Distr. GENERAL

01 May 2013

WP.15

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata
(Geneva, Switzerland, 6-8 May 2013)
Topic (ii): Case studies and tools

CANADA'S UPDATED CASE STUDY AND THE BENEFITS AND CHALLENGES OF IMPLEMENTING THE GENERIC STATISTICAL INFORMATION MODEL

Working Paper

Prepared by Flavio Rizzolo, Tim Dunstan and Kathryn Stevenson, Statistics Canada

I. Introduction

1. This updated case study outlines the current state of statistical metadata management at Statistics Canada including a proposed metadata management strategy and action plan, a description of key metadata repositories across the Agency, the implications of transitioning to a Service Oriented Architecture (SOA) and the adoption of the Generic Statistical Information Model (GSIM).

A. Metadata Strategy

2. Statistics Canada is undergoing an agency-wide modernization initiative to promote organizational efficiency, increase robustness of systems and processes and shorten implementation time for new projects. One of the key principles guiding this review is: *Create metadata at the beginning of every process and use them throughout the project life cycle.* A working group with members representing all phases of the statistical business process has been developing a strategy for statistical metadata management and an action plan to support these principles. The goals being considered for the strategy relate to four themes: drive, make available, structure and manage. Actions required to implement the strategy are expected to cover:
   (a) Establishing governance;
   (b) Creating centres of expertise;
   (c) Adopting standard structure and content specifications (including GSIM);
   (d) Developing a metadata portal to identify authoritative sources of metadata and make them available to users, and;
   (e) Identifying metadata gaps required to enable metadata-driven processes.

B. Current Situation

3. The strategy for statistical metadata management and high-level actions are being reviewed by senior management. Implementation of activities identified in the action plan is expected to begin in spring 2013. This includes the adoption of GSIM as a recommended reference model for the Agency. Some of the anticipated challenges to implementing the strategy and GSIM include limited resources, on-going work, transitioning to a SOA and Government of Canada initiatives (i.e., Open Data project). Establishing governance, assigning a project manager to lead implementation and validating GSIM through a pilot project are key steps to realize a metadata-driven environment.

II. Statistical Metadata Systems and the Statistical Business Process

A. Statistical business process

4. Statistics Canada approved the Generic Statistical Business Process Model version 4 (GSBPM) 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 business phase and corresponding recommendations.[1]

B. Current Systems

(a) Integrated Metadata Base (IMDB)

5. The IMDB is based on the ISO/IEC 11179 Metadata Registries and the Corporate Metadata Repository model (CMR). The metadata layer extends across all phases of the statistical business process and can support disseminated data, analysis, archived datafiles, and the planning and design of surveys. Metadata in the IMDB is beginning to be linked to some data warehouses, which hold both micro- and aggregate data; and can be potentially used for data analysis including data benchmarking and data confrontation.

(b) Integrated Business Surveys Project (IBSP)

6. The business statistics program includes approximately 250 surveys and administrative-based programs. The IBSP was initiated in April 2010 to make use of shared and generic corporate services and systems for collecting, processing, disseminating and storing statistical information. Content for business surveys is to be harmonized wherever possible[2] and the approach to data analysis streamlined across programs.

(c) The Common Tools Project

7. The goal of the Common Tools Project is to implement a harmonized set of processes and tools to support social surveys. The project is divided into two primary environments, the Social Survey Metadata Environment (SSME) and the Social Survey Processing Environment (SSPE). SSME uses four tools to feed information to and from a metadata repository. This allows metadata to be documented once and reused throughout the process to improve quality and generate efficiencies.

C. Costs and Benefits

8. Detailed costs associated with implementing these projects are not yet available. Benefits from establishing the IMDB include more rigorous information and metadata management, harmonization and standardization of concepts, knowledge sharing and reuse of information assets.

9. The IBSP has incorporated good metadata management and is starting to integrate systems that are metadata-driven and that optimize the use of corporate services such as collection and methodology.

[1] See Section III-B for further discussion of the Agency's use of the GSBPM.
[2] A project to harmonize content for business surveys will be undertaken in 2013.

Reuse of content modules, increased use of administrative data and the adoption of electronic questionnaires are expected to reduce micro-editing, standardize methods and processes and lead to operational efficiencies.

10. Creating a metadata environment (SSME) within the Common Tools project has facilitated the transfer of information between business processes through a suite of tools for survey documentation.

11. Overall, these projects are reducing system maintenance, development and training costs while improving sharing of standards and best practices.

D. Implementation Strategy

(a) Integrated Metadata Base (IMDB)

12. The IMDB has been implemented using a "step-wise" approach with three development phases and future opportunities to re-use metadata and expand the IMDB metadata model to link to other information systems in the Agency.

13. Phase 1 produced a set of static web pages displaying data sources and methods for each statistical program and survey. These were accessible to external users through hyperlinks from data tables and publications on the Statistics Canada website. Internal users could browse the full inventory through the Agency's Intranet site.

14. In 2000, Phase 2 began collecting reference metadata including survey methodology and data accuracy. Information was formatted and validated by subject matter areas before being loaded into the IMDB. Like the initial phase, Phase 2 information is available to external users through hyperlinks on the website and an internal version on the Intranet site. Updates are triggered by new data releases so that metadata accompanies every release.

15. The past 10 years have seen improved quality of Phase 2 content and a push to include more information in the IMDB. Phase 3 has been initiated to add definitions of concepts, variables and classifications for all subject matter areas. This work is expected to be completed by spring 2015.

(b) Integrated Business Surveys Project (IBSP)

16. The IBSP has produced a semantic model with a proposed IBSP metadata classification, along with an agreed-upon vocabulary of terms and definitions. Details of metadata by system components are being identified. Current plans are to develop one single portal to make accessible overarching metadata, the data it describes and the processes it controls. This tool will enhance reporting capabilities and built-in quality assurance.

(c) The Common Tools Project

17. Systems have been developed using the Rational Unified Process (RUP), an iterative approach. Using this method, several tools are in development concurrently. The priority is to deliver basic functionality as quickly as possible and to combine all tools into an integrated processing and metadata management environment. The initial phase of the project created a processing environment and associated standards including naming conventions and directory structures. Subsequent phases cover individual processing steps.

III. Statistical metadata in each phase of the statistical business process

A. Metadata classification

18. GSIM is being adopted to specify, design, and implement components that will easily integrate into "plug'n'play" solution architectures and seamlessly link to standard exchange formats (e.g. DDI, SDMX). It is important to note that GSIM does not make assumptions about the standards or technologies used to implement the model, which leaves the Agency room to determine its own implementation strategy.

19. Statistics Canada is beginning to use GSIM's *Concepts* and *Structures* Groups as the main classifiers of metadata. These groups contain the conceptual and structural metadata objects, respectively, that are used as inputs and outputs in a statistical business process. The *Structures* group defines the terms used in relation to data and their structure. The *Concepts* group defines the meaning of data, providing an understanding of what the data are measuring.

20. Work focuses on aligning the new GSIM-based classification with other internal metadata classification models currently in use. For instance, IBSP identifies the following types of metadata:
    - Reference metadata: Describes statistical datasets and processes.
    - Definitional metadata: Description of statistical data (with meaning to business user community) E.g., concepts, definitions, variables, classifications, value meanings and domains.
    - Quality metadata: Quality evaluation of a dataset or individual records; helps users assess the fitness of associated data for their specific purposes. E.g., CV, rolling estimates, analysts comments about the quality of a set of records.
    - Operational metadata: links between the concepts and the physical data.
        - *Processing specifications*: Capture, edit and output specifications and processing flags.
        - *Processing results*: Content, outcomes, outputs of processing.
        - *Paradata*: Data from the collection operation or statistical analysis used to support decision making in the survey process or statistical analysis. These include system logs, history files and comments.[3]
    - Systems metadata: Low-level information about files, servers and infrastructure that allows the physical IT environment to be updated without re-specification by the end user.

B. Metadata used/created at each phase

21. Metadata use is not uniform across all GSBPM phases. IMDB metadata, consisting mostly of GSIM *Concepts* and *Business* objects, is used for survey design (phase 2) and dissemination (phase 7). Survey managers use the IMDB to identify existing variables for reuse. New variables and related questions will be soon documented and stored during questionnaire design as well, which will then be used by collection processes across the Agency. In the dissemination phase, the IMDB is the primary source of summary texts describing surveys, definitions of variables, related methodology, and data quality and questionnaire images. Most products on the Statistics Canada website offer a link to the related IMDB survey records.

22. The System of National Accounts (SNA) creates and uses metadata (classifications) for the data integration sub-process (5.1) of the GSBPM. In particular, the SNA creates the Input-Output Industry Codes (IOIC), Commodity Codes (IOCC) and Institutional Sectors classifications. They are mainly used for the GDP surveys (annual, quarterly and monthly). SNA classifications are being exported to the IMDB and integrated into data warehouses for analysis via a classification web service[4]). In addition, concordances to NAICS, NAPCS and other international classifications will be maintained in a Classification Management and Coding System (CMCS).

23. The Social Survey Processing Environment (SSPE) has its own metadata repository (see Section IV-B) which is used from design through dissemination, including survey and questionnaire metadata, codesets and codebooks. The SSPE repository does not use the same metadata objects that are in the GSIM *Business* or *Structures* groups.

24. Underlying the GSBPM is the implicit need for common semantics which require some degree of harmonization and maintenance across all phases. Even common concepts like *questionnaire*, *survey*, or *classification* mean different things across the Agency. Enterprise Architecture Services (EAS), Methodology and Subject Matter areas have worked collaboratively to make progress in semantics work for the IBSP on a number of topics, including:

[3] For example: analyst comments about their analysis, output of statistical processes; respondent comments, interviewer comments or additional information about the respondent obtained during collection.
[4] See Section IV-F-(a) for more information on this service.

- variables, defined by purpose type (identification, statistical, processing and design variables) and by value source[5] (collection, administrative/tax, edit and imputation, subject matter correction);
- distinction between derived variables (calculated value that leads to a change in statistical concept) and transformed variable (change in value but same concept, usually metadata driven);
- statistical activity concepts and their identifiers (activity, program, survey, instance and questionnaire);
- collection edits, defined by edit types, specifications (decision points) and actions;
- sample design, allocation, estimation and statistical units;
- quality indicators, measure of impact, paradata.

25. 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 (submitted to the GSIM group for review). This semantic work has been the starting point for developing a canonical model for survey instrument and questionnaire for the SOA[6] .

26. IBSP has also developed a conceptual framework and naming convention for harmonized content. SSPE has developed standardized questionnaire modules for cross-cutting household survey variables. These modules contain standard concepts, definitions, classification and wording for multiple collection modes.

C. Metadata relevant to other business processes

27. Several projects have been identified as potential content providers or consumers of the IMDB. The IMDB now stores documentation for public use microdata files as part of the requirements for the Data Liberation Initiative (DLI) - an initiative between Statistics Canada and Canadian universities to share data for social science research. Statistics Canada makes available to universities and colleges, by subscription, all of its statistical products including microdata files using Data Documentation Initiative (DDI) specifications.

28. Another initiative under development is the Research Data Centre (RDC) Metadata Project. Rather than integrating with the IMDB on a case-by-case basis (point-to-point integration), authorized applications can gain access to content through IT industry standard web services in a standard based format (DDI). This approach is expected to reduce development costs, allow for code and component reuse across projects and foster the adoption of global standards across the Agency. It will also support the future establishment of standard-based data and a metadata management framework.

IV. System and design issues

A. IT architecture

29. Statistics Canada is moving towards a SOA. A key enabler of SOA is the Enterprise Application Integration Platform (EAIP) that allows the delivery of solutions based on meta-data 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.

30. Web services currently in use and under development by EAS are associated to information objects representing *core business entities* (e.g., *questionnaires*, *classifications*, *tax data*, *business registry*) that are classified into GSIM's *Concepts* and *Structures* groups. This fits nicely with GSBPM as well: services provide the inputs and outputs to GSBPM statistical processes. They satisfy a basic set of

[5] It indicates the provenance of the value for data quality purposes.
[6] See Section IV-F for more information on SOA canonical models.

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). Work continues to establish a complete framework, including discoverability (via a service registry and inventory) and governance.

31. At this point, Statistics Canada has a combination of services and silo-based/point-to-point integration that can be described as a combination of maturity levels 3 and 4 in terms of the Open Group Service Integration Maturity Model (OSIMM) maturity matrix (see Figure 1). During the transition years to a corporate-wide SOA, incremental changes are being made by applying SOA adoption and governance by segment in which cross-silo services and consumers coexist with point-to-point integration of systems and data. Early adopters of SOA services include IBSP, SSPE and SNA.

32. Developing Data Service Centres (DSC) is a key initiative that fits into Statistics Canada's emerging SOA. The objective of the DSC is to manage statistical information as an asset – to maximize its value by improving accessibility, utility, accuracy, security and transparency through the use of a centralized inventory of statistical data holdings, associated metadata and documentation. Key statistical files and associated standard metadata (i.e., file name, type, description, creators, owners, etc) will be registered and integrated into statistical processes via SOA. This integration will rely on a data access layer with common interfaces to access statistical files without the user needing to know their location, format and/or technology.

B. Metadata management tools

33. IMDB metadata discovery is performed via a Wiki-based solution and MetaWeb. Each Wiki page provides the context of the information and all available links. These pages are programmatically generated based on templates developed for the IMDB. MetaWeb is a JSP and Servlets-based application. Data are collected and populated into the IMDB via a Microsoft Excel IMDB Extraction/Loader, an Oracle PL/SQL IMDB Loader and MetaWeb.

34. The starting point for the Common Tools project (See Section VII - Figure 2) is the Questionnaire Development Tool (QDT) used to enter specifications for social survey data collection instruments. All question metadata is entered in the QDT, including questions and answer category text, interviewer instructions and conditions controlling flows. The Processing and Specifications Tool (PST) then loads variable metadata such as variable name, length and type. These are linked to question metadata already entered via QDT so no re-entering of question or answer category text is required. Finally, the Social Survey Processing Environment (SSPE) utilities use collection layouts or schema to generate variable metadata to be loaded to the metadata repository. Two projected tools will complete the picture: the Data Dictionary Tool (DDT), which will provide an interface to the metadata repository for updating descriptive variable metadata, and the Derived Variable Tool (DVT), which will allow entry of specifications for derived variables and will be used to produce detailed documentation for data users. Within Statistics Canada's SOA, the SSPE metadata repository will export metadata in a canonical model to IMDB via an EAIP service under development[7] .

35. Solutions and tools are needed to support other types of metadata, specifically in the GSIM *Structures* and *Production* groups.

C. Standards and formats

36. The following is a list of standards and formats and where they are being used:
    - BPMN – EAIP orchestrations;
    - ISO/IEC 11179 Metadata registries – IMDB;
    - CMR – IMDB;
    - DDI 2.1 – DLI and Research Data Centres;

[7] See Section IV-F for more information on SOA.

- DDI 3.0 – IMDB tool (automate metadata "wrap" for microdata files/PUMFs) and web services (extract metadata from IMDB for DLI and Research Data Centres). See Section IV-E for more details on this project;
- ISO/TS 17369 SDMX ML – Formatted data from dissemination;
- Neuchatel Terminology Model Part 1;
- Classification database object types V2.1 – Standards Division;
- ISO 3166-1:2006 Part 1: Country – Standards Division;
- ISO 19115 Geographic Information – Geography Division;
- ISO 15489-1 Part 1: General – Information management.

D. Version control and revisions

37. For web services that expose information assets, not only the underlying data evolve (both in content and structure) but also the services that expose it. As a result, (potentially) different versions of the same data will be published and exchanged by (potentially) different versions of the same service. No centralized versioning framework for data exists and many areas have customized versioning schemes.

38. For example, the IMDB allows time travel by version and effective period. A new version of a metadata item is created by copying an existing item, making necessary changes and assigning the version number to the immediate next version. Each version has an interval of validity (or effective period) associated to it. In other words, the lifespan of each version of a metadata item can be determined; conversely, the version of an item in effect at a specific point in time can also be determined.[8]

39. Service versions are identified using a three-digit versioning scheme: *major.minor.patch*. An increment in the *major* version requires some of its consumers to change their code. This happens because of a major change in the service contract, e.g., at least one operation has been removed or an operation signature has changed in a way not foreseen by the extension points defined in the Web Service Description Language (WSDL) file.[9] An increment in the *minor* version does not require changes on the consumer applications. These are implementation changes and/or backwards compatible changes to the interface, e.g., additions of operations or extensions to data types in the WSDL file. An increment in the *patch* version is only used for bug fixes.

40. Service versions are designed with the goal of making them as *forward* and *backward* compatible as possible. By making the interface extensible, *forward compatibility* makes room for future, uncertain functional requirements. This approach is guided by knowledge and best practices in SOA interface design, XML schema design and type theory (since forward compatibility of service interfaces is essentially a special case of subtyping). *Backward compatibility* is achieved in the usual way: by ensuring that consumer applications developed for older versions of the service can continue to work with the new version.

E. Outsourcing versus in-house development

41. External consultants were contracted for building DDI services and tools, specifically to develop in-house DDI expertise and a set of core SOA web services around the IMDB. These services expose IMDB content in a standard format compliant with the DDI XML specification to support applications that focus on different types of metadata (e.g. surveys, variables, classifications, concepts, etc.). Rather than integrating with the IMDB on a case-by-case basis (point-to-point integration), the web services enable applications to gain access to its content in a standard based format. This initial effort defined and implemented a core metadata service that delivers IMDB

[8] Not every change to a metadata item generates a new version: versioning of different entity types (surveys, classifications, questionnaires, etc) are handled by a different set of business rules.
[9] Interface specification that describes the functionality and data types of a web service.

content encoded in DDI XML. A testing tool was also developed based on a set of common use cases (see Figure 3) to validate the effectiveness of the approach. The service is used to support the Data Liberation Initiative (DLI) and the Canadian Research Data Centre Network (CRDCN) Metadata projects comprising 25 Research Data Centres (RDCs) from universities across the country. The services were developed with a Java technology stack, including some JPA components for database access that were reused in other in-house services.

42. In addition, EAS developed a proof-of-concept client based on JSPs, Servlets and XSLTs to transform and render the DDI XML content returned by the data service into human-readable HTML and other proprietary formats for interoperability with internal applications (e.g., SQL Server, SAS).[10]

F. Sharing software components and tools

43. Statistics Canada's emerging SOA 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.

44. 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 (and even by different data consumers) in different ways. This "impedance mismatch" has historically been addressed by *point-to-point data integration*, i.e., either the producer or the consumer has to conform to the other's data model. With SOA, canonical information models are created to which both producers' and consumers' models will *map* (*SOA data integration*). Canonical information models are enterprise-wide, common representations of information objects – a sort of "lingua franca" for data exchange. These models enable the organization to share and exchange enterprise information that is consistent, accurate and accessible. A mapping is a specification that describes how concepts from two different models relate to each other. At the physical level, it actually specifies how data are *translated* between two models. Canonical models are not intended to *replace* the disparate set of heterogeneous physical models in use across the organization. Prescribing a single model would be impractical and counterproductive. Instead, both data consumers and producers can continue to use their own models (relational database schemas, SAS files, etc.) within their own environments and just map to the canonical only when data need to be exchanged.

45. Within the SOA framework, canonical models are implemented as object models that are serialized into XML Schema Definition (XSD) types. Data producer and consumer schemas are mapped to the canonical object models used by services via schema mappings – object-relational (ORMs) or object-XML (OXMs). An inventory of canonical XSD types is currently being created; it can be referenced and reused by multiple service contracts (WSDL) in the EAIP schema registry. These XSD types will be maintained by the service developers within the governance framework set up by the EAIP.

46. When exchanging data from a source database to a consumer application, there are a number of mappings involved along the way. First, data need to be extracted from a relational or multidimensional database into the canonical object model. This could be done automatically by object-relational mapping (ORM) tools, when the source schema is close in structure to the canonical, or it may require customized SQL/MDX extraction queries. At the other end of the process, the canonical object model is serialized into XML/JSON to be shipped to the client application via a web service interface. This mapping is done automatically by the EAIP tools. Finally, the client application needs to map the XML/JSON produced by the service into its own object model via an

[10] See Section VII – Figure 4 for the overall architecture of the IMDB DDI services.

automatic de-serialization process. This process may include some XSLT transformation when the canonical model is very different from the consumer model and requires restructuring.[11]

(a) Example: Classification service

47. Classifications were one of the first core business entities to use an EAIP service. The Classification canonical model is based on GSIM and Neuchâtel. 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). This model will be extended to include Versions and Variants as necessary.

48. To expose IMDB data in this canonical 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.[12]

49. There are parent-child hierarchies defined on Classification Level and Classification Item. The Level hierarchy is linear (each level has at most one child) and the Item hierarchy is a tree (each item may have zero or any number of children). Both hierarchies are related by a constraint that ensures that two items are in a parent-child relationship only if their respective levels are in a parent-child relationship as well. This constraint ensures that both hierarchies remain consistent.[13]

V. Organizational and workplace culture issues

50. An environmental scan of metadata management practices was recently conducted by the Agency's Statistical Metadata Working Group. Strengths included broad recognition of the importance of statistical metadata, significant metadata expertise in certain areas, use of authoritative sources for some metadata (e.g. conceptual metadata from the IMDB) and a range of metadata currently being created and used in the statistical business process. The challenges highlighted metadata documented at the end of the statistical process, difficulty finding metadata because of numerous storage locations, storage formats and localized value domains and few formal mechanisms to coordinate, review or implement metadata management across the Agency.

A. Overview of roles and responsibilities

51. Each change activity identified in the proposed Strategy for Statistical Metadata Management action plan will be assigned a lead division or area to take responsibility for the business plan, resources and reporting regularly on progress.

52. Centres of expertise or responsibility for each statistical domain in the Agency will also be formally recognized. These centres would provide a stewardship function for statistical metadata within their domain including providing advice related to design and standards development and ensuring that the

[11] Section VII - Figure 5 depicts the entire process of exchanging data from a source database to a consumer application.

[12] Section VII - Figure 6 shows the relationship between both models (the canonical entities are those starting with the word "Classification"). For the purpose of defining a mapping, Classification Schemes and Levels can be viewed as a subtype of Enumerated Value Domain (EVD), whereas Classification Items are a subtype of Permissible Value (PV). Classification Schemes are a special type of EVD with no PV (i.e., Item) directly associated to them – Items are only associated to Levels. All Items associated to a given Level have different Values.

[13] Section VII - Figures 7 and 8 show the actual physical mapping for Classification Scheme, Item and Level. The mapping is defined by UML notes (the boxes with the bended corners). The syntax of the mapping is straightforward: the "<<" symbol indicates an assignment from the attribute on the right to the attribute on the left. In addition, there are constraints on code sets from the IMDB content model.

authoritative source is comprehensive, up-to-date and accessible to the users (people and systems) who need it.

53. Roles and responsibilities for all areas of metadata management will become clearer as the key activities in the action plan are implemented.

B. Metadata management team

54. A steering committee is expected to oversee implementation of the operational phase of the proposed Strategy for Statistical Metadata Management action plan. This committee would be comprised of senior managers from key stakeholder areas and would focus on the activities necessary to build capacity. A project lead will be assigned to monitor implementation projects and lead the communication and learning initiatives.

55. Once the first phase of implementation is complete, a permanent governance structure will be established to ensure the use of standard metadata structures, content and exchange protocols; to provide guidance on creation of metadata at the appropriate stage of the statistical business process and to identify opportunities to align the Agency's on-going work with the adopted enterprise architecture.

C. Training and knowledge management

56. Communication and learning initiatives are essential to realize statistical metadata management across the Agency. Building capacity among staff about the scope of statistical metadata, its role in the statistical business process and use of reference models like the GSIM and the GSBPM are key priorities. Roles and responsibilities and best practices for the planning, design, creation, use and management of statistical metadata will also be communicated across the Agency.

D. Partnerships and co-operation

57. Statistics Canada continues to support and participate in international initiatives such as the GSIM and the GSBPM. The Agency was involved in the original GSIM sprints and is currently participating in the Plug & Play sprint and three separate task teams: GSIM-DDI mapping, GSIM-SDMX mapping, and GSIM implementation – all initiatives have senior management commitment.

58. The Agency is also represented in the UNECE METIS, UN Expert Group on International Statistical Classifications, UNECE High-level Group on the Modernization of Statistical Processes and Services, SDMX Statistical Working Group, DDI/SDMX Dialogue and the Extensible Business Reporting Language (XBRL) Canada Working Group.

59. Evaluating DDI and the role of other international format standards in the statistical business process has been flagged for further discussions.

E. Other issues

60. A major challenge has been integrating international work with Statistic Canada's internal projects. In particular, (1) the GSIM development with IBSP and ICOS Questionnaire/Survey Instrument (described in Section III-B), and (2) the GSIM implementation task team (*Production* group) with current work on IBSP. The IBSP project is beginning to look at *Production* metadata and it is hoped that involvement in the GSIM Implementation task team can be leveraged to support this stage of development.

VI. Lessons learned

61. Embarking on a wide-ranging modernization initiative has focused the Agency's attention on the role of statistical metadata. Developing an enterprise-level statistical metadata management strategy has been the first step to engage stakeholders in discussions about the scope of statistical metadata, stewardship and a vision for creating a metadata-driven environment. The next step is to identify opportunities to align current business practices with the action plan.

VII. Attachments and links

Figure 1: Open Group Service Integration Maturity Model (OSIMM) matrix

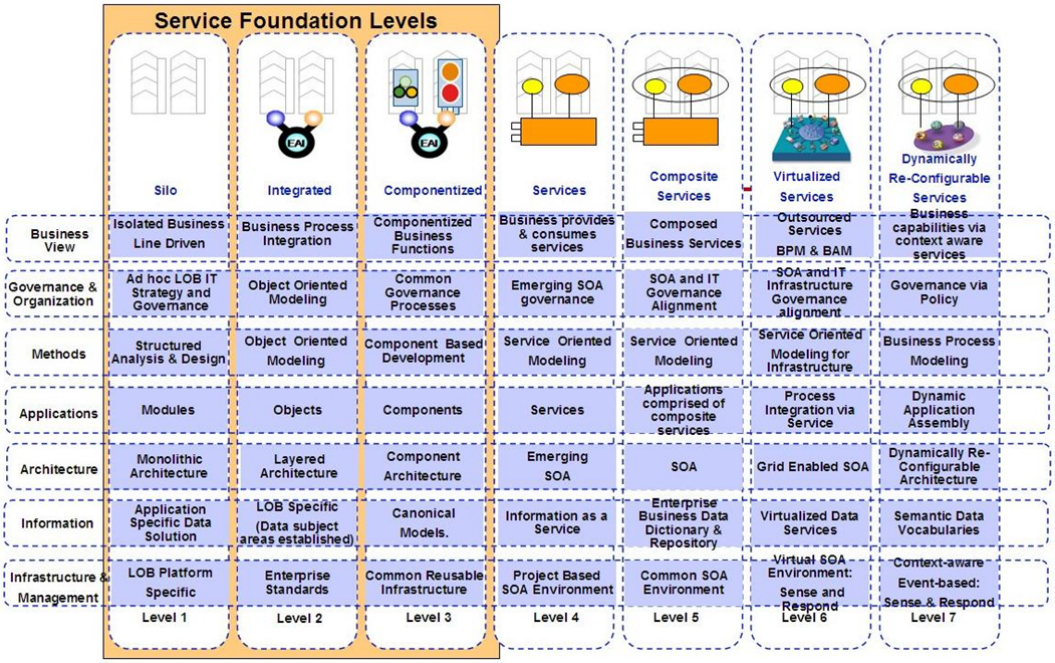

Figure 2: Common Tools Project Social Survey Metadata-driven Environment

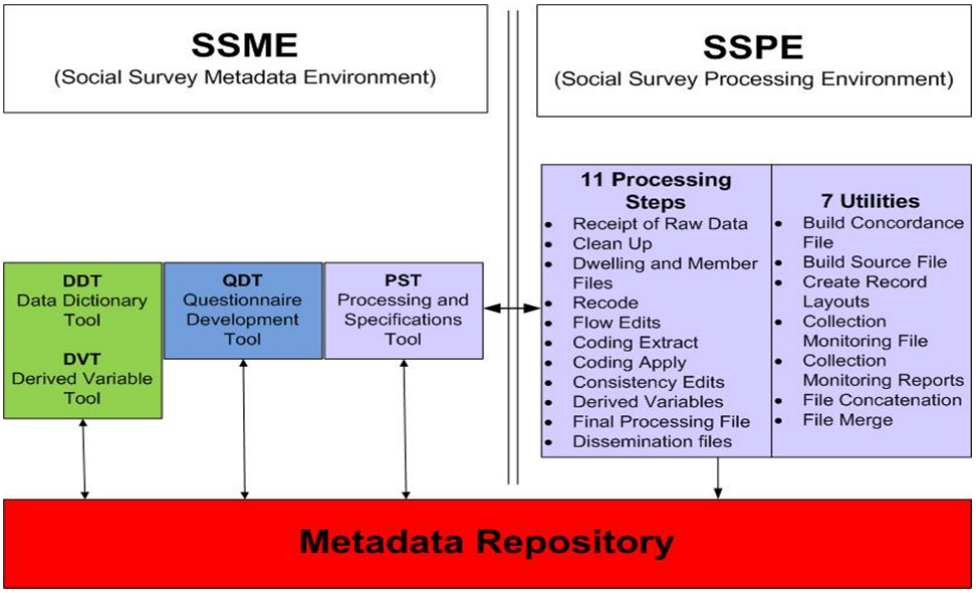

Figure 3: Integrated Metadata Base (IMDB) DDI Services Use Cases

Use case	Input	Output (includes bilingual information)
Get Survey	Identification and governance information	Identification, version and governance information + abstract, frequency, state in life cycle, summary of changes in the survey version, methodology, universe.
Get Survey Instance	Identification and governance information	Identification, version and governance information + reference periods for data collection, publication
Get Value domain	Identification and governance information	Includes classifications and other value domains commonly called codesets. Identification, version and governance information + codes, value meanings, classes (levels of the hierarchy).

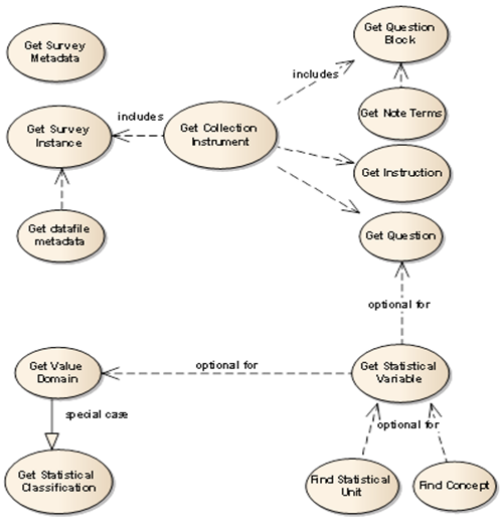

Figure 4: Integrated Metadata Base (IMDB) DDI Services Architecture

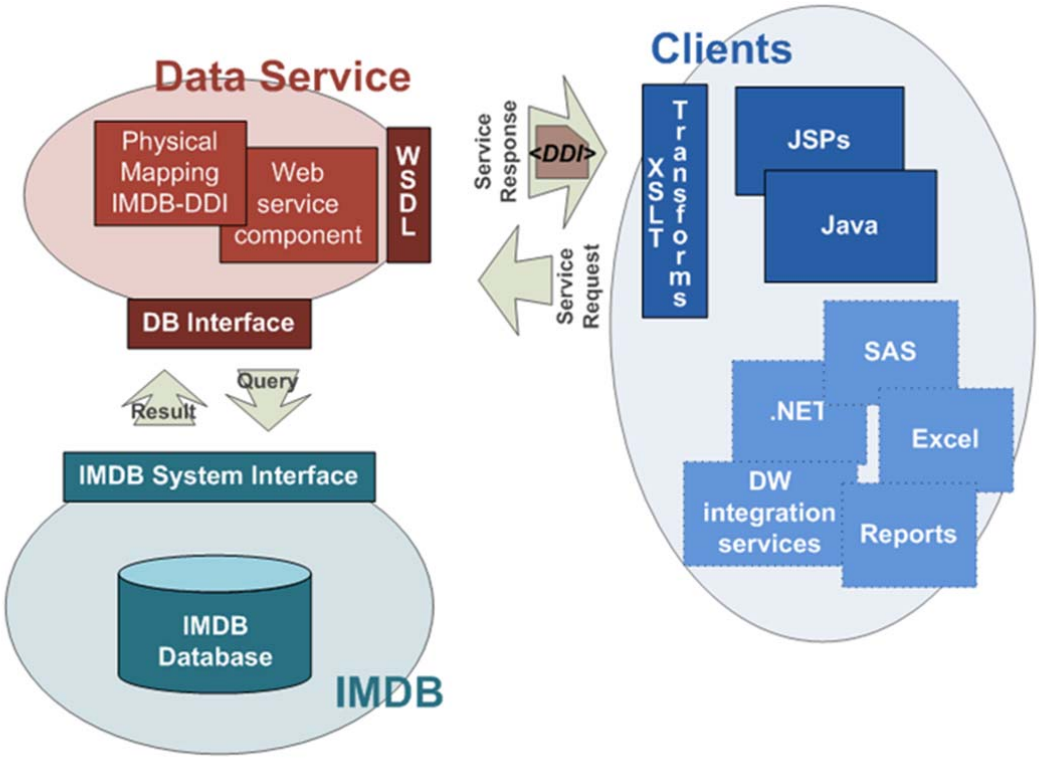

Figure 5: Statistics Canada's Service Oriented Architecture Models and Mappings

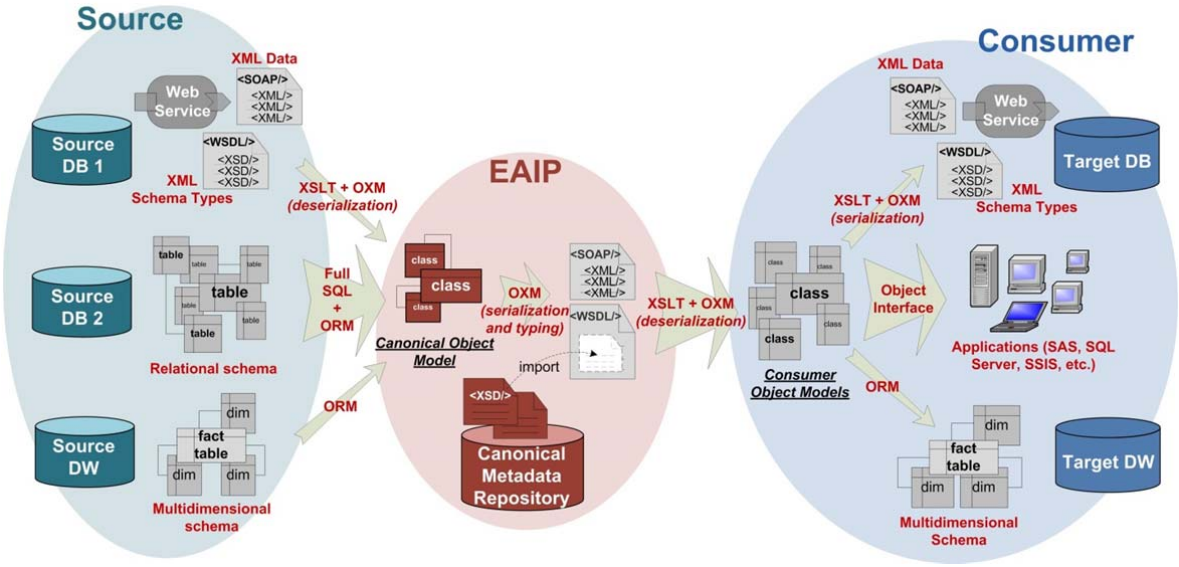

Figure 6: Conceptual containment mapping from Classification to ISO/IEC 11179 Metadata Registries

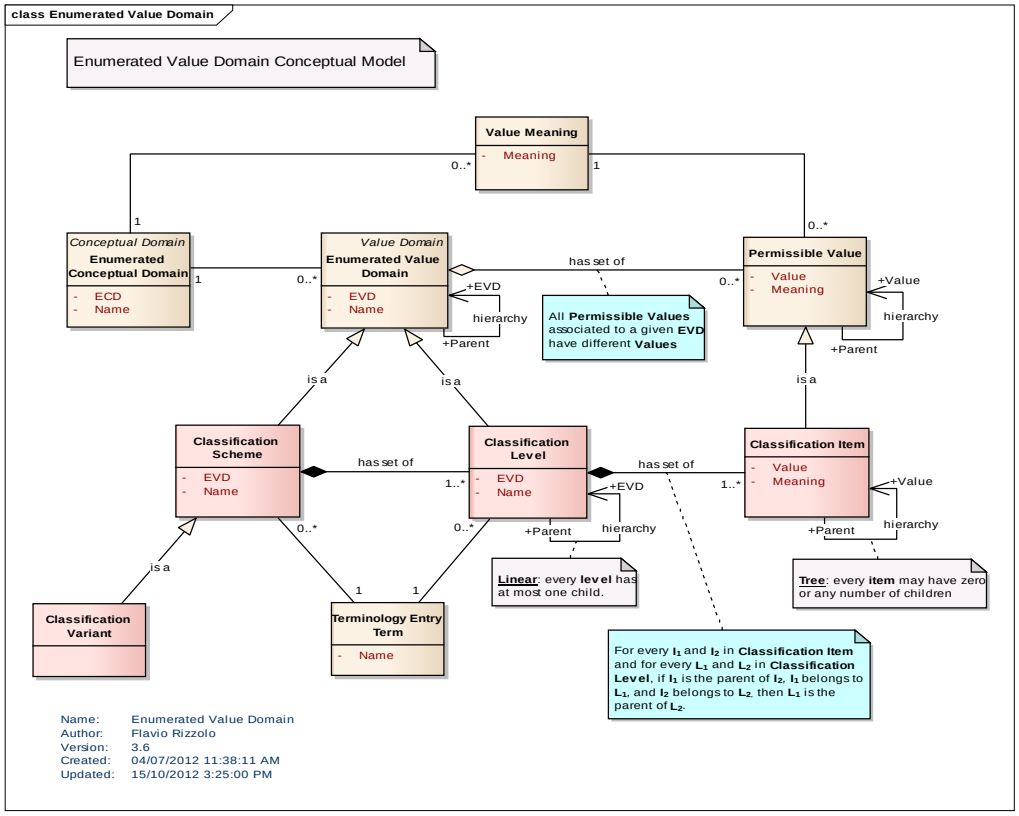

Figure 7: Physical Mapping from Canonical Classification Model to Integrated Metadata Base (IMDB) ISO/IEC 11179 Metadata Registries

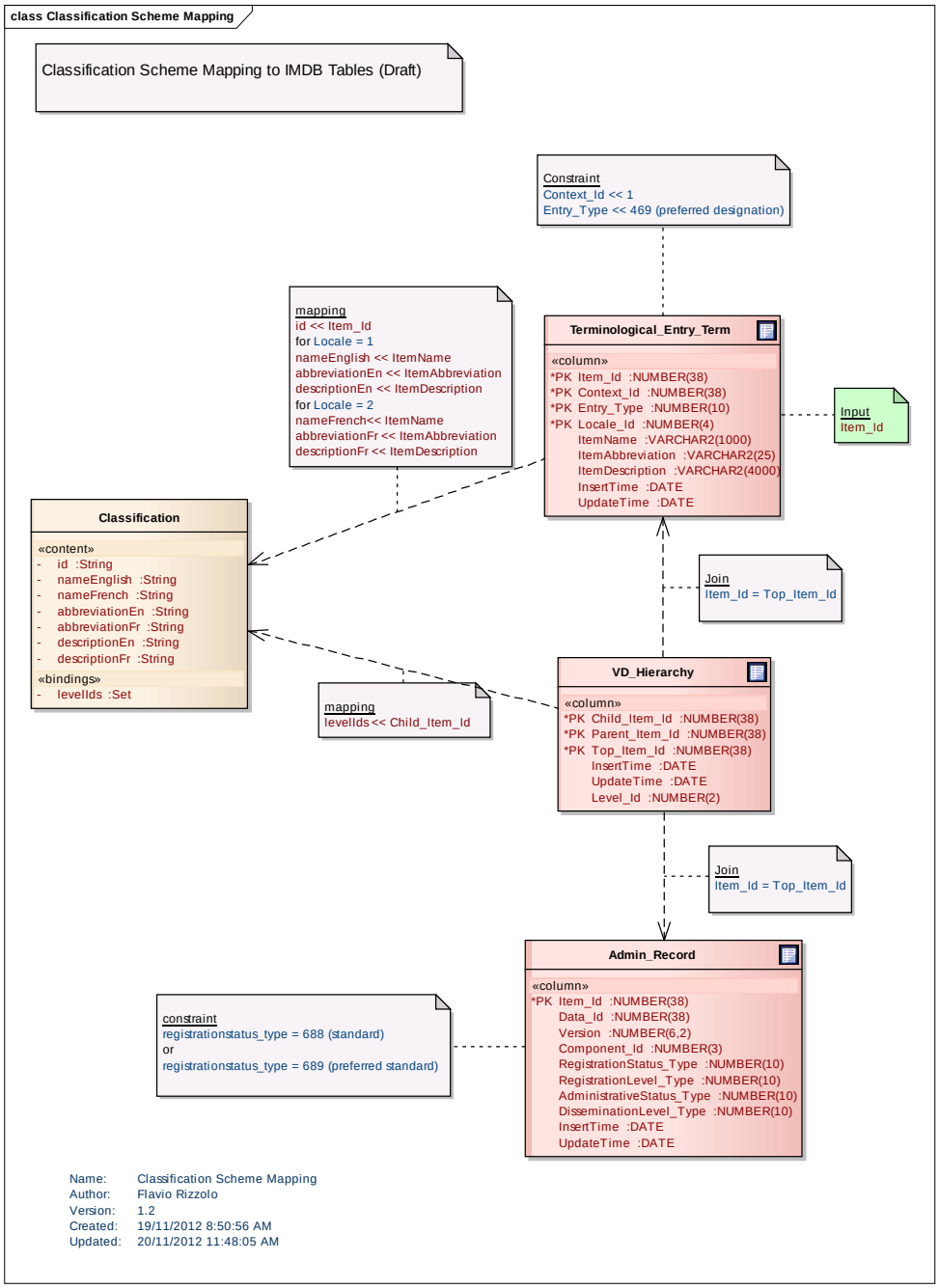

Figure 8: Physical Mapping from Canonical Classification Model to Integrated Metadata Base (IMDB) ISO/IEC 11179 Metadata Registries