STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 3
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Washington, D.C., United States, 28-30 November 2000)

Topic (i): Statistical metadata for dissemination

STATISTICS CANADA'S INTEGRATED METADATABASE CURRENT STATUS AND FUTURE PLANS

Submitted by Statistics Canada [1]

Invited paper

I. INTRODUCTION

1. Statistics Canada is in the second year of a multi-year project to implement a central metadata repository in support of its on-line data dissemination activities. This paper reports on the current status of this project and outlines some directions for future development.

II. CURRENT STATUS – PHASE 2

2. The second phase of the Integrated metadatabase project is now completed. Statistics Canada has designed a database in which information on all 400 of its active statistical programs is stored, as well as information on another 400 discontinued programs. The database was implemented in Oracle 8 and is resident on a central server. Metadata was collected from a variety of pre-existing metadata stores, reformatted 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. They can be accessed through hyperlinks from our output online database, known as CANSIM 2, or 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, implemented in Java, and deployed over the Intranet to the desktops of the authors in the data producing programs. Updates are quality assured and registered before being made available for generation of the external HTML pages.

3. The content of the metadatabase has been selected to suit its primary purpose, which is to provide users with the information required to interpret the statistical data we disseminate. The IMDB data model was inspired by much of the previous work in this area, in particular by the work of Dan Gillman at the United States Bureau of the Census. It can be thought of as consisting of two regions. The first, the statistical information region, defines all of the entities and the relationships among them required to describe Statistics Canada's statistical programs, their content and their methods. Each of these entities is referred to as an administered component. As administered components, they all share an identical set of relationships and entities for the stewardship of the component. (ANSI X3.285:1999). For each administered component, this stewardship region includes such information as theme, topic, keyword, time frames, contact, organization and documentation

[1] Prepared by Paul Johanis.

4. The main headings under which information has been collected for Phase 2 are shown in Table 1. The indentations of each entry in the table illustrate the hierarchical relationship between the entities. The definitions for each entity on the list are provided in the pages following the table.

Table 1 - IMDB headings

Statistical activity
Survey
    Universe
    Data element concept
        Property
        Statistical unit
    Frame
Survey Instance
    Survey instrument
        Survey instrument image
    Methodology summary
        Questionnaire design
Sampling plan
        Collection and capture method
        Editing procedures
        Imputation method
        Quality control procedures
        Estimation method
        Time series processes
    Disclosure control method
Quality evaluation method
    Quality measures
Data files

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

    Time frame
    Theme
    Topic
    Keyword
    Organization
    Contact
    Additional documentation

6. Definitions:

Statistical activity: An organizing concept that groups surveys that share a common processing system or conceptual framework. Examples include Vital Statistics or the System of National Accounts. Not every survey is necessarily part of a statistical activity.

Survey: A statistical data producing activity administered by the statistical agency. Includes direct surveys, in which data are obtained directly from respondents, administrative surveys, in which data are collected from administrative data sources, and derived surveys, which make use of existing direct and/or administrative surveys to create new statistical information.

Universe: The set of units about which statistics are to be produced by one or more surveys.

Data element concept: A description of the meaning of a piece of data.

Property: A characteristic that can vary across the members of a population.

Statistical unit: The definition of the unit about which data are collected, for example persons, households, businesses, crimes, etc.

Frame: The file from which units to be surveyed are selected.

Survey Instance: Each cycle of a survey.

Survey instrument: In the case of a direct survey, the questionnaire; for an administrative survey, the record layout of the input record; not applicable in the case of a derived survey.

Survey instrument image: In the case of a direct survey, an image of the questionnaire(s).

Methodology summary: A text description of each of the following aspects of the methodology of a survey.

Questionnaire design: The method used to design and test the survey instrument.

Sampling plan: Sample design and sampling method

Collection and capture method: Includes coding procedures where applicable.

Error detection procedures: Methods used during collection, capture and processing to identify errors in the data.

Imputation method: Methods used to correct for error or missing values.

Quality control procedures: Procedures used to monitor and measure errors committed during collection and capture

Estimation method: Method used to produce estimates from collected data.

Time series processes: Methods used to adjust estimates in relation to same estimates for prior periods. Includes seasonal adjustment, calendarization and benchmarking.

Disclosure control method: Methods used to modify the data so as to maintain the anonymity of respondents.

Quality evaluation method: Methods used to measure the quality of the data that are produced in terms of relevance, timeliness, accuracy, coherence, accessibility and interpretability.

Quality measures: The measures of data quality that result from the quality evaluation methods, for example in the case of accuracy, the coefficients of variation for the main variables in a survey.

Data files: Information on the location, format and content of the clean record masterfile for a survey.

Time frame: A time reference that is relevant for the specified component. Includes periods (reference period, collection period) and dates (effective date, start and end dates).

Theme: A broad topic that describes the subject matter of the specified component. Used for searching the database in a browse mode. Selected from a standard list.

Topic: A more specific category within each theme that describes the subject matter of a given component. Used to search the database in a browse mode. Selected from a standard list.

Keyword: A word or phrase that describes a given component. Used for keyword searches of the database. Selected from a standard list (thesaurus).

Organization: Information on organizational units and their role in relation to a given component. Selected from a standard list.

Contact: Name, phone number and email address of persons who can contacted regarding a given component. Selected from a standard list (departmental directory).

Additional documentation: URL of additional documentation related to a given component.

III. CURRENT DEVELOPMENT – PHASE 3

7. In the next phase of the project, planned for completion in March 2001, the database content will be expanded to include information on the variables collected in each survey, their definition and classifications. This is an extension of the information included in the Phase 2 database under the heading Data Element Concept.

8. As shown in Table 1, the data element concept is a definition that comprises two parts: an object class (known in our model as a statistical unit) and a property. In a statistical agency, there is a finite list of objects about which data are produced and this list of items, each with its definition, will be loaded in the database and made available for selection by author divisions. The list of properties, or variables, is much more extensive. Many are standard, however, and can be loaded at the time of the initial load of the database. The act of selecting a statistical unit in association with a property will create a data element concept. This provides the definition for a data element, which is the data item about which metadata is to be provided in the first place. In the data model, the data element is a map relating a survey instance to a data element concept. Taken together, the metadata for a data element contains the following information: in a given instance of a given survey, this data element describes property x for statistical units y in universe z. Translated to a real case, this would read as follows: in the October cycle of the Labour Force Survey, this data element describes the gender of persons in the labour force. The data user can obtain further information on each part of this statement: what do we mean by the October cycle? What is the Labour Force Survey? What do we mean by gender? How do we define persons? What do we mean by the labour force? The answers to these questions are the kind of statements we want the Integrated Metadatabase to provide. Most can be provided by the content of IMDB Phase 2. However, the question, “What do we mean by gender?” can only be answered by Phase 3 content.

9. In addition to the definition of the data element, a full answer also requires the list of values that the data element can assume and their meaning. In the cases of gender, for example, those values are “male” and “female”. In the model, the set of values that a data element can assume is known as a value domain. These are divided into non-enumerated and enumerated domains. The former correspond to data elements that are continuous variables. This means that these variables can assume any value in a set of rational numbers. Data elements such as income, sales, revenues or age have non-enumerated values domains. Statistical data are rarely, if ever, disseminated in this form. Usually, data collected as continuous variables are converted to categorical variables for dissemination purposes, for example age groups, or income ranges. Other data are collected directly as categorical variables, for example gender, mother tongue or industry. The set of values that this type of variable can assume is known in the model as an enumerated domain. In statistical agencies, enumerated domains are often organized into classifications, that is, arrangements of classes that are mutually exclusive, exhaustive of the universe of interest and that can be aggregated into successive levels in a hierarchy. As statistical data are almost always disseminated according to given classifications, standard or otherwise, data users need to know the meaning of each category or class within a classification and the hierarchical relationship between

them. In IMDB, every enumerated domain will be named, each of its constituent classes will be defined and the level and position of each in a hierarchy described.

10. In most cases, even the most disaggregated level in a value domain that is organized as a classification is itself an aggregate of more detailed descriptors. The set of all such detailed descriptors ("value meaning" in the data model) forms the conceptual domain of a value domain. Many value domains may be formed from the same conceptual domain. The data element "industry", for example, has many enumerated value domains, or classifications, including the ISIC, NACE and NAICS. All belong, however, to the same conceptual domain, which is made up of many thousand descriptions of economic activities, which will also be stored and managed through the IMDB. It is through these detailed economic activities that the correspondence between classes of each of these classifications can be established.

11. Once loaded into the IMDB, searching and linking to Phase 3 content will be provided to data users on the Statistics Canada website. In addition to the themes, topics and keywords loaded in Phase 2 IMDB, each data element concept will become a search term in Phase 3. A search on "mother tongue", for example, will return all surveys for which data using that data element concept have been disseminated. Linking to Phase 3 content will be provided either through a direct search of the metadatabase or through hyperlinks from CANSIM 2, the principal on-line output database at Statistics Canada or Canadian Statistics, a collection of preset statistical tables on our site. In the case of CANSIM 2, links to the IMDB will be provided during the table specification dialog. For example, if a user wishes to include the data element "mother tongue" in a custom table that he or she is specifying online but is not sure of the definition or of the values it can take, a hyperlink to the IMDB will provide the answers. In the case of Canadian Statistics, the definitions of the data elements included in each table will be a click away. With respect to public use microdata files, the IMDB will be used to store record layouts and to generate codebooks, according to emerging standards such as the DDI or others. This is part of the basic information service that the IMDB is designed to provide. Having built such an infrastructure, however, opens the door to other uses of the metadatabase.

IV. FUTURE DEVELOPMENTS - PHASE 4 AND BEYOND

12. Metadata can support at least three broad functions within a statistical agency: data dissemination, data production 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. It can, however, in its current design or with some extensions, support data production and management objectives.

13. With respect to the latter, as mentioned earlier, many of the definitions of properties, statistical units and value domains recorded in IMDB will be standard, in the sense that their use within the agency is widespread and customary. Many others, however, will not. Statistics Canada has a Policy of Standards, which establishes an objective and a process for using coherent and standard definitions of concepts, variables, classifications, units and populations across the agency. Each of these entities is represented in the IMDB data model, which also provides for recording whether a definition is a departmental standard, a recommended standard or a program specific standard, the three levels of standards recognized in the Policy.

14. The IMDB can therefore also be used in its current state as a tool for evaluating the extent to which standard definitions are used in the statistical programs, for identifying opportunities for further harmonization and for promoting the use of departmental standards. In this respect, one of the known obstacles to greater coherence is simply a lack of information regarding departmental standards in author divisions. As a result, many areas may be "re-inventing the wheel" when they would be perfectly happy to use pre-existing, standard definitions for new or existing survey programs. A fairly large number of standard definitions have been approved under the Policy and it is believed that by loading these upon the

initial load of IMDB Phase 3 and making them available for selection by author divisions, the extent of use of standards will increase throughout the agency.

15. Other aspects of the management of the statistical system can be performed using the IMDB as currently designed. For example, it has incorporated the data requirements of an earlier metadata store, which it now replaces, the Meta-Inventory of Data Assets (MIDAS), which was used by the agency to fulfil regulatory obligations regarding the archiving and storage of data banks containing personal and business information (clean statistical data masterfiles).

16. The IMDB can also currently support the maintenance of an inventory of software and application systems used in the various stages of collecting and processing survey data. This could permit the analysis of software diversity throughout the agency, which would help to promote coherence and reuse of software components across survey programs.

17. The current design would also support to production on a regular basis of standard reports on various aspects of data quality associated with the agency's programs, including the production of time series. For example, in reference to the timeliness dimension of data quality, it is now possible to analyze the evolution of the elapsed time between reference periods and release dates over time, across all programs or for specific programs. Quality measures related to accuracy are also stored and could be analyzed over time and across programs. Survey performance measures such as response rates, refusal rates and frame adequacy are also available. With very little extension to the model, it would be possible to track and to analyze over time response burden and survey overlaps.

18. There are many other management uses to which the metadata stored in IMDB can be put, limited only by the imagination and ingenuity of the managers themselves. This metadata has just as great potential however for supporting the third broad function mentioned above, data production.

19. A number of services could be operated from or through the IMDB in support of various phases of the production of statistical data. One of the first that will be developed is a coding service. The value domains and related conceptual domains provide the required metadata infrastructure to support computer assisted manual coding and also automated coding systems. For example, an industry classification coding system would use the descriptions of economic activities stored as value meanings as a reference file against which response phrases to be coded can be matched. A positive match would lead to the assignment of an industry code from any of the various value domains supported by the system.

20. This same infrastructure can be used to produce concordance tables between value domains. Using the industry example, tables relating the current vintage of NACE and NAICS to each other could be produced automatically.

21. Basic range and validity edits are other data production support activities that can be supported from Phase 3 content with no extensions. Instead of downloading or otherwise creating look–up tables for such edits, it would be possible to consult the IMDB directly, thereby eliminating the potential for error

22. Beyond these activities, extensions to the IMDB content, but still within the data model, would be required to support other data production activities. In the immediate future, Statistics Canada is committed, as part of the Government On-Line initiative, to expanding its data collection activities over the Internet. The metadata for online collection could be stored and managed through the IMDB. First, the data model provides a framework for recording information for all survey programs, over time. In other words, a full metadata history for all survey programs will be maintained. This provides the basis for reuse of survey instruments or parts thereof. Secondly, the model currently includes survey instrument, survey questions and response categories, though his aspect has not been developed for Phase 3. Thirdly, definitions of variables and value domains, which would form a significant portion of response guides and user help files, are currently stored in IMDB. It would require little extension to the

data model and the creation of new interfaces to collection systems to bring about the integration of IMDB in support of on-line data collection.

23. In the longer term, the IMDB could support other aspects of data production. One strategy might be to associate the IMDB with existing or emerging production systems. For example, the Unified Enterprise Survey Program in Statistics Canada is in the process of developing a centralized survey processing system for much of the agency's business statistics program. These systems make use of metadata throughout their operation, and, in the absence of a corporate metadata system, these have had to be built in within the processing systems. Now would be the time to examine the potential interface between this production system and the IMDB. Another strategy involves generalized systems. Statistics Canada has over time developed a suite of generalized systems, which support different phases of the survey process. Each of these is parameter driven. These parameters are in effect metadata. It would therefore be possible to imagine a scenario in which a generalized system is invoked, and the parameters for its operation provided, through the IMDB interface. For example, to draw a sample, the Generalized Sampling System (GSAM) requires the following information: the type of sample unit, the target universe, the survey frame from which units are to be selected, the sample size, the sample plan, the sample selection method. Almost all of this information resides in Phase 3 IMDB. Only the sample selection method would require extension, based on work that has already been done by survey methodologists assigned to the project team. With an appropriate interface between the IMDB, GSAM and one or more of the central survey frames (for example Statistics Canada's Business Register), it would be possible to execute a metadata driven sampling process. A major advantage of this approach from a metadata perspective is that the actual specifications for the process become the documentation of the process. In other words, active metadata replaces passive metadata. Similar scenarios could be built for the other generalized systems, such as the Generalized Estimation System and the Generalized Edit and Imputation System.

24. A broad system architecture study that situates the IMDB, corporate registers and survey systems and business processes within a web-enabled environment is currently underway. If all goes well, it should provide a road map, or a town plan, for the integration of these corporate resources over the next several years.