Metadata at Statistics Canada: implementation opportunities and challenges

Michel Cloutier[1] and Alice Born[2]

Introduction

The Integrated Meta Data Base (IMDB) is a corporate repository of information on each of Statistics Canada's nearly 400 active surveys. The IMDB was first developed to support the interpretation of data disseminated by the agency and to present the metadata in a consistent manner. The IMDB is one of the principal mechanisms used by Statistics Canada to fulfil the requirements of the policy on informing users of data quality and methodology. In fact, the metadata contained in the IMDB are a critical component of the agency's communication strategy.

For those outside Statistics Canada (data users), access to metadata is needed primarily to understand the data and surveys produced by the agency. The IMDB is the primary source for the information they need to interpret the statistical products published by Statistics Canada. This information includes a description of data sources and methodology, definitions of concepts and variables and indicators of data quality. For internal users, the IMDB serves as a source of information to support knowledge management, the development of survey content, the management of surveys, classifications, coding and a variety of other statistical functions in addition to dissemination.

The role of the IMDB has continued to evolve as more and more uses are made of this central infrastructure system. These new uses leverage our investment in the IMDB and provide added value to the organisation, while eliminating the need to develop new software systems to document all the aspects of the national statistical system. This paper will outline how this evolution has occurred and some of the challenges and opportunities that national statistical offices (NSOs) face with regard to documenting information about their data and their statistical programmes.

Metadata at Statistics Canada

Metadata have always existed in Statistics Canada, but our approach was at first very much less structured than it is today. Before the age of the computer, a limited amount of documentation regarding surveys was most often published, with the data, in notes at the end of paper publications. As electronic data became more popular, documentation was sometimes lacking and was certainly not kept systematically in a central repository. The meta information was often not easily available to most users. With the advent of the internet, it soon became evident that users needed convenient online access to metadata to help them interpret published data.

The IMDB is the descendant of previous initiatives used to document our survey activities. It started in 1998, partly in response to observations made in a report by the Auditor General of

[1] Statistics Canada, Ottawa, Canada K1A 0T6. E-mail: michel.cloutier@statcan.gc.ca.

[2] Statistics Canada, Ottawa, Canada K1A 0T6. E-mail: alice.born@statcan.gc.ca.

Canada. In this report, it was underscored that Statistics Canada should put greater emphasis on providing quality documentation of survey processes and programmes. Before 1998, Statistics Canada had a number of databases and systems for storing metadata. The first step in the creation of the IMBD was therefore the consolidation of existing data stores into one central store.

The following systems were retired and their metadata consolidated into the IMDB:

- Statistical Data Documentation System (SDDS), consisting of primary textual descriptions of the statistical business processes
- Meta Inventory of Data Assets Systems (MIDAS), consisting of metadata describing the confidential master data files
- Thematic Search Tool and Paradox Meta System for Social Statistics
- Questionnaire Inventory.

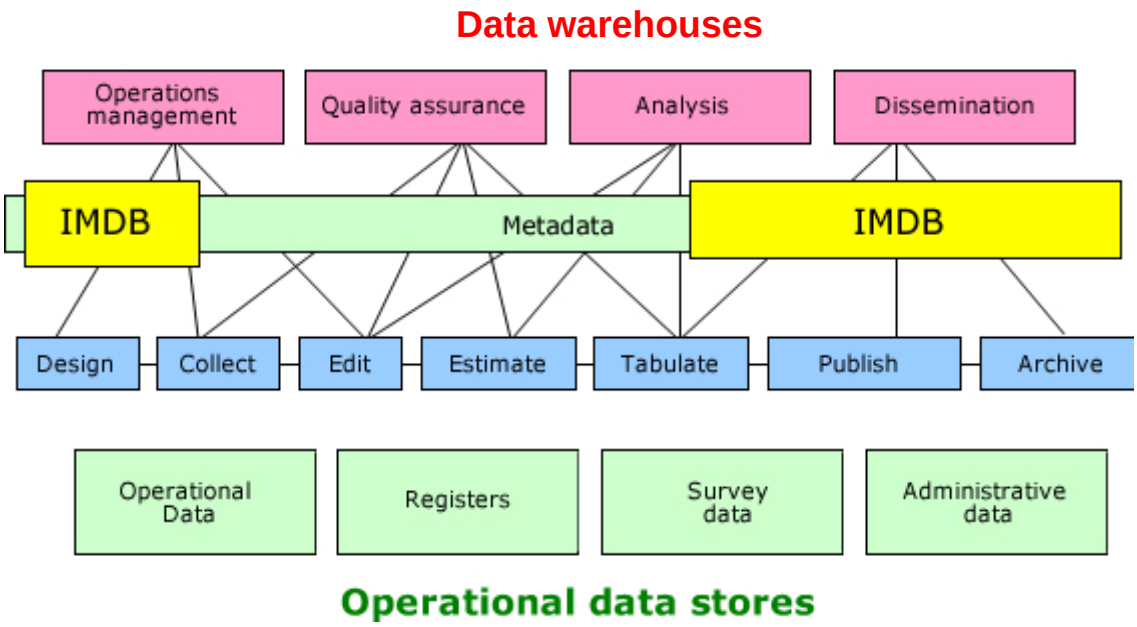
Today, the IMDB information is made available to outside users on the Statistics Canada website. Every data release on the site (in the Daily) includes hyperlinks to metadata from the IMDB. In addition, the Statistics Canada online database (CANSIM), containing millions of statistical time series and data tables, is also linked to the IMDB. The IMDB information can also be accessed directly through the “Definitions, data sources and methods” module on the website. The database is kept up to date through an input system deployed over the departmental intranet and also using the STC Wiki application. Updates are quality-assured and registered before being made available to the external website. The potential also exists for metadata stored in the IMDB to be exported to other meta information systems in any output format that suits these other systems (ie SDMX, DDI, HTML, Wiki).

The short history above illustrates the practical approach that Statistics Canada has used by seizing opportunities and responding to challenges in order to gradually develop its metadata systems and content. Figure 1 illustrates how metadata in the IMDB currently support the statistical system. While the metadata layer extends across all phases of the statistical business process, metadata in the IMDB currently support the design, publishing and archiving phases, and some of the collection and tabulating phases. Metadata for the other phases of the survey life cycle (such as data production) occur in other meta information systems in the agency. The consolidation of these other systems represents an opportunity for further expansion of the IMDB.

The opportunity for increasing the centralisation of metadata management in the agency has occurred through pressure from external stakeholders, the need for survey documentation both internally and externally, and the urgency of updating existing systems that had come to the end of their life cycle. Further examples of this approach are documented in the following section.

Figure 1

The role of the IMDB in the survey life cycle



Opportunities

The scope and use of the IMDB has grown extensively over the years. Although this growth has been planned and managed to meet the needs of the agency, it has also occurred in part owing to events that were not foreseen when the system was originally conceived. Examples will be presented to illustrate the advantages of remaining vigilant and flexible when faced with these opportunities. Statistics Canada has been well served by not limiting the content and functionality of the IMDB to the original requirements and specifications of the system.

Dissemination

As stated previously, the first driver for the IMDB was the need to provide information to data users as part of our dissemination activities. In the dissemination phase, the IMDB is the source of summary texts describing surveys and statistical programmes, their methodology, the quality of the data produced and the definition of the variables they measure, as well as of the images of the questionnaires used in the survey. The content of the IMDB is reused wherever possible to document methodology and data quality in electronic and paper publications. This information is essential to users when they are analysing data from surveys so that they can determine whether the data are fit for their use. This includes analysis of aggregate data as well as public use micro data files (PUMFs) utilised by more sophisticated users, such as university researchers. Documentation for Statistics Canada's Data Liberation Initiative (including PUMFs) uses IMDB metadata reformatted under the Data Documentation Initiative (DDI) standard.

Collection

During the collection phase of the survey, there is often a demand by respondents to confirm the authenticity of the survey they are being asked to complete. By using the IMDB through

the website, we have been able to fill this need. The “Information for survey participants” module on the Statistics Canada website uses the IMDB to provide respondents with basic information about the survey. The module links to the IMDB to display the images of questionnaires as well some of the descriptive text about the survey.

Adopting standards and reducing diversity

The IMDB is becoming a great tool to help us reduce diversity in the statistical system. It is helping us to first document and then control the number of variants we have in terms of common variables, procedures, classifications and systems. When all documentation is stored in a central database, the diversity of various elements of the statistical system becomes evident very quickly in the metadata.

One of the clear benefits of a centralised metadata store is the opportunity to improve data quality (coherence, interpretability, accessibility). The IMDB has allowed us to get a more complete picture of what is being released in our many surveys. For example, we have now documented the large number of variants of the industrial classifications used in the time series published by Statistics Canada. We are now working on reducing the number of variants. Thus the IMDB is enabling us to promote greater standardisation of the classifications used in our surveys. This will allow us to improve coherence and interpretability in the medium to long term.

Another good example of this is our content harmonisation project for household surveys. This project has been an important strategic opportunity to help us expand the use of the IMDB and further metadata implementation in Statistics Canada. Currently, the IMDB intervenes mostly at the analysis, dissemination and post-survey evaluation phases of the statistical cycle. However, with this project we have started using the IMDB for the survey planning and design phase. During this phase, the IMDB can be consulted and used by survey managers to identify existing variables for reuse and to support the development of new questionnaires.

The objective of this project is to define standard concepts and variables to be measured by new or redesigned surveys, as well as to create standard questions and question blocks to collect the data necessary to measure these concepts. The naming of the variables is being done jointly by the IMDB and household surveys staff according to the IMDB naming convention. The Statistics Canada Wiki (a collaborative authoring tool) allows survey managers an easy way to document new variables during questionnaire design and to store these variables and the related questions. Other systems can then use this information as inputs to avoid duplication of metadata in subsequent steps of the statistical process such as collection or dissemination. This metadata will also allow us to better support multimode questionnaires and questionnaire automation, since all collection systems will access the same source for survey questions. This integrated approach to metadata (eg concepts, variables, classifications, questions, question blocks and response choices) will allow the entire statistical system to become much more efficient and effective.

Other information systems

The metadata associated with various phases of the statistical process are managed by multiple systems distributed throughout the agency. The IMDB serves as the source for variables and classifications associated with many of these systems and functions. This is particularly the case for data warehouses, which support the aggregation and analysis phases of the survey life cycle. Data warehouses are being developed across Statistics Canada including in the system of national accounts (SNA), education and health fields. The design and architecture for these projects include links with the IMDB as the authoritative source for metadata. This saves costs, avoids duplication and ensures coherence within the statistical system.

One of the challenges faced by the warehouse architects was how to provide warehouse users with access to metadata in a way that allows them to navigate quickly through the rich metadata environment and submit updates to Standards Division related to metadata that are incomplete or inaccurate. The Data Warehouse Centre resolved this issue by leveraging the Statistics Canada Wiki application to act as a link between the IMDB and the Data Warehouse Framework.

The Government of Canada has implemented, in consultation with departments and agencies, a common look and feel (CLF 2.0) for all federal internet/intranet sites and electronic networks. As part of this policy, there are accessibility rules that no longer allow the use of PDF files on government websites. In particular, at Statistics Canada, survey questionnaires have historically been presented to users in PDF format. To meet the new requirements of CLF 2.0 the IMDB will be used as the source for new XHTML versions of survey questionnaires. These questionnaires will therefore be completely accessible.

Information management

As part of Statistics Canada's information management plan, we have recently decided to use the IMDB as the source for metadata when archiving statistical data. To ensure that archived data is usable by future generations, we need to include not only the data themselves but also information such as the survey objectives, definitions of variables, record layouts, questionnaires, quality indicators, classifications and methodology. The IMDB is the ideal source for this information. By using the IMDB for this purpose, we are again avoiding recreating another system and another database for this particular phase of the survey cycle.

Document management (paper and electronic documents) at Statistics Canada brings together all documents relating to a particular survey from its conceptual development phase to its dissemination phase. The Document Management Centre uses classification numbers from the IMDB to organise all of their files, thus ensuring consistency with other sources for this metadata repository. The documents stored include the following: questionnaires (including test questionnaires), methodology documentation (scientific, technical, operational), correspondence and internal memoranda related to statistical activities (scientific, technical, operational), statistical analytical papers (analysis and quality measurement), promotion and marketing material, reports on survey costs, etc.

Planning and management

We continue to look for strategic opportunities such as the redesign of systems to further promote the use of the IMDB. Every year Statistics Canada goes through a long-term planning (LTP) exercise where new projects and programmes are proposed to meet new user needs or to address problems with existing programmes (ie improving quality and relevance, upgrading existing systems, etc). Many of these proposals are designed to produce efficiencies over the long term (savings that can be reinvested in the statistical system). In many cases, the IMDB is allowing us to avoid developing new systems to meet the metadata needs of these LTP proposals. The following are examples of the development projects that have used the IMDB.

As previously stated, the IMDB was originally designed to be a repository of the metadata describing our surveys and survey outputs. However, since metadata by definition can be collected on any group of objects, Statistics Canada has begun to expand the scope of the IMDB to be a repository of objects other than our data outputs. In particular, we have begun to view structured information describing our management information, IT systems and enterprise architecture as metadata.

It is possible to expand the scope of the information housed in the IMDB because the IMDB conforms to a strongly defined metadata standard, ISO/IEC[3] 11179 metadata registries. Although this standard was designed specifically to support metadata on data, the standard can easily be applied to any group of objects for which metadata are being collected. One such application of the IMDB has been to replace an ageing system known as the Statistics Canada Software Register (SR). The SR was a database containing a list of all software used and developed in Statistics Canada along with support levels and dependencies for survey programmes. This information will now be stored in the IMDB, eliminating the need to redevelop and maintain a separate system.

Another example of this type of implementation within the IMDB is the documentation required for the Government of Canada's Management Accountability Framework (MAF). With the MAF reports and indicators entered and stored in the IMDB, our corporate planning and evaluation programme can take advantage of the functionality in the IMDB such as time travel, metadata classification and registration. This application helps the agency accumulate information to report to central agencies under the MAF while linking the management framework to metadata on survey and statistical programmes.

In addition, we are planning to use the IMDB to store management assessment information related to the ongoing Quality Assurance Reviews of surveys and statistical programmes.

Classification management and coding

Recently, after reviewing our requirements and completing a business case, the agency decided that it needed to centralise and rationalise the management of classifications and computer-assisted coding. It was decided that, as part of this new system, we would leverage the IMDB infrastructure to house the required information. The initiative will reduce the number of individual coding systems across the Bureau as well as provide a computer-based coding tool for divisions where manual coding (using paper versions of classifications) still occurs. The new interactive coding tool will use the redesigned Automated Coding by Text Recognition (ACTR) system as the search engine, and support a centralised set of reference files for several classification domains, thus improving the quality of coding activities and coherence of data throughout Statistics Canada. The generalised interactive coding tool will be offered as a web service, which will make it accessible to any computing platform. This should reduce systems costs for surveys requiring access to an interactive coding tool, and reduce training costs as the agency moves towards increased centralisation of coding in its new collection model.

A generalised coding tool with a standard, centralised set of reference files will enhance the coherence of the agency's statistical outputs since programme areas will be able to interface with the Classification Coding System, and its reference files, through a web service. This will allow access to the most up-to-date version of reference files that have been coded, approved and registered by Standards Division and user groups. Standardising reference files should improve the rate and quality of both automated and interactive coding data. The redevelopment of these systems has been a perfect opportunity to further expand the scope and usefulness of the IMDB.

---

[3] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC).

Challenges and implementation issues

Development and maintenance of metadata

One of the greatest challenges related to the implementation of metadata systems in statistical organisations is ensuring that survey managers prioritise the development and maintenance of metadata (content) for their surveys. With the time and resource pressures faced by many programmes, taking time to understand how to update the IMBD and complete documentation is often a much lower priority than releasing survey results and dealing with quality issues.

Therefore, ensuring that metadata are up to date, of good quality and useful to users must be made a priority by senior management. It also helps if the metadata are used as an integral part of the data release, as they are with the Daily releases by Statistics Canada. When Statistics Canada first implemented the IMDB, the Chief Statistician made it a priority for all programmes to provide accurate and up-to-date metadata. The agency quickly followed up with areas that were deficient in this regard. The result was a marked improvement in the quality and quantity of metadata in the IMDB.

Survey managers now see the usefulness of accurately maintaining metadata since they are immediately available to their clients when they view new survey results. It is also critical that we create a user-friendly environment that allows easy updates. The key is showing survey managers that metadata have value to their users and are a critical component of communicating with them. Successful implementation depends on increasing the importance of metadata as an integral part of the statistical process. Metadata need to be shown to have a practical value for analysts. They must help them to convert statistics and numbers into meaningful information. As described above, once the metadata are stored in the IMDB, they can be reused for many purposes in the statistical system. Internal and external analysts (re)use metadata in questionnaire design, data tables, publications, data warehouses, micro data files and data archives.

Standardisation and centralisation

When the IMDB was first implemented, significant effort had to be put into improving the quality of the metadata. This was fully supported by senior management and has resulted in good “buy-in” from survey managers. Now that we have developed the content of the IMDB, the challenge is to enhance and maintain the quality of the metadata themselves. Getting programmes to use common definitions and tools is one way to make it easier to improve the quality of our metadata. We are now encouraging the use of generic concepts instead of survey-specific ones and up-to-date revised classifications instead of continuing to use historical versions of these classifications. Nevertheless, there is still widespread use of variants of our industry classification and multiple definitions of common variables and concepts. The issue of maintaining historical continuity makes it difficult for survey managers to make the change in many cases. The use of concordances and historical revisions helps to facilitate transitions, but it still requires a great deal of time and resources to move to more current standards.

Getting survey programmes to understand and use metadata systems, content and tools requires training and communication throughout the organisation. Using new technologies such as the IMDB’s MetaWeb and Statistics Canada Wiki helps to reduce the learning curve and makes updating content much easier for survey managers. More and more, the IMDB team is permitting survey managers to enter their own metadata updates directly into the IMDB using the MetaWeb interface, thereby allowing staff in the IMDB team to focus on the quality of the metadata. Once entered, survey managers can immediately see their metadata in the IMDB portal on the Statistics Canada Wiki. The advantage of the wiki technology is

that internal users can see all of the metadata stored in the IMDB and then develop a customised metadata presentation depending on their requirements.

Metadata management and governance

Standards Division in Statistics Canada is the principal area in charge of supporting and developing statistical metadata for the agency. This work is also guided by the Methods and Standards Committee. Survey programmes are responsible for keeping the content on the IMDB up to date. Despite this organisation of responsibilities, individual survey areas often find it challenging to adopt and converge on revised standards and classifications. The use of multiple variants of variables, concepts and classifications makes coherence and interpretability difficult among various survey outputs. It is also very difficult for Standards Division to monitor all of this activity with a very limited budget. The challenge is to find the right balance among all of the stakeholders in achieving the best quality possible for the agencies' metadata.

Standards Division is currently developing a proposal that will outline roles and responsibilities for the central functional areas (such as Standards Division and the IMDB) versus those of the subject-matter areas. The IMDB is helping the agency to distinguish between the functional responsibilities of each division as opposed to its corporate roles as they relate to metadata and their contribution to the statistical system. To improve the coherence of metadata, including classifications and variable definitions, across the statistical system, we are looking at creating a Standards Governance Board with a mandate to examine and monitor the implementation of standards across survey and statistical programmes. Membership will be drawn from survey programmes, registers (eg business register and tax data) and statistical programmes such as the System of National Accounts. The objectives will be to increase awareness of the challenges of incoherence in the system and to develop a planned approach for implementing coherent metadata.

As part of the harmonised content for household surveys project, we have already seen the benefits of a corporate approach to developing and approving standardised definitions for variables and classifications, and their related standardised question and response choices. For example, we will have only one definition for household income and one set of questions that will be used on questionnaires to measure household income. This should lead to greater efficiencies by reducing design and processing costs, and to better coherence since the concept will be the same across most household surveys.

Conclusion

Metadata management and governance in statistical institutions are important issues that can have a significant impact on the entire statistical system. Statistics Canada has taken a pragmatic approach to metadata implementation by looking for strategic opportunities – redesign of collection systems, survey systems and information management, as opposed to a complete redevelopment of our statistical system. This approach has allowed the agency to cost-effectively gain many benefits, including some that were unexpected, from a centralised metadata store. The integrated approach of the IMDB has been very effective in promoting and creating better metadata for all of Statistics Canada's programmes (surveys and management information). We have also been successful in developing an environment that promotes efficiencies through the reuse of content, improved data quality and better information management.

References:

Bellerose, P-P, T Dunstan, C Greenough, A Lee and A Born (2007): “Case study – Canada: UNECE Workshop on the Common Metadata Framework”, Vienna, 4–6 July.

Born, A, and T Dunstan (2008): “Metadata requirements for archiving structured data (METIS)”, Luxembourg, 9–11 April.

Doherty, K (2008): “Metadata architecture at Statistics Canada: Meeting on the Management of Statistical Information Systems (MSIS 2007)”, Luxembourg, 7–9 May.

IMDB (Statistics Canada website) (www.statcan.gc.ca).