Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

ROLE OF THE INTEGRATED METADATABASE AT STATISTICS CANADA

Paul Johanis[1]

ABSTRACT

The Integrated Metadatabase is a corporate repository of information on each of Statistics Canada's surveys. This information includes a description of data sources and methodology, definitions of concepts and variables measured and indicators of data quality. It provides an effective vehicle for communicating data quality to data users. Its coverage is exhaustive of Statistics Canada's data holdings, the information on data quality provided complies with the Policy in Informing Users of Methodology and Data Quality and it is presented in a consistent and systematic fashion.

KEY WORDS: Integrated Metadatabase; data sources definition

1. INTRODUCTION

The Integrated Metadatabase is a corporate repository of information on each of Statistics Canada's nearly 400 active surveys. These surveys are the Agency's core activities and the IMDB is the principal mechanism by which they are documented, providing a key information resource for corporate knowledge management and for data users.

Metadata can support at least three broad functions within a statistical agency: data dissemination, data production, including collection, and management of the statistical system. IMDB was designed specifically to support data dissemination, that is, to provide users with the information they need to interpret the statistical data we disseminate. To meet this objective, the content of the IMDB was based in large part on the requirements of the Policy on Informing Users of Data Quality and Methodology (PIUDQM).

The database is resident on a central server. Metadata were collected from a variety of pre-existing metadata stores, reformated and validated and loaded into the new metadatabase. On a regular basis, an HTML generator reads the database and produces formatted HTML pages, which are made available on the Statistics Canada website. These pages were designed to meet the requirements of the PIUDQM. They can be accessed through hyperlinks from our output online database, known as CANSIM 2, from our online catalogue or from statistical tables on the website. The pages can also be accessed directly through a search engine in the metadata section of the website. The database is kept up to date through an input system deployed over the departmental Intranet. Updates are quality assured and registered before being made available for generation of the external HTML pages.

2. INFORMATION ON DATA QUALITY AND METHODOLOGY IN IMDB

According to the Policy, a specific set of summary information on data quality and methodology must be presented or made available to users for each statistical product. The summary documentation required by the Policy is to be organized according to a number of standard headings, each containing a generally

[1] Paul Johanis, Director Standards division, Statistics Canada, Ottawa, Ontario, Canada,K1A 0T6

consistent set of information. These headings and the guidelines for the content of each section are reflected exactly in the web pages produced from the IMDB.

Upon accessing the entry point to the IMDB on the STC website, users are presented with a standardized message introducing the information on data quality and methodology and emphasising the importance of taking it into account. Meta-information is organized around an entity known as the survey. A survey can be a direct survey, a statistical program that uses administrative data or a data integration activity. There are approximately 400 surveys in the agency, each of which is documented in the IMDB. Each survey can have one or more survey instances, each representing one cycle of the survey. For example, a monthly survey has 12 instances per year. Metadata are collected for each survey instance.

It is at the level of the survey instance that the information on data quality and methodology is provided to data users. The challenge in defining the information to be provided for each survey instance is to achieve an appropriate balance between comprehensiveness and excessive detail. There is a constant temptation to list every possible attribute of these entities but such a detailed database could never be kept up-to-date as its maintenance would represent too large a burden on survey managers to provide the necessary information. Our guiding principle has been to refer to the fundamental objective of our project, which is to produce information necessary for data users to correctly interpret and to place in context the data we disseminate.

Under the Data Sources and Methodology heading, an introductory paragraph provides the purpose, objectives and general nature of the survey and a statement on the time frame or reference period of the data. A section on the conceptual universe and the target population of the survey follows, describing the statistical units covered by the survey. In addition to this basic information, users require information on the methods used to carry out statistical activities. This, in conjunction with information on the quality of the data produced, enables users to judge the extent to which the data source responds to their needs.

In the IMDB, a methodology entity has been defined, which can assume one of several methodology types. These types include the sampling plan, collection and capture method, error detection procedures, imputation method, estimation method, time series processes and disclosure control method. It also includes quality evaluation procedures, from which links to various reports and studies on sources of error and other aspects of data quality can be made. In this way, standardized descriptions of survey methods can be obtained. It is also possible to link from these standard descriptions to more extensive documentation regarding the appropriate survey method.

The next major heading required under the policy is Concepts and Variables Measured. For each survey program, the list of variables produced, along with their definitions and classification, will be included in the database. To model this aspect of the database, we have complied with the ISO 11179 standard. The basic components of that standard have been retained: data element concept, object class, data element and value domain. We have used, however, terminology that is more familiar to statistical users. The data element concept is simply known as a concept in Statistics Canada, the object class is known as a statistical unit, the data element is a variable while the value domainwith associated value meanings corresponds to what we call a classification. Nevertheless, the basic structure and relationships implicit in the standard have been preserved in our model. Each of these elements has been defined in the departmental Policy on Standards. Under this policy, standards can be declared regarding the definition of concepts, variables, classifications, units and populations. Statistical programs that have adopted these standard definitions will therefore have no need to provide additional documentation for their variables as these will already have been documented as standards, which will be resident on the database. The final mandatory heading under the Policy is Data Accuracy. Under this heading, various quality measures have been defined at the survey instance level. They include the components necessary to calculate the response rate, coverage error and imputation and sampling error for key variables. The sampling error will be expressed as coefficients of variation. CVs for selected important variables only will be recorded in the metadatabase. Again, it is possible to link to more extensive documentation on data quality. The full data model is contained in appendix 1.

3. CONCLUSION

The Integrated Metadatabase provides a very effective vehicle for communicating data quality to data users. Its coverage is exhaustive of Statistics Canada's data holdings, the information on data quality provided complies with the Policy in Informing Users of Methodology and Data Quality and it is presented in a consistent and systematic way. Having created such a system, the challenge will be to enhance and maintain the quality of the metadata itself.

Appendix 1: Meta-information headings of IMDB

- Statistical activity
- Survey
    - Universe
    - Data element concept
        - Concept
        - Statistical unit
        - Classification
    - Frame

- Survey Instance
    - Questionnaire design
    - Survey instrument
    - Sampling plan
    - Sample
    - Collection and capture method
    - Editing procedures
    - Imputation method
    - Quality control procedures
    - Estimation method
    - Time series processes
    - Data files
    - Disclosure control method
    - Quality evaluation method
    - Quality measures
        - Coverage error
        - Response rate
        - Imputation rate
        - Sampling error for key variables
        - Other quality rating

For every item on this list, the following generic information is collected:

- Time frame
- Theme
- Topic
- Keyword
- Organization
- Contact
- Additional documentation