**Metadata in the modernization of statistical production at Statistics Canada**

*Carmen Greenough, Kaveri Mechanda and Flavio Rizzolo, Statistics Canada*

**1. Introduction**

Statistics Canada has started the modernization of its statistical production through a corporate business architecture approach. Four of the architectural principles of this transformation of particular significance to statistical metadata are: metadata-driven processes; maximizing reuse of existing corporate systems; enforcing reuse of concepts and classifications and governance of data and metadata architectures.

These principles are achieved by accessing statistical metadata in the Agency’s corporate integrated metadatabase – the IMDB. The IMDB is a well-structured and authoritative source of conceptual and other types of metadata such as referential and exchange metadata. Its structure is based on ISO/IEC 11179 *Metadata registries.* Building web services to exchange information among metadata repositories using common exchange models that align with the Generic Statistical Information Model (GSIM) is a prime focus for current development and enhancement of the IMDB. The lack of consistency and standard structures across repositories present challenges to integration. GSIM provides the framework to establish coherence, reduce duplication and support interoperability between and among these data and metadata repositories.

This paper describes the advantages of a centralized metadata registry in promoting the architectural principles presented above and coherence in the Agency’s statistical programs. Section 2 provides an overview of metadata modernization at Statistics Canada. Section 3 focuses on the IMDB metadata repository, and explains the advantages of having a centralized metadata repository, particularly for conceptual metadata. Examples of integrating metadata via services are presented in Section 4. We conclude in Section 5.

**2. Metadata modernization**

Statistics Canada is undergoing an agency-wide initiative to modernize its statistical production. A review of how the Agency manages its statistical metadata and its metadata architecture is part of this initiative. A wide range of statistical metadata is created, used and shared in national statistical offices. Statistics Canada recognizes the importance of managing the metadata that is created and used, as shown by the local processes to manage metadata that are in place within different areas and systems. The intent of the metadata modernization initiative is to establish common semantic structures, robustness of systems and efficiency of statistical production through a corporate business architecture approach. The next sections will describe how Statistics Canada has taken a multi-faceted approach to metadata modernization.

*2.1 Working with international organizations on metadata management*

Statistics Canada has been a key player in the recent international work in the development of GSIM. [1] GSIM is the result of a collaboration involving statistical organizations across the world in order to develop and maintain a generic reference model suitable for all of the organizations and meet the strategic goals (in particular the modernization effort) of the statistical community. This generic reference model is now being used to establish common semantic metadata structures and exchange of metadata in the Agency, which allows for more efficient exchange of data. The Statistical Classification Model of GSIM is being used as a conceptual model for the exchange of classifications among corporate metadatabases, resulting in improved coherence in how the statistical classifications are structured and managed within the Agency. The corporate metadatabase at Statistics Canada (described in Section 3) was developed in alignment with international standards (in particular, ISO/IEC11179 *Metadata registries*). The development of GSIM, with its common structure and semantics, provided the *lingua franca* to successfully integrate the corporate metadatabase with the diverse and heterogeneous set of metadata stores across the Agency for centralized registration, management and harmonization. This GSIM-based integration is described in Section 4.

Statistics Canada has also been heavily involved in the development of the Common Statistical Production Architecture (CSPA) [2] and its implementation. CSPA provides a framework to develop metadata-driven processes and systems based on GSIM and other international metadata standards to facilitate interoperability and reuse. More details about the Agency involvement in this project can be found in Appendix I.

*2.2 Agency-wide approach to management of business architecture*

Statistics Canada has taken an agency-wide approach to the management of its business architecture. This initiative at Statistics Canada has set out principles and direction to ensure efficient, reliable and responsive statistical data and processes. Some of the basic architectural principles of particular significance to statistical metadata are:

* Metadata driven: This principle specifies that metadata should be an integral part of the process and precede the data. In other words, statistical metadata, from survey specification, sample design and edits to products and services, must be captured, managed and used to drive the entire statistical process.
* Maximize re-use: This principle is driving the organization towards common business processes and enabling computer systems. To take full advantage of this, statistical data and information structures supporting the metadata should be re-used.
* Governance: Statistics Canada has identified the need for corporate level thinking about architecture, data management and metadata that spans the statistical business process. This identifies metadata management as a key element in the statistical business process.

This agency-wide vision and approach to metadata management has resulted in (1) initiating a strategy and a project specifically mandated to propose and implement activities for the modernization of metadata; (2) focusing on ways to better manage metadata and finding efficiencies as part of new project and initiatives and (3) enhancing existing metadata management structures and databases.

*2.3. Agency-wide approach to metadata modernization*

To create a coherent corporate approach to metadata management that improve quality and efficiency, Statistics Canada has embarked on the modernization of metadata management with the Metadata Architecture Modernization project. Included as part of the first stream of activities is the identification of gaps in metadata management. With the increasing focus on metadata management influencing design (or redesign) of data storage and exchange in Statistics Canada, many of these gaps in metadata management are being filled. The project also aims as part of its second stream of activities to identify and implement structures for the corporate management of metadata specifically reviewing the way metadata is managed in corporate repositories. It is expected to move Statistics Canada toward greater coherence in terminology used for metadata and promote the reuse of systems and components that are based on this terminology.

*2.4 Policy framework for metadata management*

The policy framework for metadata management aims to revise existing policies and develop directives and guidelines on metadata management. The intention is to regulate the implementation of metadata modernization. The specific changes include revising the policies to reflect the common semantics of GSIM, updating the responsibilities to reflect the expected fluid exchange of metadata from areas of data creation to those of data dissemination. Specifically, two policies being updated refer to the creation and structuring of conceptual metadata and informing users of data quality and methodology.

*2.5 Redesign of the corporate metadatabase*

The redesign of the corporate metadatabase is expected to take place after a review of corporate metadata needs and current management of metadata in various parts of Statistics Canada. The (IMDB) is the corporate metadata repository for metadata on surveys, questionnaires, variables and classifications used in Statistics Canada. The need for corporate metadata management is so urgent that some initiatives related to the corporate metadatabase have already started or are beginning to influence the redesign (for example, see Section 4 for the use of web services to integrate metadata).

**3. Statistical metadata systems**

3.1 *Integrated Metadatabase (IMDB)*

The IMDB model is based on the ISO/IEC 11179 *Metadata Registries* and the *Central Metadata Repository Model* (CMR). The information objects included in the IMDB are used to provide enough information to enable users of data and metadata to interpret the results of surveys and statistical programs.

 Statistics Canada approved the Generic Statistical Business Process Model version 4 (GSBPM) [5] as a reference model in March 2010. The GSBPM has since been used to model several large projects as part of the modernization initiative. For each of the top level phases of the GSBPM, there is a description of the core phase and corresponding recommendations.

Metadata for each step of the survey process as well as collection instruments, complementary documentation, statistical variables and classifications are represented by administered items in the IMDB. The IMDB model is set up for version control of each administered item. Each time a survey releases data, a new version of the instance administered item is created. The instance represents the reference period and brings together all of the metadata related to the data release.

All metadata are structured the same way for all surveys and statistical programs. This allows for reusability and sharing of the metadata across surveys and programs. The IMDB contains an administration layer that documents the source of the metadata (responsible organization) and allows for the registration of each item. Registration serves to identify the status of completion of the metadata as well as its level of application.

The IMDB provides a way of organizing all the statistical units, variables and related classifications used in processing and dissemination within the Agency. It allows for improved coherence of the definitions of variables and classifications across subject matter areas and also promotes reusability. The IMDB can serve as a management tool for monitoring the extent to which standard definitions are used in the metadata, and the extent to which referential metadata are documented in a standard format to both internal and external users.

* 1. *Integrated business statistics program (IBSP)*

The IBSP, which includes approximately 250 surveys and administrative-based programs, was initiated in April 2010 to make use of and share generic corporate services and systems for collecting, processing, disseminating and storing statistical information. Content for business surveys is to be harmonized wherever possible and the approach to data analysis streamlined across programs.

A number of repositories for the different types of data, paradata and metadata are used throughout the process act as Data Service Centres. The main Data Service Centres are: the Business Register which serves as the frame, Tax warehouse which contains all tax files; the Integrated Metadata Base (IMDB) which contains metadata related to content; the IBSP data mart which contains all data processing files; and the Collection warehouse which contains raw data from respondents. The integration of generic services is facilitated by using the Enterprise Architecture Integration Platform **(**EAIP) which allows seamless data transformations between core business services.

The IBSP extract standard classifications from the IMDB with the use of a web service (refer to Section 4). Other classifications being accessed by the IBSP include Industry and Geography classifications required for the sampling, collection and other phases throughout the survey process.

* 1. *Common tools for social surveys*

The Common Tools Project was mandated to develop generalized tools and systems in view of harmonizing the business processes used across the Social, Health and Labour Statistics Field; to help in the transition of all social surveys to these tools; and to support and maintain the tools and processes. The common tools include the Questionnaire Design Tool (QDT) which was developed for the purpose of allowing subject-matter to specify and disseminate questionnaires using a standard approach, the processing and specifications tool (PST) which facilitates the creation and management of metadata related to variables within the Social Survey Metadata Environment (SSME), the Data Dictionary Tool (DDT) developed to consolidate, manage and standardize the work involved in the development of a data dictionary, and the Derived Variables Tool (DDT) which will be used to create, manage, develop and verify variables created during processing.

The IMDB uses the questionnaire design tool (QDT) from the suite of common tools to load questionnaire content. A questionnaire load web service was developed to allow for the load of question blocks, questions and response choices used by social surveys. In this case, the IMDB becomes the Agency’s repository of all questions and questionnaires for both business and social surveys.

* 1. *New Dissemination Model (NDM)*

The NDM is a current initiative to revise the organization of the Agency's data holdings to enable better discovery and enhanced navigation through the Statistics Canada web site, which is the Agency's primary dissemination vehicle.

The IMDB will have several roles within the NDM. For instance, data tables will include hyperlinks to the IMDB thus allowing users to access all definitions for the disseminated represented variables and classifications; the statistical variables which are linked to the data tables are also linked to the surveys that disseminated them; users have access to reference metadata for all surveys and statistical programs; and users will be provided with a drill-down approach to navigate and/or search for data using variables and classifications across surveys and statistical programs (please refer to figure A.4 below).

**4. Service oriented approach to metadata integration**

Statistics Canada is implementing a service approach based on our Corporate Business Architecture (CBA) principles. One of the goals of these principles is to foster interoperability and reuse by reducing the numbers of decentralized and specialized systems and processes. A key enabler of our service approach is the Enterprise Application Integration Platform (EAIP) that allows the delivery of solutions based on metadata driven, reusable software components and standards. Most business segments will benefit from the common core business services, standard integration platform, workflow and process orchestration enabled by the EAIP. The platform also simplifies international sharing and co-development of applications and components.

Web services currently in use and under development for the EAIP are associated to information objects representing core information entities (e.g., surveys, questionnaires, classifications, tax data, business registry) and statistical business functions (e.g. coding, editing, imputation). The formers correspond to GSIM’s *Concepts* and *Structures* groups. The latter, usually map to sub-processes in the GSBPM. Our services satisfy a basic set of SOA principles, i.e. they are loosely coupled (consumer and service are insulated from each other), interoperable (consumers and services function across Java, .NET and SAS), and reusable (they are used in multiple higher-level orchestrations and compositions). We are in the process of establishing a complete framework, including discoverability (via a service registry and inventory) and governance.

*4.1 Common models and data exchange*

Statistics Canada’s emerging service approach is providing the next generation of software components to be shared across the Agency. Services are reusable: they are designed to be combined with other services to create more complex solutions. In addition, generalized systems are being wrapped with a service interface to increase interoperability by shielding users from older technologies and multiple platforms.

One of the main challenges of this approach is that the same abstract information object (e.g. questionnaire, classification, T1 tax data) can be physically implemented by different data producers in different ways. Historically, this has been addressed by using point-to-point data integration in which the producer or the consumer has to conform to the other’s data model. In contrast, our service-oriented data integration relies on *common information exchange models*, i.e. enterprise-wide, common representations of information objects – a sort of “lingua franca” for data exchange (i.e. data in transit). These models, based on GSIM, enable the organization to share and exchange enterprise information that is consistent, accurate and accessible. It is important to note that common exchange models are not intended to replace the disparate set of heterogeneous physical models in use across the organization. Prescribing a single organization-wide model would be impractical and counterproductive. Instead, both data consumers and producers can continue to use their own models for data at rest (relational database schemas, SAS files, etc.) within their own environments and just translate them to the common exchange model only when data is in transit between systems.

*4.2 Integrating metadata repositories via services*

Domain-specific metadata repositories, such as IBSP and Common Tools, need to be integrated with the IMDB. One of the reasons for this is that even though metadata can be produced locally and stored in different repositories, it has to be centrally registered in order to facilitate governance, harmonization and reuse. These systems are currently being integrated with the IMDB via EAIP services that automate the metadata integration and exchange. The next are examples of metadata integration services developed or under development within the Agency.

*4.2.1 Classification exchange service*

We used Neuchâtel (now part of GSIM 1.1) to define the common classification model used by the EAIP service that extracts classification metadata from IMDB. This service is currently being used by IBSP. The first version contains the basic classes needed to support a classification structure, namely Scheme, Levels and Items. Each Scheme consists of one or more Levels (i.e., classes), each of which consists of one or more Items (i.e., members). To expose IMDB data in this common model, the IMDB’s ISO/IEC 11179 Metadata Registries entities need to be mapped to GSIM/ Neuchâtel. The IMDB data model does not have Scheme, Level and Item concepts (at least not with the usual GSIM/Neuchatel semantics), so a mechanism identifies and extracts them from the IMDB physical model via SQL mappings. At the conceptual level, this can be done by defining containment mappings that are expressed as subtypes between both models.

The Classification Management System (CMS) provides a GUI to allow users to develop new classifications and maintain existing ones. To exchange CMS classifications, the existing common classification model (and service) will be extended with additional metadata, e.g. concordances/ conformances, inclusions/exclusions, definitions, indexes, etc. to support the additional metadata requirements. This extended common classification model will then be used for integration between systems (e.g. CMS and IMDB) and to send classifications to clients internally, e.g. National Household Survey (NHS) and New Dissemination Model (NDM), and externally, e.g. Employment and Social Development Canada (ESDC). Figure 4.1 shows the preliminary conceptual architecture.



**Figure 4.1**: Common Classification Service and related components and flows

*4.2.2. Survey and questionnaire metadata exchange services*

Statistics Canada’s involvement in the development of the GSIM has influenced both GSIM and the Agency’s internal semantic work. A case in point is the work done by EAS with IBSP and the Integrated Collection and Operation System (ICOS) on survey instrument and questionnaires, which helped identify the need for a flow decision object separated from flow action that was included in version 1.0. This GSIM-based semantic work has been the starting point for developing a common questionnaire model for survey instrument and questionnaire in use by the EAIP questionnaire loading service that transfers data from both the Social Survey Metadata Environment (SSME) and the IBSP database into the IMDB.

Another service is being designed to integrate survey metadata from the IMDB and NDM metadata DB. NDM metadata DB is a CKAN-Solr solution that contains metadata about products together with discovery metadata to facilitate product search in the dissemination website. Some of this metadata, for instance about surveys, is being registered and maintained in the IMDB and need to be sent periodically to NDM in order to power the dissemination website. This exchange is going to be handled by new service based on a common survey exchange model. Figure 4.2 shows the current architecture with the questionnaire and survey services and their clients.



**Figure 4.2**: Common Questionnaire and Survey Services and related components and flows

**5. Conclusion**

Managed and coherent statistical metadata at Statistics Canada will drive and support all phases of the statistical business process efficiently; give convenient access to the statistical information that internal and external users need, in the way that they need it; and lead users to the appropriate interpretation, integration and use of statistical data.

Statistics Canada’s efforts in standardization provides for improved quality by maximizing reuse of existing corporate systems and enforcing reuse of concepts and classifications. Statistical metadata requirements have been identified and standardized within common processes and systems, such as the IMDB, IBSP and Common Tools. There is an opportunity to re-use or adapt the requirements and standardized metadata that can apply to other processes and systems.

Introducing a service oriented approach to metadata integration ensures the reusability by reducing the number of decentralized and specialized systems and processes. The centralized approach towards a corporate metadata repository allows for coherence as all metadata are structured the same across surveys and statistical programs. It is with the implementation initiatives such as the NDM, where subject-matter create the linkages between the data and the variables that we will see an increased reusability and a natural evolution towards harmonization.

**6. References**

[1] United Nations Economic Commission for Europe (UNECE) 2013, Generic Statistical Information Model (GSIM): Specification Version 1.1, December 2013. http://www1.unece.org/stat/platform/display/gsim.

[2] United Nations Economic Commission for Europe (UNECE) 2013, Common Statistical Production Architecture (CSPA) Version 1.0, December 2013. http://www1.unece.org/stat/platform/display/CSPA

[3] Bell, Michael (2008), Service-Oriented Modeling: Service Analysis, Design, and Architecture, John Wiley & Sons, Inc.

[4] Sweeney, Rick (2010), Achieving Service-Oriented Architecture: Applying an Enterprise Architecture Approach, John Wiley & Sons, Inc.

[5] United Nations Economic Commission for Europe (UNECE) 2009, Generic Statistical Business Process Model (GSBPM) Version 4.0, METIS Workshop on the Statistical Business Process and Case Studies.

**Appendix I – Statistics Canada and CSPA**

Many statistical organizations are currently facing similar challenges and are trying to implement common approaches. Historically, they have developed their own business processes and IT-systems for producing statistical products. As a consequence, every technical solution was built for a very specific purpose and platform, hence limiting interoperability, reuse and information sharing. To address this problem, many statistical organizations are in the process of modernizing and transforming their organizations using enterprise architecture to underpin their vision and change strategy. As part of the modernization effort, the High Level Group for the Modernization of Statistical Production and Services (HLG) has prioritized the development of the CSPA, currently under way.

CSPA will assist in the modernization of statistical organizations by providing a framework, including principles, processes and guidelines, to help reduce the cost of developing and maintaining processes and systems and to foster sharing and reuse of process components across the industry. CSPA is the industry architecture for the official statistics industry. An industry architecture is a set of agreed common principles and standards designed to promote greater interoperability within and between the different stakeholders that make up an “industry”, where an industry is defined as a set of organizations with similar inputs, processes, outputs and goals (in this case official statistics).

**CANCEIS Proof of Concept**

As part of the CSPA initiative, Statistics Canada is in the process of designing and implementing statistical services that will perform one or more functions in the statistical process. Statistical services can be developed at different levels of granularity: fine-grained statistical services encapsulate small pieces of functionality, e.g. a step within a GSBPM sub-process, whereas coarse-grained or aggregate statistical services encapsulate a larger piece of functionality, e.g. a whole GSBPM sub process, which in turn may be composed of a number of fine-grained services.  Inputs and outputs of statistical services are modeled using GSIM at the conceptual level, and DDI and other XML formats at the implementation level.

As part of a proof-of-concept for CSPA, we developed a service wrapper for CANCEIS,The Canadian Census Edit and Imputation System. CANCEIS is a tool used by Statistics Canada to perform edit and imputation of variables that contain data from the Canadian Census. For the proof-of-concept, all metadata related to variables were represented using DDI 3.1. Configuration and data files were maintained in native CANCEIS format.



**Figure A.1**: Service layer for the CSPA CANCEIS wrapper

Figure A.1 illustrates the levels of abstraction in the development of statistical services. The top level (in blue) is the Statistical Service Definition and corresponds to the conceptual (or business) layer. The capabilities of a statistical service are described in terms of the GSBPM sub process that it relates to, the business function that it performs and the GSIM information objects that are the inputs and outputs. The bottom level (in red) is the Statistical Service Implementation Description and corresponds to the implementation (or technical) level. In this layer, the capabilities of the statistical service are refined into detailed operations whose inputs and outputs are GSIM implementation objects (in this case DDI 3.1). This layer defines the service contract, including communications protocols, a precise description of all dependencies to the underlying infrastructure, non-functional characteristics and any relevant information about the configuration of the application being wrapped. For more granular services, CSPA also define an intermediate, logical layer (not shown in the picture) that fleshes out the business level capabilities into more detailed business functions before refining them into implementation operations.

**Appendix II – IMDB tools**

Internally, the IMDB has its own portal on the Statistics Canada Wiki portal which displays its content. The metadata are centralized and can be viewed from a survey perspective, a variable perspective, a question perspective, etc. The wiki technology facilitates metadata discovery and metadata search of the IMDB.

For example, a search on the statistical variable “*Expense of establishment*” produces two represented variables: “*Expense of establishment, value*” and “*Expense of establishment, type*”. By choosing the former, the Wiki identifies the surveys that disseminate the variable and for which reference period; the questions asked to collect the variable and the data files in which the variable is included.



***Figure A.1*** *shows a screenshot of a Wiki page.*

External users have several entry points to access the metadata from the IMDB. The [Definitions, data sources and methods](http://www.statcan.gc.ca/concepts/index-eng.htm?MM)[[1]](#footnote-1) web module on the Statistics Canada web site provides access to web pages for each of the active and inactive surveys (see Figure A.3). Each page contains the procedural metadata, links to definitional metadata as well as access to data tables and other related publications and announcements.

Also included on this web module are standard definitions for variables, statistical units and classifications which are also documented in the IMDB.



***Figure A.2:*** *The Definitions, data sources and methods web module on Statistics Canada web site*



***Figure A.3:*** *Data table in CANSIM with links to the North American Industry Classification System definition in the IMDB.*

1. http://www.statcan.gc.ca/concepts/index-eng.htm? [↑](#footnote-ref-1)