WP.4
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Workshop on the Common Metadata Framework
(Vienna, Austria, 4-6 July 2007)

Topic 1: Models of the Statistical Cycle

INTEGRATED METADATABASE (IMDB) – A METADATA REPOSITORY TO SUPPORT THE SURVEY LIFE CYCLE

Statistics Canada[1]

I. INTRODUCTION

1. Statistical metadata is relevant to all stages of the statistical processing and there are a growing number of statistical agencies viewing statistical metadata as part of the whole survey life cycle, commonly known as end-to-end (E2E) metadata.

2. At Statistics Canada, a corporate repository and registry of statistical metadata, the Integrated Metadatabase (IMDB) is based on a model that supports the metadata required for the complete survey life cycle – from planning and design of a survey to archiving of master datafiles. The objective of this paper is to describe the metadata model that is being developed in Statistics Canada to support the metadata requirements of the agency's 590 active surveys and statistical programs. Although the metadata was first developed to meet the requirements for disseminating statistical data, there is growing internal pressure to reuse existing metadata and the administration layer of the model in other phase of the survey life cycle. Currently, we are developing the administered items for the questionnaire part of the metadata model as well as expanding the model for archiving data

3. The IMDB model is based on the ISO/IEC 11179 Metadata Registries and the Corporate Metadata Repository (CMR) from the U.S. Census Bureau. The CMR model consists of a data dimension model, business dimension model, administration and document dimension model, and terminology and classification dimension model.[2] For purposes of this paper, Statistics Canada's application in the IMDB of the CMR is described in detail: the data dimension model, business dimension model and questionnaire model, which links the business and data dimensions. Also, registration and classification of the metadata are described.

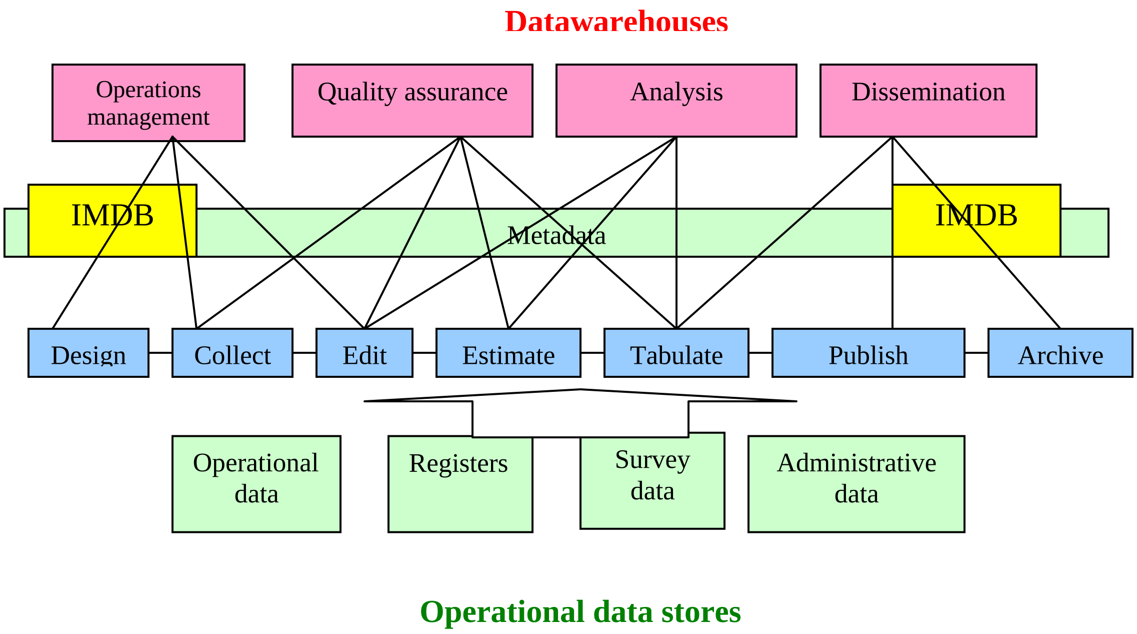
II. SURVEY LIFE CYCLE AND THE IMDB

4. Figure 1 shows where the IMDB currently supports the survey life cycle. While the metadata layer extends across all of the phases of the survey life cycle, metadata in the IMDB currently supports or will support disseminated data, archived datafiles, and the planning and design of surveys. However, metadata are

[1] Prepared by Alice Born, alice.born@statcan.ca
[2] Johanis, Paul and Dan Gillman, 2006: Metadata Standards and Their Support of Data Management Needs, Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS), Geneva, April 3-7, 2006.

derived from the different phases of the survey life cycle and stored in the IMDB. Also, metadata in the IMDB are linked to the Agency's various data products such as datawarehouses, which hold both micro- and macrodata; and may be used for data analysis (i.e., data benchmarking and data confrontation). The operational datastores hold the raw data collected from questionnaires (operational data), the registers (e.g., business register, address register, farm register, geographies) used for survey frames, imputed and estimated data (survey data) and administrative data. The relationship between the IMDB and the operational datastores has not been established.

Figure 1. The IMDB in the survey life cycle.



III. THE IMDB ARCHITECTURE

5. The IMDB data model is based on a modified version of the Corporate Metadata Repository (CMR) model, which is designed to support the metadata necessary to describe the survey life cycle, the linkages between the concepts and processes used across surveys, and the use of metadata to drive systems in support of the survey life cycle.[3] The model supports the metadata for a survey or statistical program over time and through the use of relationships.

6. The CMR is an extension of the international metadata standard ISO/IEC 11179 *Metadata registries-3*, which covers the full conceptual model for a metadata registry. The CMR model supports the production-oriented and output-oriented purposes of statistical information systems. Output-oriented information systems are commonly called data dissemination systems and are accessible on the Internet. Production-oriented information systems are related to design and processing systems. Increasingly, both types of systems have a strong metadata component to their design.

7. The IMDB model is organized as a set of overlapping dimensions. Each dimension is a model on its own, but the combination is greater that the sum of its parts. The main dimensions included in the IMDB are:

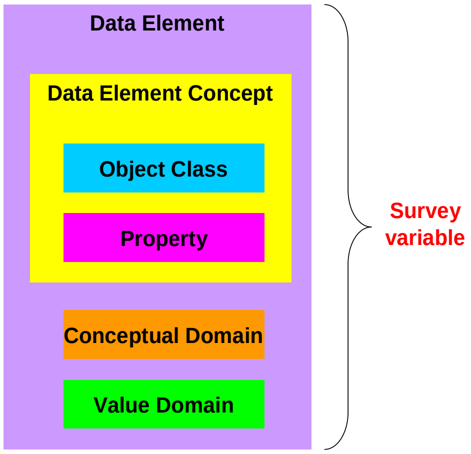
[3] Daniel Gillman, 2000: Corporate Metadata Repository (CMR) Model, U.S. Bureau of Labor Statistics.

1. Data dimension model;
2. Business dimension model;
3. Questionnaire dimension model; and
4. Administration and documents dimension model.

A: Data dimension model

8. The IMDB has adopted the data dimension model to describe the data and is based on the ISO/IEC 11179 standard. There are four classes in this dimension of the model: data element, data element concept, value domain and conceptual domain (Figure 2).

9. Data elements or variables are specified by an object class (statistical unit), a property (characteristic) and the value domain (set of permissible values, classifications). A detailed description of the data elements in the IMDB is provided in Part B of the Common Metadata Framework:[4]



10. Each of these components can stand on its own and is reusable in the construction of other data elements. By adopting this strategy that promotes the reuse of each of the components in combination, there are approximately 1,000 variables covering the entirety of the statistical output disseminated through CANSIM, Statistics Canada's online database (Table 1). However, there are additional data elements in the IMDB to support survey planning and design.

[4] Johanis, Paul and Dan Gillman, 2006: Metadata Standards and Their Support of Data Management Needs, Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS), Geneva, April 3-7, 2006.

Figure 2: Data dimension model.

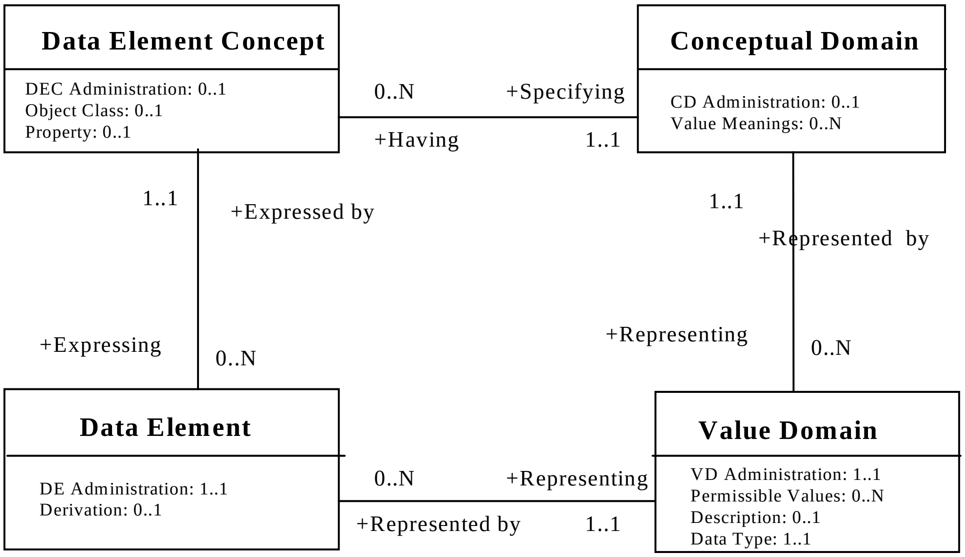


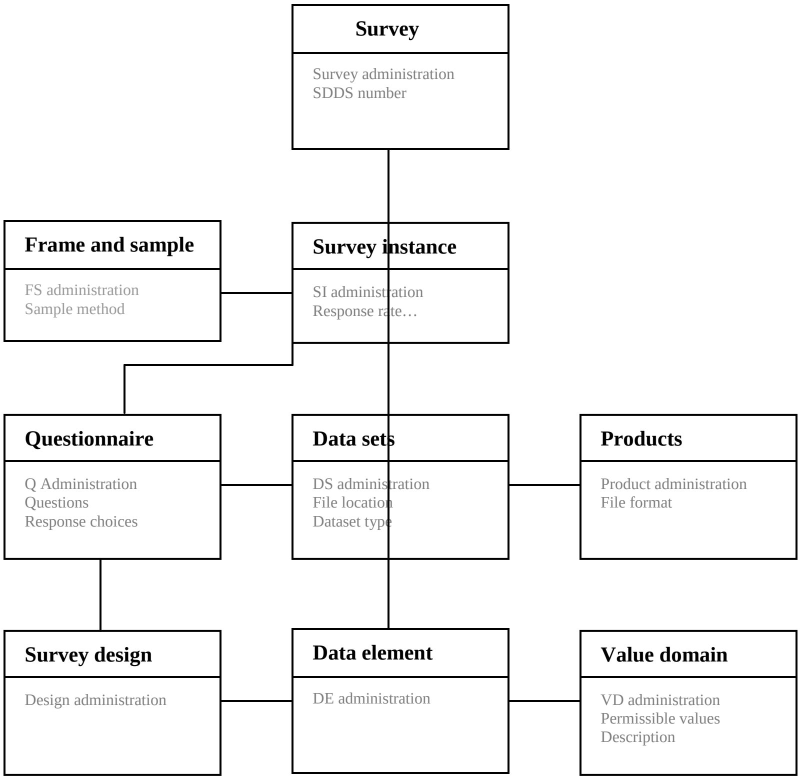
Table 1: Metadata in the IMDB supporting the data dimension model.

Object Class items	85
Property items	290
Data Element Concept items	506
Conceptual Domain items	202
Value Domain items	1509
Data Element items	1034

B. Business Dimension Model

11. The business dimension describes the survey designs, questionnaires, processing, data sets and products. It contains the metadata for the different phases of the survey. The IMDB has adopted a modified version of the CMR business dimension model to store metadata describing a survey and its documentation (Figure 3). The business dimension model has been expanded to show its link to the data dimension model.

Figure 3. Business dimension model in the IMDB.



12. The IMDB model defines the entities for describing Statistics Canada's surveys and statistical programs, their content and their methodology, and the relationships between them. The model supports the metadata requirements of many of the phases of the survey cycle including survey design, dissemination, post-survey evaluation and data archiving. The *Systems* model is outside the IMDB or part of the administrative layer in the metadata model. For example, the Software Register contains metainformation on the Agency's software and applications, which is linked to the *survey* in the IMDB. There are also metainformation systems that support the data collection and data processing phases of the survey life cycle, which are not part of the IMDB.

13. The basic structure of the metadata in the IMDB is illustrated in Figure 4. Each entity is referred to as an *administered item*. Each of the administered items currently in the IMDB represents a part of the survey life cycle[5] and the Data Dimension of the CMR (i.e., data elements and value domains). Administered items are defined, and may be reused or shared; and they are also managed, tracked and organized. In order to complete the latter, each administered item is supported by the following "regions", outlined in red in Figure 4. The *stewardship* region (e.g., organization, contact and documentation) supports the administration aspects of the administered item such as the responsible division and information for registration as well as supporting documentation. The *identification* region (e.g., identification and time frame) manages the name of the administered item and the time context for the administered item. The *classification* region (e.g., keyword and themes) manages the classifications and keywords to which administered items are assigned. In Statistics Canada, some administered items (e.g., surveys and

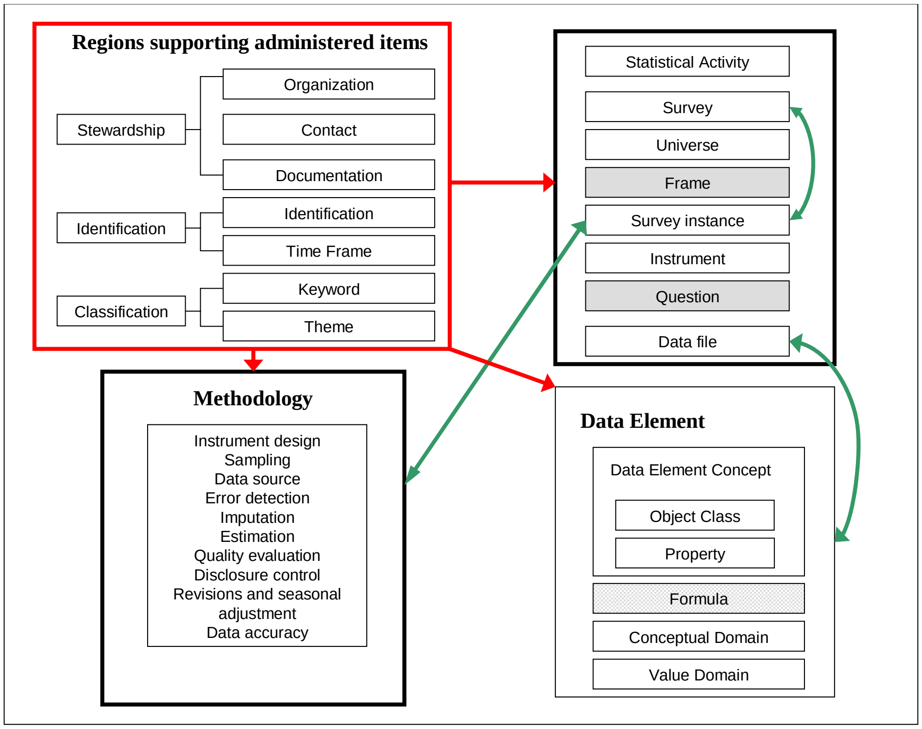
[5] The administered items supporting the survey life cycle in the IMDB match well to the proposed components in Part C of the METIS framework for statistical metadata (see Graeme Oakley, 2006). The IMDB also closely follows the administered items in the Business Dimension Model of the CMR.

questionnaires), data tables, data releases and publications are organized around 27 top themes and 221 sub-themes. Those administered items shown in grey have not been implemented in the current version of the IMDB.

14. In Figure 4, the administered items have been grouped into items that support information about the survey and its “umbrella” statistical activity; the survey methodology; and data elements. The green arrows show some of the relationships between these administered items. In the model, all the administered items describing data sources and methodology (i.e., methodology box) are attached to the survey instance; survey instances are linked to the survey; and data elements (variables) and value domains (classifications) are linked to the data file.

15. The administered items in the current version of the IMDB are: 1. Statistical activity; 2. Survey; 3. Instance; 4. Universe; 5. Instrument; 6. Methodology; 7. Documentation; and 8. Data Files. These administered items reflect the mandatory requirements for reporting information on data sources, methodology and data accuracy for each survey as stated in the *Policy on Informing Users of Data Quality and Methodology*.

Figure 4: Administered items in the IMDB.



C. Statistical Activity

16. The Statistical Activity administered item in the IMDB represents groups of surveys that share some common features and for which some common explanatory text would be useful to data users. For example, a Statistical Activity record was created for Statistics Canada's Unified Enterprise Survey (UES) Program, which contains general information applicable to 200 separate business surveys into a single master survey program. Another example is the Canadian System of National Accounts.[6] Not every survey needs to be linked to a Statistical Activity and is only created in consultation with subject-matter areas.

D. Survey

17. The content of the IMDB is organized around the survey entity as opposed to datasets as in other metadata models such as SDMX. A survey, in order to be considered as a record in the IMDB, is defined as a statistical activity that involves the collection, compilation and publication of statistical data measuring characteristics of a population. In the IMDB, surveys are defined as three types:[7]

- **Direct**: microdata are collected directly from a respondent with the use of a Statistics Canada collection instrument (e.g., Labour Force Survey);
- **Administrative**: microdata are extracted from administrative sources from an external organization, which were originally collected their own purposes (e.g., vital statistics from provincial and territorial governments); and
- **Derived**: data are derived from other Statistics Canada surveys or other data sources to produce datasets of new derived variables (e.g., national accounts, Gross Domestic Product, price indexes).[8]

18. The following guidelines are used to determine whether or not a "statistical activity" is a survey, and therefore requiring a record in the IMDB.

19. Activities producing clean microdata serving as data sources to surveys or to analytical studies, and for which no aggregated data are published, do not constitute surveys for the IMDB purposes. Direct surveys and administrative data can be active, discontinued or be conducted one-time only. Derived surveys can only be active or discontinued. One-time only derived statistics are considered as an analytical study and are therefore considered out of scope for IMDB. A compilation of selected data collected from direct surveys or administrative sources does not constitute a derived survey, even if it is produced on an on-going basis. In general, these statistical compendia, such as Statistics Canada's *Canadian Economic Observer*, are treated as a product and not as a survey.

20. Contrary to direct and administrative surveys, it is difficult to establish clear operational criteria for the designation of derived statistics. The extent of the transformation of the source data to produce new information is the critical factor, which cannot be quantified to establish an absolute rule. In the case of an activity drawing on data collected by others to produce a new dataset, one must ask oneself if referring the users to the metadata in the IMDB on the source surveys will adequately inform them on the quality and methodology of the product. If the answer is yes, the creation of a new derived survey and associated metadata is not called for; if the answer is no, then the statistical activity leading to the product should be designated as a derived survey.

[6] http://www.statcan.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=1735&lang=en&db=IMDB&dbg=f&adm=8&dis=2.

[7] For purposes of presentation, surveys are referred to as "surveys and statistical programs" on the IMDB web pages as a way of representing all three types of surveys.

[8] Derived statistics are referred to as "statistical programs" in the IMDB.

21. The survey administered item contains the following information: the title and acronym of the survey, (i.e., Monthly Survey of Manufacture (MSM)); an overview of the survey that provides a description of the objectives of the survey, the survey population, for whom the data are intended and the use of the data; and the status of activity on the survey (e.g., active, discontinued, transferred or one time only). All surveys are assigned an identification number, known as the Statistical Data Documentation System (SDDS) number in the IMDB.

E. Other Administered Items

22. A Survey consists the administered items related to the survey life cycle that are grouped together through an Instance administered item for each reference period of the survey. Table 2 shows these administered items and their respective definitions. The indentations of each item in the table illustrate the hierarchical relationship between the entities. These administered items are reused or updated in successive survey instances or shared with other surveys and statistical activities in the generation of web pages. Through the Policy on Informing Users of Data Quality and Methodology, the IMDB has a **common metadata set**[9] that is reused for each survey.

23. On the Statistics Canada website, users have access to metadata for each instance of each survey or statistical program for which data are disseminated. The administered items stored in the IMDB have been organized to present general information on the survey (survey title, status, frequency, record number and survey mandate) and metadata related to the survey life cycle for a survey instance (e.g., reference period, data release date, survey instrument (questionnaire), variables, survey description, data sources, methodology, data accuracy, documentation and data file (available internally only)). Quality metrics such as response rates and coefficients of variation are disseminated under Data Accuracy.

Table 2. Administered items in the IMDB and their definitions.

IMDB administered items	Definition
Statistical activity	The statistical activity is groups of surveys that share some common processing system or conceptual framework. For example, a statistical activity was created for the Unified Enterprise Survey Program, which contains general information applicable to all surveys in the UES program. Not every survey needs to be linked to a statistical activity. A statistical activity record is created in consultation with subject-matter areas.
Survey	A survey is a statistical activity that involves the collection, compilation and publication of statistical data measuring characteristics of a population. Includes direct surveys, administrative surveys and derived surveys. Provides administrative details about the survey.
Target population (or universe)	The population of units that is actually covered by the survey. Identifies the statistical unit and any of its relevant characteristics that are used (e.g., Canadian population aged 15 and over and not residing in institutions; or, establishments in NAICS industry XXXXXX with revenues over a certain threshold). Where applicable, any differences between the survey population and the target population (i.e., population for which information is desired) of the survey are described.

[9] The term metadata set is taken from the SDMX initiative. **Metadata set** is a set of information regarding almost any object that describe the maintainers of the data and structural definitions; describe the schedule on which data is released; describe the flow of a single type of data over time; describe the quality of the data, etc. In SDMX, the creators of reference metadata may take whatever concepts they are concerned with, or obliged to report, and provide a reference metadata set containing that information. Statistical Data and Metadata Exchange Initiative (SDMX), 2005: Framework for SDMX Technical Standards (Version 2.0), November, 2005.

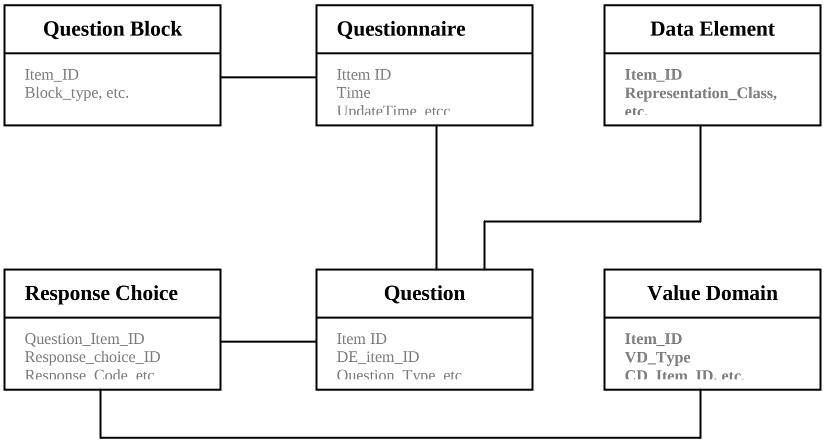
Survey Instance	Refers to each time the survey process occurs (i.e., each cycle of a survey). For each reference period, a new version of the instance record is created. For example, for a monthly survey, the IMDB will contain one instance record for every monthly cycle of the survey.
Collection instrument	The vehicle used for the collection of data. For direct surveys, it is a questionnaire. For administrative surveys, it is the record layout of the input record. It is not applicable for derived surveys. A questionnaire can be in many forms such as paper version, electronic, etc. Each questionnaire is linked to the instance record to which it pertains. The questionnaire image is copied into IMDB in pdf format.
Methodology	A text description of each of the following aspects of the methodology of a survey.
Instrument design	Method used to design, test and implement survey instrument. It only applies to direct surveys. Description of the design; methods for testing the questionnaire (e.g., review committee, focus group, pilot survey, etc.); and date of last revision of the instrument.
Sampling	Description of the survey units, any stratification used and the sample selection methods. It does not apply to derived surveys.
Collection method	Details on the collection methodology and the type of instrument. Details on method of initial contact and follow-up. Also included is a description of the capture method. For administrative and derived surveys, this item can be used to describe data sources. Collection methods include: data collected directly from respondents with a use of a collection instrument include: electronic data interchange; respondent completed – paper format (mail or fax); respondent completed – touch-tone telephone; Computer Assisted Telephone Interview (CATI); or Computer Assisted Personal Interview (CAPI) methods; or data extraction from administrative files provided by an external organization; or data derived from other Statistics Canada surveys.
Error detection	Methods used to detect errors during collection, capture and processing of the data. Provides details on the types of edits used, ratios applied, etc. and identifies at which stage of the survey process it is done (i.e., during collection, or as part of processing.)
Imputation	Process used to replace missing microdata, and invalid or inconsistent responses identified during editing. Provides details on the type of imputation (e.g., manual, automatic, etc.), imputation rates, the method used (e.g., historical, hot deck, donor, etc.), and software used. Usually does not apply to derived surveys.
Estimation	Methods used to produce estimates for the survey population from collected data. It includes non-response macro adjustments, post stratification, calibration, weight-share methods, and variance estimation methods (e.g., direct, Taylor, Jackknife, Bootstrap, etc.). In the case of administrative and derived surveys, procedures and models used to produce the indicators are described.
Quality evaluation	Methods undertaken to evaluate the quality of the final data. Procedures include data confrontation with other published sources, re-interviews, reverse record checks or historical trend analysis.
Disclosure control	Measures taken to ensure that data from the survey does not disclose information concerning any identifiable respondent thus maintaining confidentiality of the respondent data. This summary can include for micro data, removal of respondent, content reduction and content modification; for tabular data, sensitive cells, correction methods such as collapses/suppress cells; and revisions by committees.
Revisions and seasonal adjustments	Methods used to adjust estimates in relation to same estimates for prior periods including benchmarking, calendarization or seasonal adjustments; and procedures for regular revisions to data.

Data accuracy	Data accuracy indicators for the survey. These measures include the coefficient of variation for the key variables in the survey, information on coverage error, response rates and any other relevant data accuracy indicators. Includes response bias and error, and processing errors.
Documentation	Documents useful for the users' understanding of the data can be linked and include user guides, data dictionaries, technical notes, etc.
Data file	Information on the location, format and content of clean data master files that are used as inputs to surveys or as outputs of surveys. The IMDB stores information on clean data master files that are produced for each instance of a survey.
Variables	Description of the meaning of a data point. Based on ISO 11179.
Statistical unit	Definition of the unit about which data are collected (e.g., establishment, household, person and births).
Property	Definition of the characteristic of the statistical unit.
Representation class	Describes the specific form of the representation of the property (e.g., type, name, category, value, area, index)
Classification or unit of measure	A set of allowed values that a variable may take. Classifications are used to represent categorical data and units of measure are used to represent quantitative data (e.g., dollars, tonnes)

F. Questionnaire Model

24. The questionnaire model illustrates the links between the questionnaires, questions, question blocks and data elements in the IMDB (Figure 5). Data element is linked to questions because it expresses the concept that is being measured by the question. Value domains are linked to response choices because each describes the valid values the data can take.

Figure 5. Questionnaire model in the IMDB.



25. The role of the IMDB is to support metadata discovery of existing data elements (variables), object classes (statistical units), properties, value domains (classifications); and questions and question blocks related to these administered items. By storing these items in the IMDB, it will promote the re-use of variables, questions and response choices as well promote coherence across surveys at the time of questionnaire development or redesign. Also, the IMDB will allow survey managers to register new survey content – variables, questions, etc. at the time of planning and design.

26. The IMDB does not store the specifications for questionnaires nor their applications. Procedural or operational metadata (i.e., paradata) will be in other metainformation systems in Statistics Canada such as Survey Specification Manager for household surveys and the Integrated Questionnaire and Metadata System for business surveys. Work is underway to link the metadata from IMDB's questionnaire model to the Survey Specification Manager, which is a toolkit of questions, question blocks and code for household surveys. Ideally, the SSM would "pull" approved questions and their response choices already registered in the IMDB, and other related metadata (e.g., data element concepts, data elements, value domains) as the starting point for questionnaire development or redesign. The goal of linking the Agency's metadata and metainformation systems is to reduce the risk of applications having incompatible concepts and models and making integration ex post resource intensive or impossible as suggested by Statistics Austria.[10]

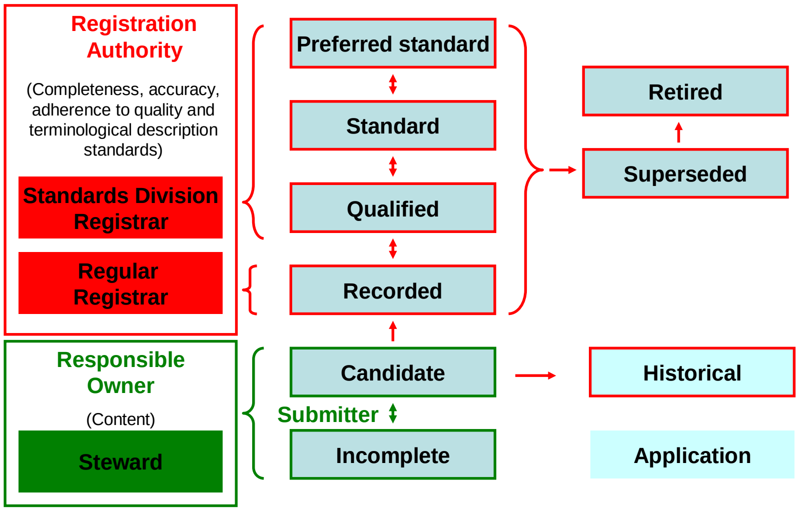
IV. REGISTRATION IN THE IMDB

27. Registration is a set of rules, operations and procedures that apply to the IMDB. Registration is part of the Administration and Documents Dimension Model of the original CMR. The purpose of registration is to monitor the source of the metadata, quality of the metadata and assigning an identifier to each object described. In the IMDB, there are three registration attributes: 1. registration status, 2. registration level, and 3. administrative status. It should be noted that the IMDB does not exactly conform to the ISO/IEC 11179 Part 6 - *Registration*. Currently, object classes, properties, data element concepts, data elements and value domains are following this registration process.

28. The registration status category identifies the quality or progression quality of administered items (AI) in the IMDB. It identifies the AI's completeness, accuracy and conformance to syntax and format within the registry. IMDB's use of the registration status categories is as follows: An AI entry is ready for progression from the "incomplete" status to another registration status category only when the all the information content of the AI entry is complete, conforms to quality and terminological description standards in both official languages and is recorded with the quality as though it is a "Standard" or "Preferred Standard" AI entry. The steward's responsibility is to ensure the accuracy and completion of the information. An AI entry will remain at an "incomplete" status until the information content is deemed of sufficient quality by a steward or submitter for promotion into the registration process. Once promoted into the registration process, the registrar determines the registration status of the item based on the compliance of IMDB's assignment criteria. The Registration status determines the registration privileges granted to a registrar (Figure 6).

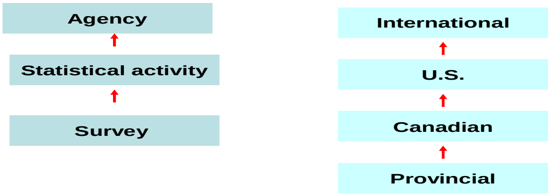
[10] Guenther Zettl, 2007: IMS (Integrated Metadata System) – An architecture for an expandable metadata repository to support the statistical life cycle; UNECE Workshop on the Common Metadata Framework, Vienna, Austria, July 4-6, 2007.

Figure 6. Registration status



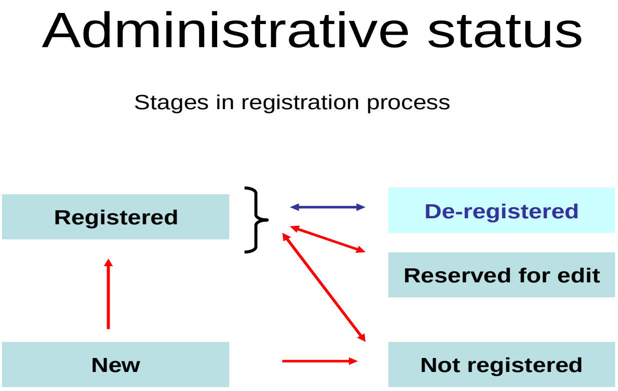
29. The registration level identifies the level of conformance within the Agency, or domestically and internationally (Figure 7).

Figure 7. Registration level.



30. Finally, the administrative status describes that stage of the registration process (Figure 8).

Figure 8. Administrative status



V. CLASSIFICATION OF ADMINISTERED ITEMS IN THE IMDB

31. In order to organize administered items, the UNECE Classification of International Statistical Activities has been adopted. There are three levels to the original classification. The first level consists of five *statistical domains* of which three are related to subject-matter areas: 1. Demographic and social statistics; 2. Economic statistics; and 3. Environment and multi-domain statistics.[11] The second level specifies the *statistical areas* (24 classes) in the statistical domains, and the third level (12 classes) indicates more detail. The third level is used only where necessary, and additional items can be added to the classification.

32. This classification has been adopted by the Statistical Data and Metadata Exchange (SDMX) initiative for the development of its own statistical domains needed for the registration of data and metadata. The UNECE classification reflects the themes of data and statistical metadata generally produced by national statistical offices, and was selected by Statistics Canada for this reason. It is also being considered by Australia, France, Japan, Korea, Iceland, The Netherlands, and Sweden; and has been implemented by the U.S. Federal Reserve Board.

33. The classification is easy to use, and lends itself to being tailored at the most granular level for the specific needs of the NSO. The classification follows the rules of classification, and most data elements in the IMDB do not need to be cross-classified. All of the data elements presently in the IMDB have been assigned to a class in the classification (a few to more than one) and the classification has been modified to better correspond to the subject-matter of the data elements in the IMDB. Approximately 1020 disseminated data elements have been classified according to this list with 429 in the Demographic and Social Statistics Domain and 576 in the Economic Statistics Domain. The IMDB-version of the classification will be used to classify other administered items such as value domains and surveys.

VI. CONCLUSION

34. The IMDB had adopted a modified version of the Corporate Metadata Repository (CMR) to describe the statistical data and the survey life cycle - questionnaires, methodologies, data accuracy and other related attributes of surveys and statistical programs. The ISO/IEC 11179 standard is used to describe the concepts and terms of the data. The IMDB metadata model has the capacity to expand to support other part of the survey life cycle, however, for now it will continue to support the planning and design, the dissemination and archive phases. However, as part of the Agency's enterprise architecture, it is encouraged that other metainformation systems reuse the administered items already registered in the IMDB.

[11] The statistical domains, 4. Methodology of data collection, processing, dissemination and analysis; and 5. Strategic and managerial issues of official statistics, have been omitted for purposes of classifying administered items in the IMDB for now.