STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint UNECE/Eurostat Work Session on Statistical Metadata
(6 - 8 March 2002, Luxembourg)

Working Paper No. 19
English only

Topic (iii): Metadata and quality

ASSESSING THE QUALITY OF METADATA: THE NEXT CHALLENGE

Submitted by Statistics Canada[1]

Invited paper

ABSTRACT

Over the last several years, many National Statistical Offices have been busily constructing metadata systems of various designs and contents. Once constructed and loaded, however, the challenge is to maintain the quality of the metadata stored in these repositories. To this end, we have turned to Statistics Canada's Quality Assurance Framework (QAF), which is used to assess the quality of data. The QAF comprises six dimensions of data quality, or fitness for use, which can be adapted for the assessment of metadata quality. The six dimensions are relevance, accuracy, timeliness, accessibility, interpretability and coherence. The paper establishes that the Quality Assurance Framework provides a sound basis for establishing criteria for assessing the quality of metadata. While the various dimensions of the framework need to be re-interpreted in a metadata context, they provide useful insights into the "fitness for use" of metadata. Based on this cursory preliminary analysis, it would seem possible to develop a systematic set of indicators that can be used to assess the quality of metadata within and across sources of statistical data.

I. INTRODUCTION

1. Over the last several years, many National Statistical Offices have been busily constructing metadata systems of various designs and contents, helped considerably by the mutual sharing of information through forums such as METIS. Statistics Canada is no exception. Since 1998, it has developed its Integrated Metadatabase (IMDB) as 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.

2. 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 to initially support data dissemination, that is, to provide users with the information they need to interpret the statistical data we disseminate.

3. 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. 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 external use. Several times a day, an HTML generator reads the database and produces formatted HTML pages, which are made available on the

[1] Prepared by Paul Johanis.

Statistics Canada website. 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.

4. The development of the database and accompanying systems has been recognized by the senior management of Statistics Canada as a significant technical achievement. However, it was also noted that the information it contained was of uneven quality, in many cases not fulfilling its primary objective, that is to provide data users with information they need to interpret and assess the quality of the data we disseminate. As a result, the main focus of the project over the last year has been the ongoing improvement of the quality of the metadata loaded in the IMDB.

II. STANDARDS FOR METADATA QUALITY

5. To improve the quality of metadata, it is necessary to have a standard that establishes what constitutes "good" quality. To this end, we turned to Statistics Canada's Quality Assurance Framework (QAF), which is used to assess the quality of data. Metadata are, after all, fundamentally data and perhaps share the same quality attributes. The QAF comprises six dimensions of data quality, or fitness for use, which can be adapted for the assessment of metadata quality. The six dimensions are relevance, accuracy, timeliness, accessibility, interpretability and coherence.

III. RELEVANCE

6. For metadata, relevance means providing the right metadata at the right level of detail, for its intended purpose. As the intended purpose of the IMDB is to provide information for users to interpret the data STC disseminates, the challenge then is to define what information users need for this purpose. To address this issue, the content of the IMDB was largely dictated by Statistics Canada's Policy on Informing Users of Data Quality and Methodology. According to this Policy, a specific set of summary information on data quality and methodology must be presented or made available to users for each statistical product, under specified standard headings.

7. 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 unit and the population of such units intended to be 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.

8. In the IMDB, a methodology entity has been defined, which can assume one of several methodology types. These types include the sampling plan, collection 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, descriptions of survey methods can be obtained under a standardized set of headings.

9. 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.

The final mandatory heading under the Policy is Data Accuracy. Under this heading, various quality measures have been defined. They include the components necessary to calculate the response rate, coverage error and imputation and sampling error for key variables. Again, it is possible to link to more extensive documentation on data quality.

10. Exactly what information to provide under each these headings is not self-evident, however, and our experience has been that it is necessary to provide additional guidelines and instructions to authors in order to collect a consistent set of information under each. In addition to providing guidance to authors, these guidelines can be used as benchmarks for evaluating the completeness of the information provided (see Appendix 1).

11. In the end, however, market testing and other consultation with users will help us determine if this is the right information to provide, at the right level of detail.

IV. ACCURACY

12. In data quality terms, accuracy is measured through the detection and quantification of various types of error. A similar approach can be taken for metadata, in which case, accuracy refers to the extent to which the information stored in IMDB is correct and up to date. While there are very many specific quality checks that could be made, they can be grouped into two basic categories: coverage errors and measurement errors. In each of these categories, it has been our experience that a few specific types of errors are more common and these will be described.

13. Coverage error is a measure of the extent to which the metadatabase contains information on all the objects it is intended to cover. For our purposes, these objects are "surveys", defined according to some agreed upon convention (see the section on coherence below). The simplest error, and perhaps the most difficult to detect, in this respect is the completeness of the list of surveys covered by the metadatabase. In a large decentralised organization, it can be difficult to ascertain whether every survey has been identified. In Statistics Canada, there exists a list of officially approved surveys, but it only covers direct surveys, that is surveys in which information is collected from respondents. As we also disseminate information collected from administrative sources (administrative surveys) as well as from the integration of a variety of sources (such as in the System of National Accounts, which we designate as derived surveys), this official list does not cover all the "surveys" for which users require metadata. Another source that can be used to gauge the coverage of the metadatabase is the list of official data releases. While this list is broader in scope than the list of direct surveys, its unit - releases - is not necessarily co-equivalent to a survey. Many surveys could be combined in one release, or the results of one survey can be spread out over numerous releases. Products are also announced in the releases (publications, CD-ROMS, compendia) and these are sometimes the primary release vehicle for certain surveys. There is therefore a significant learning required to understand the relationship between releases and surveys to make full use of this source to assess the coverage of the metadatabase.

14. In terms of measurement error, common types of error occur in basic survey attributes and in the alignment of methodology texts and headings. Over the last year, we have detected many errors in basic survey attributes. These include the survey type (direct, admin, derived), the mandatory or voluntary nature of the survey, whether a survey is active, discontinued or one-time only and the frequency of the survey. These constitute important "tombstone" information for surveys, for which anomalies can be detected through the production and analysis of counts and cross tabulations of the metadata (for example, the number of surveys, by frequency, by activity status, by division). Knowledgeable program staff can quickly identify errors from such detailed disaggregations. The other frequent type of content error is the misalignment of methodology texts and headings. Much of the text in IMDB comes from pre-existing sources and has been converted or cut and pasted into the database. Due to a lack of understanding of the headings or an inadequacy of the guidelines, inappropriate texts were sometimes included under certain headings. Frequently occurring errors include data accuracy texts under the Quality Evaluation heading, general survey descriptions under the Instrument Design heading and just about anything under the Estimation heading. These errors can only be detected through the systematic review of IMDB records over time.

V. TIMELINESS

15. Ideally, metadata would always accompany the data it describes and there would be no delay between the availability of the data and of the metadata. However, data and metadata are frequently on different production paths and it is possible for them to become disassociated temporally. Timeliness of the metadata, therefore, refers to the extent to which its availability lags the availability of the data it describes. A measure of timeliness regarding metadata is the availability of updated metadata at the time of the release of the associated data. Failing simultaneous release, a measurement of timeliness would be the time lag between release of the data and the availability of the associated metadata. Currently at Statistics Canada, a new metatdata record is created for every instance, or cycle, of a survey. Release calendars are used as triggers for updating metadata in advance of data release and, assuming co-operation of authors, metadata are generally available on the same day as the data release, though as of yet not simultaneously with the data release. Other methods of detecting timeliness problems include the production of tabulations of the metadata by frequency and number of instances. For example, as the IMDB went into production one year ago, there should by now be 12 instances of every monthly survey on the database. If not, there likely is a timeliness problem. Maintaining the correct alignment of the vintage of the data and its associated metadata is a constant quality challenge.

VI. ACCESSIBILITY

16. Accessibility in this context refers to the ease with which users can access the metadata that supports the data they wish to use. Assuming users find the data in which they are interested, how easily can they access the necessary interpretative information? Or, if they have not found the data in which they are interested, how well can they use the metadata to help them find it? At Statistics Canada, users can access individual IMDB records on the Agency's web site through hyperlinks from CANSIM, our main online data dissemination system, and the on-line catalogue. In both these cases, the metadata are one click away from the data. It is also possible to get to the metadata from Canadian Statistics, a collection of free data tables on the web site, but it is often two clicks away, via CANSIM. Users can directly access the contents of IMDB through the Statistical Methods module of the web site, where they can access the full list of Statistics Canada surveys, arranged alphabetically or by theme. However, it is currently not possible to get from the metadata to the data. While searching the metadata can inform users of what data has been produced by Statistics Canada, it does not provide very good pointers to where these data might be found. These are elements of metadata quality that we will seek to address in the next round of enhancements to the system.

VII. INTERPRETABILITY

17. All of the work of Standards Division in the area of metadata is in direct support of the interpretability of data from a corporate perspective. Metadata must accompany data so that users can understand the sources, definitions and limitations of the data they are consulting. But what of the interpretability of metadata? At the risk of sounding too arcane, interpretability in the context of metadata relates to the availablility of meta-metadata, that is, definitions and contextual information regarding the metadata itself. For example, a convention has been adopted in the IMDB regarding the meaning of a survey. There are direct surveys, administrative surveys and derived surveys. Unless these terms are defined somewhere, they can be misconstrued by users. Currently, this type of glossary or metadata dictionary is not available to users on the Statistics Canada web site. The guidelines and instructions for authors (see Appendix 1) contain the appropriate information but would need to be reformated and loaded onto the website, with appropriate hyperlinks to make them easily accessible to users.

VIII. COHERENCE

18. Coherence, as regards metadata, can refer to the extent to which metadata can be found in one, central corporate repository, the extent to which standard definitions and concepts are used in formulating metadata, and the extent to which metadata are presented to users in a consistent, standard format.

19. For the IMDB, one of its principal objectives was to bring about greater coherence in corporate metadata by integrating pre-existing metadata systems into one, corporate repository. This has now been achieved, but unless the IMDB is actively used, and the quality of the information it contains is maintained, there will ever be a risk of new metadata repositories cropping up elsewhere in the organization.

20. As regards standard definitions and concepts, the coherence of the Integrated Metadatabase itself is assured by pinning its contents on a solid policy foundation, the Policy on Informing Users of Data Quality and Methodology and the Policy on Standards. These provide a framework for the information that is collected and maintained in IMDB. In turn, these policies reflect emerging international standards regarding metadata, as exemplified by ISO 11179, a standard for documenting data elements. Above and beyond such standards, however, many conventions need to be adopted regarding basic definitions. For example, at its base, 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, so defined, 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.

21. Conventions regarding the definition of what is a "survey", what constitutes a new version of a survey, how surveys and products relate to each other and how complex surveys are to be modelled are essential to achieving coherence of the metadata. Regarding complex surveys, we have found it useful to use an entity known as "statistical activity", which groups related surveys. An example of a complex survey requiring this kind of modelling is the General Social Survey. This is a survey that is conducted every year, in which the same content is cycled every five years. There are many ways of representing this survey in the metadatabase but we have agreed on a convention whereby the GSS would be identified as a statistical activity, comprised of five surveys, each representing one content cycle. Each survey is considered to be of quinquennial frequency, with one survey instance created every five years. There are many other instances of complex surveys, in which the concepts of statistical activity, survey, instance and survey instrument must be used in an agreed up way to represent their complex nature.

22. In terms of presentation, the summary documentation required by the Policy is to be organized according to a number of standard headings, each containing a generally 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. While metadata are presented in a coherent fashion on the STC web site, the same cannot be said across all STC products. We have not reached the point yet when the metadata contained in the IMDB is used as the single source for all published metadata, regardless of format and of medium.

IX. CONCLUSION

23. The Quality Assurance Framework, be it Statistics Canada's or any of the other variants that have been developed elsewhere, provides a sound basis for establishing criteria for assessing the quality of metadata. While the various dimensions of the framework need to be re-interpreted in a metadata context, they provide useful insights into the "fitness for use" of metadata. Based on this cursory preliminary analysis, it would seem possible to develop a systematic set of indicators that can be used to assess the quality of metadata within and across sources of statistical data.