Extraction and Data Analysis. Within each stage, multiple tasks were identified, each with its own information requirements. For example, the Data Discovery stage consisted of fourteen separate tasks (see Figure 1.) The type of information being sought was then described within each of the tasks. For example, the metadata content in the first task in Figure 1 was itemized as a list of commonly accepted subject headings, keywords or themes to categorize each variable. The elements in ISO/IEC 11179 that would hold this information would be THEME, TOPIC, KEYWORD, while the DDI elements are <varGRP><labl> and <varGRP><txt>. The complete listing of tasks, metadata content and ISO/IEC 11179 and DDI elements are available from the authors.

Figure 2

General Social Survey Cycle 17 Variables Used to Test Metadata Implementation

ACMYR	*Main activity of the respondent in the last 12 months*
EDUSTAT	*Full-time or part-time education status for the respondent*
MAR_Q125	*Are you looking for paid work?*
MAR_Q130	*Did you have a job or were you self-employed at any time during the past 12 months?*
AGE_LSTPDWK	*Age of respondent when did last paid work. (suppressed in the public use file)*
AGE_LSTPDWKC	*Age of the respondent when they last did paid work.*
MAR_Q140M	*In what month did you last do any paid work? (suppressed in the public use file)*
MAR_Q140Y	*In what year did you last do any paid work? (suppressed in the public use file)*
MAR_Q140A	*How old were you when you last did any paid work? (suppressed in the public use file)*
LS_Q210	*Using the same [10-point] scale, how you feel about your life as a whole right now?*
LS_Q120	*Please rate your feelings… What about your job or main activity?*
LS_Q130	*Please rate your feelings… What about the way you spend your other time?*
EDUC10	*Highest level of education obtained by the respondent – 10 groups.*

Once these elements were identified for the two metadata standards, thirteen variables from the GSS 17 master file were selected as representative of the types of variables typically found in microdata files. Variables were chosen to meet one of the following criteria.

- Variables from questions involving branching or skip patterns;
- Variables that could be used to derive new variables;
- Variables that have the same category labels but use different scales;
- Variables from an open-ended question; and
- Variables that appear across microdata files for which standard categories have been established and where the distribution of responses may be compared.

Each of these variables was then described applying both metadata standards. For DDI, the NESSTAR Data Publisher® was used to generate this metadata.[4] The source of the information came from the GSS 17 Data Dictionary distributed by the author division, the questionnaire for GSS 17, the SPSS® version of the data file and personal classifications for variable group names and keywords.

6. STUDY FINDINGS

All of the test variables from GSS 17 were described in detail in DDI according to the information needs of the secondary data analysis model for the Data Discovery Stage. Metadata for four of the five tasks for the Data Extraction Stage were also captured by DDI. However, the two tasks under the Data Analysis Stage were beyond both metadata standards.

ISO objects or combinations of ISO objects can be used to produce the variable metadata required of the secondary data analysis model. ISO 11179 can be mapped to produce equivalent DDI compliant documentation.

[4] The Data Publisher is an MS Windows application licensed by NESSTAR Inc. that outputs an XML file of DDI-compliant metadata. This application uses forms to identify information fields that are associated with DDI elements. The advantage of this tool is that documentation can be completed without having to tag the information directly. Data Publisher can also read SPSS internal files to capture the information for many of the elements used to describe variables.

7. CONCLUSIONS

Version 2 of the DDI standard was used in this study, which consists of a five-part tag library to mark-up the text of traditional social science codebooks.[5] In this study, the elements of Part IV, which describe variables, were the most relevant to the end-user information model. The implementation of Version 2 results in a large volume of redundant coding across variables with identical information. For example, variables using the same value labels are each encoded with the set of tags and labels. This repetitive coding not only makes the XML file containing the metadata very large but also raises processing issues in mapping information between the two standards in this study. Propagating DDI-tagged value labels from the ISO/IEC 11179 field containing this information is much more straightforward than reversing the process from DDI to ISO/IEC 11179. This asymmetry arises because of the nature of the element models of these two standards.

Late in the fall of 2005, Version 3 of DDI was released by the DDI Alliance, the oversight body maintaining the DDI standard.[6] A substantial restructuring of the standard occurred in this new version moving away from the original hierarchical five-part codebook design to a modular design allowing more complex relationships among the metadata elements. An advantage of this new object-oriented version of DDI is its capacity to incorporate and compare external metadata. For example, external vocabularies and thesauri can be linked at the variable level without duplicating or replicating content. Consequently, classification databases containing more abstract information at a conceptual or definitional level can exist outside the DDI metadata. Such external registries can be used to make comparisons of the similarities and differences among variables, concepts and definitions in a standardized way. For example, the IMDB is an ISO/IEC 11179 metadata registry maintaining standard concepts and definitions. These elements can be linked to variable-level microdata documentation for Statistics Canada surveys, which tends to be at a less abstract level and require in many instances repetitive and redundant links to describe microdata variables. DDI and ISO/IEC 11179 are complementary standards that can serve separate but related functions. Together, the metadata potential for microdata has been greatly enriched.

[5] These five parts consist of elements to describe the metadata document, the survey or study in which the data were created, the data files in the survey, the variables in each file and other study-related materials. A DDI-compliant metadata file is one large ASCII document containing tags from each of these five parts and the content describing a survey. Wrapped in XML, these documents can be processed over the Internet and displayed by employing an endless variety of style sheets.

[6] The DDI Alliance is a membership organization with subscribing institutions having a say in the development of the DDI standard. More information about the Alliance can be found at www.icpsr.umich.edu/DDI/org/index.html