3. Statistical Metadata in each phase of the Statistical Business Process (Statistics Canada)

3.1 Metadata Classification

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.

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.

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.[1]

-       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.

---

[1] For example: analyst comments about their analysis, output of statistical processes; respondent comments, interviewer comments or additional information about the respondent obtained during collection.

3.2 Metadata used/created at each phase

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.

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[1]). In addition, concordances to NAICS, NAPCS and other international classifications will be maintained in a Classification Management and Coding System (CMCS).

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.

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:

- variables, defined by purpose type (identification, statistical, processing and design variables) and by value source[2] (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.

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[3] .

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.

---

[1] See Section IV-F-(a) for more information on this service.
[2] It indicates the provenance of the value for data quality purposes.
[3] See Section IV-F for more information on SOA canonical models.

3.3 Metadata relevant to other business processes

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.

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.