

Distr.
GENERAL

CES/SEM.47/23
16 January 2002

ENGLISH
ENGLISH AND FRENCH ONLY

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE
EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint UNECE/Eurostat Seminar on Integrated Statistical
Information Systems and Related Matters (ISIS 2002)**
(17-19 April 2002, Geneva, Switzerland)

Topic IV: Ways of making statistical information systems more responsive to users

METADATA AT STATISTICS CANADA

Invited paper

Submitted by Statistics Canada ¹

I. INTRODUCTION

1. Over the past few years, there has been a proliferation of data repositories and data warehouses and with the maturation of technology, the data are readily accessible via the Internet. Data users are faced with the enormous task of sifting through the information, evaluating the comparability of the disparate data sources and then determining the relevance and the quality of the data for their needs. A data producer is faced with similar tasks-"Is there existing data that can provide additional value or streamline the production process?" "Has anyone else already produced the same or similar data from which I can benchmark the quality?" Although there are search engines and end-user analysis tools to facilitate data search, data analysis and data presentation, it is the meta information and the metadata² that give the data meaning and context.

2. This paper discusses the evolution of meta information and metadata at Statistics Canada, the software tools used for development of the Integrated Meta Database (IMDB), the passive and active roles of the information, integration of the IMDB with existing systems and their business processes and access paths to the information stored in the IMDB.

3. The IMDB is a corporate registry of meta information and metadata that are necessary to clearly describe, inventory, analyse and classify statistical information and statistical data at Statistics Canada.

¹ Prepared by Amie Lee.

² Meta information refers to the concepts and definitions describing the business processes that create the data. Meta information is used to evaluate and understand the survey data. Metadata refers to the description of the physical data. Metadata facilitates access to the survey data and supports data sharing across systems and organizations. The term *metadata* will be used to refer to meta information and meta data collectively.

II. HISTORY OF META INFORMATION AND METADATA

4. Extensive and good quality *metadata* are available for Statistics Canada's data, although it is distributed and self contained among various media such as in design specifications in a Computer Assisted Software Engineering tool, sampling, editing or imputation applications/systems, data bases, data warehouses, data repositories or in the form of a publication or document.

5. At Statistics Canada, there were a number of initiatives to create a *metadata* repository. The first initiative was the Statistical Data Documentation System (SDDS). This system contained descriptive information about the surveys conducted at Statistics Canada. Standards Division updated the survey information annually by contacting program managers in each of the data producing areas. The content of the Statistical Data Documentation System did not adequately provide timely meta information and did not serve the metadata needs of the agency.

6. In the 1990s, the Thematic Search Tool using Folio Views was developed and centrally maintained for social statistics. This system provided good thematic access to the meta information for social statistics, however the system did not contain standard use of vocabulary or terms, required an enormous amount of manual processing and maintenance for data collection and data extraction to maintain the thematic classification. The Paradox Electronic Collection Template was developed in the 1990s to alleviate some of the collection and maintenance issues of the Thematic Search Tool. This tool was deployed to only the social statistic subject matter areas and it was the responsibility of the program managers to ensure the data was updated annually. The Paradox Electronic Collection Template did not support data sharing because each of the subject matter areas had their own instance of the database. The hardware platform, software tools and database design used to build the collection template were not sufficiently robust to support use in the entire agency.

7. In addition, in the 1990s, Data Access and Control Services Division developed and maintained the Meta Inventory of Data Assets System. This system stored archival information for all of Statistics Canada's clean statistical master data files. Data Access and Control Services Division updated the information annually by contacting the program managers in the subject matter areas. The data served the requirements of archivists from National Archives.

8. Each of the systems described above collected their own *metadata* separately. The information in addition to being inconsistent from system to system, was a response burden for the program managers because each system collected similar information at different times during the year.

9. In 1998, the Integrated Meta Database Project was started to create a corporate wide *metadata* repository. The project's scope was to identify *metadata* requirements and to consolidate the various sources of meta information and metadata into one central repository for all statistical activities of the corporation. The goal was to eliminate duplication of information among the various systems, to provide better access to information to both internal and external clients, to minimise the response burden by the subject matter areas and to assist in harmonisation of concepts and variables. The Integrated Meta Database (IMDB) repository and system is in production with update currently limited to Standards Division³.

³ The system is to be deployed corporately once user access and security, registration and versioning modules are completed.

III. CONTENT DEVELOPMENT

10. The meta information content of the IMDB was guided by the following:
 - Statistics Canada's Policy on Informing Users of Data Quality and Methodology;
 - Statistics Canada's Quality Guidelines;
 - Statistics Canada's Policy on Standards;
 - Statistics Canada, An Overview of Business Survey Processing, April 1999;
 - United States Bureau of Census's statistical metadata repository model (MDR) ⁴.
11. From June 1998 to September 1999, the meta information requirements were determined through consultation with Dissemination Division, Data Access and Control Services Division, Project to Improve Provincial Economic Statistic representatives, Data Liberation Initiative representatives, Business Statistics Methodology Division, Social Statistics Methodology Division, Standards Division, Library and Information Centre, various subject matter areas and Systems Development Division⁵. The issue faced during these consultation sessions was a timing issue. Subject matter areas had their own specific *metadata* requirements for their production schedules and as a result, they developed their own *metadata* systems during the design and development of the IMDB. The *metadata* in these systems will be addressed in the future⁶.
12. In June 2001, the metadata content of the IMDB was guided by the following:
 - United States Bureau of Census statistical metadata repository (MDR);
 - ISO/IEC CD 11179-3, Information Technology – Data Management and Interchange – Metadata Registries (MDR) – Part 3: Registry Metamodel (MDR3), June 6, 2001-12-11;
 - ISO/IEC PDTR 20943-3, Information Technology – Achieving metadata registry (MDR) content consistency – Part 3: Value domains, February 26, 2001-12-11;
 - Neuchâtel Group, The Neuchâtel Terminology: Classification Database Object Types and their Attributes.
 - Statistics Canada, An Overview of Business Survey Processing, April 1999.
13. Standards Division, Systems Development Division and Culture, Tourism and the Centre for Education Statistics⁷ reviewed the metadata requirements and extended the IMDB model to provide concordance across Value Domains and to provide future automated coding capabilities for Statistics Canada Industry Classification Coding System (ICCS) system.

IV. COLLECTION PROCESS AND COLLECTION TOOLS

14. The IMDB collection screens and special load utilities were developed for the collection and verification activity⁸. The collection and verification of *metadata* is a major task because of the sheer quantity that is required to describe the survey activity and its data outputs. Although collection and update is presently done centrally through Standards Division, *metadata* update triggers have been identified in the existing dissemination process. Integration of the *meta data* collection with the existing workflow results in

⁴ Dan Gillman sent the model via e-mail.

⁵ Overview of the IMDB content is described in a paper prepared by Paul Johanis, "Statistics Canada's Integrated Meta Database Current Status and Future Plans", Work Session on METIS, 28-30 November 2000, Washington, D. C.

⁶ The *meta data* in these systems could be a data source used to populate the IMDB or the IMDB would replace these systems entirely. The issue will be discussed once business processes for interaction with the IMDB has been fully defined.

⁷ The representatives were from the education program that is participating as a pilot program for development of education content.

⁸ As of December 2001, about 80% of the input screens have been delivered for production in Standards Division.

the *metadata* being collected passively, however it increases the probability of the *metadata* being kept up to date since these triggers act as *metadata* checkpoints. The active collection of the *metadata* will become a reality when *metadata* takes a more active role in the planning, development and management of survey activities and survey products.

A. Meta Information Load

15. The meta information from the following systems was incorporated into the IMDB and the systems were retired:

- Statistical Data Documentation System;
- Thematic Search Tool;
- Paradox Electronic Collection Template;
- Meta Inventory of Data Assets System;
- Questionnaire Inventory System⁹.

The meta information from each of these systems was consolidated and loaded into the IMDB using specially developed conversion tools in Access97 for input and analysis and Oracle PL/SQL procedures to load data into the IMDB. Standards Division was responsible for the management of the conversion activity, information consolidation and verification of the data loaded. The Access97 tools and Oracle PL/SQL procedures were retired when the load into IMDB was complete. Meta information content was also gathered from publications and manually entered using the IMDB collection screens.

16. The verification was very labour intensive and time consuming because:

- The overlapping content from the retired systems required verification [and approval by each of their respective system managers and program managers from the subject matter areas.]
- The information must be available in both official languages¹⁰ and the textual descriptions must be equivalent in both official languages.

17. The verification process was streamlined by using the existing processes/vehicles for dissemination of the meta information. Standards Division had already published the SDDS content on Statistics Canada's intranet and internet via manually produced HTML pages. Program managers used these pages to review and forward updates to Standards Division for update of the SDDS system. These manual HTML pages were replaced with static HTML pages automatically generated from the IMDB repository and the update process via the program managers remained the same.

B. Metadata Load

18. The initial load of IMDB metadata content is restricted to the output variables disseminated on CANSIM. This load activity is another labour intensive process because it involves the development of standardised and harmonized content, which requires a review of all CANSIM variables and their definitions. The harmonisation is attained through the development of standard value meanings and then the identification of the set of permissible values that link to the standardised value meanings. The content development includes definition of Object Class, Property, Data Element Concept, Conceptual Domain, Value Domain and Data Element for the CANSIM output variables. Standards Division is responsible for the review and development of this content. The verification process is very labour intensive for the same reasons as described for the meta information load.

⁹ The Questionnaire Inventory System consists of questionnaires used for each fiscal year. Standards Division maintains the inventory.

¹⁰ French and English

19. The advantage of a centralized load, update and verification by Standards Division is:
- Conformance to Statistics Canada's policies and guidelines;
 - Information content is consistent for all *metadata* across subject matter areas;
 - Information quality is consistent for all *metadata* across subject matter areas;
 - *Metadata* providers have a concrete example of the level of detail desired;
 - *Metadata* providers have a concrete example of delivery style;
 - *Metadata* providers have one point of contact, therefore minimising their time commit required.
 - Harmonisation is attainable.
20. The disadvantage is Standards Division is faced with the enormous task of entry and management of the information in addition to content analysis and development. A number of tools were developed to assist in this activity.
21. A loader/exporter was developed using a Visual Basic 5.0 and XML-DBMS version 1.01 to create an XML based solution to load and export standard classifications (i.e. NAICS)¹¹.
22. An Access97 Gathering Tool was developed to gather metadata from CANSIM and to allow Standards Division to analyse and prepare the metadata for load into the IMDB. An Access solution was chosen to leverage the knowledge of Access. This tool will be retired once the CANSIM metadata are loaded into the IMDB.

C. IMDB Collection Screens

23. Design goals of the *metadata* collection screens were to:
- Divide the amount of information into manageable pieces which will provide a modular delivery for testing and production;
 - Facilitate collection by optimizing the amount of information on one screen, but yet providing enough information for intuitive collection by subject matter areas;
 - Minimise the time required for development;
 - Minimise the time required to complete system testing;
 - Minimise impact of changes on system as a result of changes in requirements;
 - Minimise the time of user training required to use the input system.
24. Feedback received from users of the Paradox Electronic Collection Template was incorporated into the IMDB collection screens. The items addressed were:
- Users have the choice of a French or an English user interface with simultaneous display of both the French and English¹²;
 - Provide an alternative to question and answer style collection¹³;

¹¹ XML-DBMS is a system for transferring data between XML documents and relational databases. It views an XML document as a tree of objects and then uses an object-relational mapping to map these objects to a relational database. This version of the importer/exporter is used for the initial load of the classification data into the IMDB. The decision to load and export only the standard classifications was based on the standard classifications used in CANSIM and by ICCS. As of December 2001, the IMDB loader/exporter was in test phase. Load of standard classifications into the IMDB is expected to be completed by March 2002. The IMDB importer/exporter functionality and environment is to be reviewed and expanded to cover other components in the future.

¹² French or English interface means the directives, data labels, user messages and help message appear in the language chosen by the user. French and English text data fields are always displayed together in the interface and meta *data* providers are required to always provide both the French and the English text.

- Effective access to input areas without extensive scrolling;
- Ability to customize labels and questions.

25. The object-oriented methodology addressed the design goals of the input system. All administered components are managed in the same manner. From a collection point of view, search screens and input screens for all administered components are identical in presentation and behaviour with exception of the information content that varied based on the type of administered component. This minimised the training required to use the input system. Once a user is familiar with the entry of one administered component, entry of other the administered components varied only in information content. The testing time is minimised because once one administered component has been tested, only the changes in content need to be tested for subsequent administered components. From a systems development perspective, the development time is reduced for each new administered component. The Administered Component class handles the administration aspect¹⁴ and all objects that are of type administered component are subclasses of the Administered Component class and therefore inherit its classes and methods. Developers only need to concentrate on the object's own classes and methods. The development time is reduced for each new administered component added to the system, however, the initial design time is increased because the designer must understand the entire system to design an architecture, which will effectively make use of object-oriented techniques and leverage the reuse of objects. The development team applied standards as much as possible at every stage of design and development through use of classical object-oriented design patterns and design and code reviews. The Unified Modelling Language (UML) was used to model the input application. The object-oriented development environment is significantly more technical and is a complete change of paradigm at Statistics Canada.

26. The IMDB collection screen is a two-tier thin client application. The data layer and the application layer are on the same tier using Oracle 8i relational database and stored procedures in PL/SQL. The presentation layer is a JAVA application¹⁵ developed using IBM Visual Age 3.0 (JDK 1.1.7, Swing 1.0.3)¹⁶, and Oracle JDBC drivers for the middleware. Other tools used for development include System Architect 2001 V7.12, TOAD VI V6.2.10.29 (Tool for Oracle Applications Developers) for development of the PL/SQL and Jprobe Profiler and Memory Analyser.

V. DISSEMINATION

27. The IMDB dissemination objective was to support the dissemination via the Statistics Canada web site (www.statcan.ca). The web site offers a broad range of information and statistics, which targets a variety of data users.

¹³ Once users were familiar with the system, the screen area used by questions was cumbersome and a hindrance for navigation.

¹⁴ Stewardship, identification and classification regions.

¹⁵ It was deployed initially as an applet, however because of the load time; it is deployed as an application on the server.

¹⁶ IBM Visual Age 3.5 (JDK 1.2, Swing 1.1) uses Persistence Builder Data to build data access classes in this version. Upgrade to this version requires redoing the data access classes because Data Access Builder was dropped in version of Visual Age 3.5. Conversion tools were not provided by IBM for Data Access Builder users to allow for upgrade from JDK 1.1.7 to JDK 1.2. Data Access Builder was the only tool available in Visual Age to build the database access classes when development of the IMDB input screens started. The development team is in the progress of building a code generator class used to build data access classes directly through JDBC. The access classes will have to be redone for the entire application before the upgrade can be done.

A. Passive Use

28. Static HTML pages are automatically generated from the IMDB daily for both the intranet and the internet.¹⁷ The pages contain up to date information for the latest survey cycle¹⁸. With the implementation of the IMDB, the content of the pages is augmented to include archival information, methodology summaries, images of the questionnaires and links to on-line documents. The generated HTML pages are accessible from Statistics Canada's web site via CANSIM, Canadian Statistics and Statistical Methods.

29. Other planned passive uses are:

- On-line Questionnaire Inventory generated from the IMDB planned for fiscal year 2002/2003.
- IMDB data element definitions accessible through the CANSIM search screens at the end of fiscal year 2001/2002. The function is currently in development using CGI-PERL 5.0 and a XML based solution.
- On-line classifications manuals generated from the IMDB planned during the fiscal year 2002/2003.
- Integration of the Corporate Software Registry and the IMDB is planned.

B. Active Use

30. A Cansim Transaction Formatter which will load data into CANSIM and metadata into the IMDB is scheduled for development in fiscal year 2002/2003. This will provide users of Information Retriever Meta-information Administrator (IRMA)¹⁹ the ability to link metadata in the IMDB with CANSIM data.

31. Other planned active uses of *metadata* are:

- Coding services through integration with Industry Classification Coding System (ICCS) system.
- Smart publishing of electronic publications using *metadata* and data from various sources.

C. Other dissemination factors

32. An architectural study was done in 2001²⁰ to identify interactions among Statistics Canada key reference databases²¹ and to propose technologies to support service delivery, data sharing and co-operative processing. This study is to serve as a guide for long-term future development efforts at Statistics Canada.

33. The Government On-Line Initiative (GOL) is the federal government's commitment to provide Canadians with on-line access to all federal information and federal services by year 2004 through the GOL Canada web site (canada.ca). The IMDB will be a key information source for this initiative.

34. Technology changes rapidly and data users and data producers are faced with the challenge to migrate to new technologies to make their information and data more accessible. The challenge is to choose

¹⁷ The internet version contains information designated for public consumption. The intranet version contains information designated for internal use.

¹⁸ Information on previous survey cycles is stored in the IMDB; however, this information will be accessible when the direct database search for information retrieval has been implemented. The search that exists presently in the input system is a search specifically for update.

¹⁹ IRMA is used to assemble micro data, aggregated data and metadata from various sources to build holdings of statistical products and to retrieval data and metadata for analysis or export.

²⁰ Hutton, T. and Graves, R. (2001), "Coherence of Reference Databases Architecture Study", Systems Development Division, Statistics Canada, Final Version, June 11, 2001

²¹ Integrate Meta Database (IMDB), Spatial Data Infrastructure, Business Register, Address Register and Tax Microdata

the technology or technologies, which offers the maximum longevity to balance the time and knowledge acquisition required to support the new technology. The Corporate Software Strategy Committee and Web Development Centre are key areas which can influence the selection of software tools available for use at Statistics Canada. With a harmonized set of software tools, data producers and data users would be able to manage their information and data more effectively and provide consistent access for supported technologies.

VI. CONCLUSION

35. The Integrated Meta Database Project concentration over the last three years has been in the development of *metadata* and loading into a central repository, the Integrated Meta Database. Existing dissemination access paths were used to provide users access to the IMDB data. The effort has been directed in the collection of the metadata as opposed to investigating new access paths because the underlying metadata is required before access paths are useful.

36. The collection and update activities are enormous tasks that require a major effort and commitment by the corporation. Object-oriented technologies, the intranet and XML were used to assist in these activities in an attempt to streamline the tasks for the IMDB project. Overall this has been a positive but challenging experience for the IMDB project.

37. The challenge now is to keep the IMDB up to date, both in terms of the data it contains and the technologies it uses (having embraced an object-oriented technology, the challenge is to avoid becoming obsolete as the object-oriented technologies are changing rapidly in the industry). Directions for the future include moving from a passive to a more active collection and use of *metadata* and providing users with improved access paths for both internal and external users to the *metadata* stored in the IMDB.

REFERENCES

Carnaham, L., Gallagher, L. (2000), "A General Purpose Registry/Repository Information Model", Draft, National Institute of Standards and Technology, Information Technology Laboratory.

Ehrenström, B. (2000), "Neuchâtel Group, The Neuchâtel Terminology: Classification Database Object Types and their Attributes", UN/ECE Metis Work Session, Washington D.C. United States, November 28-30, 2000.

Graves, R., Hutton, T., (2001), "Coherence of Reference Databases Architecture Study", Final Version, Statistics Canada, Systems Development Division.

Hutton, T. (1999), "An Overview of Business Survey Processing", Draft, Statistics Canada, Systems Development Division.

ISO/IEC 11179-1, "Specification and Standardization of Data Elements, Part 1, Framework for the Specification and Standardization of Data Elements", Committee Draft, October 1997.

ISO/IEC CD 11179-3, "Information Technology – Data Management and Interchange – Metadata Registries (MDR) – Part 3: Registry Metamodel (MDR3)", June 6, 2001-12-11.

ISO/IEC PDTR 20943-3, "Information Technology – Achieving metadata registry (MDR) content consistency – Part 3: Value domains", February 26, 2001-12-11.

Johanis, P., (2000), "Statistics Canada's Integrated Meta Database Current Status and Future Plans", UN/ECE Metis Work Session, Washington D.C. United States, November 28-30, 2000.

Smiderle, G., (2000), "Storing UES Metadata on the IMDB", Preliminary Report, Statistics Canada, Unified Enterprise Statistics Program.

Systems Development Division, (2001), "Pilot Project to Interface IMDB with the Cansim II Transaction Formatter", Specifications, Statistics Canada.

Statistics Canada, (2000) "Policy on Informing Users of Data Quality and Methodology".

Statistics Canada, (2001), "Policy on Standards".

Statistics Canada, (1998), "Statistics Canada Quality Guidelines", Third Edition.

Statistics Canada, (2001), "The Government On-Line Strategy and Statistics Canada". Statistics Canada's Public Report to Treasury Board on Government On-Line.

Systems Development Division, (2001), "Pilot Project to Interface IMDB with the Cansim II Transaction Formatter", Specifications, Statistics Canada.