WP. 16
ENGLISH ONLY

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

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Luxembourg, 9-11 April 2008)

Topic 2 (iii): Metadata and the Statistical Cycle

METADATA REQUIREMENTS FOR ARCHIVING STRUCTURED DATA

Submitted by Statistics Canada, Canada[1]

I. INTRODUCTION

1. The final step of data stewardship is that of archiving the final master file. In Canada, the Library and Archives of Canada Act along with the federal government's Policy on Information Management requires that government institutions fully protect and safeguard for future generations those government records deemed by the National Archivist as warranting preservation. In the statistical program, a Record Disposal Authority is required before a survey master file can be destroyed.

2. Statistics Canada initiated in the mid-1980s an archival system for final master datafiles, which has subsequently been incorporated into the Agency's Integrated Metadatabase (IMDB). Among other statistical metadata, this metadatabase contains descriptive information concerning the location, format and content of the confidential master datafiles for all of Statistics Canada's surveys. The confidential master datafiles and all electronic metadata elements are preserved and routinely exercised to ensure their continuing accessibility. The Agency is currently revaluating the metadata requirements for archiving all structured data, and developing business rules for archiving and disposing of datafiles.

3. During the METIS Workshop on Part C of the Common Metadata Framework (CMF), held in July 2007, a generic model of the statistical business process was adopted. It was also agreed that with the addition of 'Archive' and 'Evaluate' phases the model currently used by Statistics New Zealand would provide a better basis for the generic CMF model. In this paper, the processes and sub-processes have been developed for the archive phase of the model. This is the first attempt at providing users the processes for archiving structured data and is subject to further revisions.

4. The focus of this paper is the metadata requirements for archiving of structured data from national statistical offices (NSOs). The paper begins with a description of where archiving fits into the IMDB metadata model. It also presents a proposal for the archive process as part of the statistical business process model and what are the sub-processes. Finally, it describes the potential role of the IMDB in documenting and managing the metadata for these archived datafiles, and provides examples of metadata elements necessary to support archived datafiles. This phase in the development of the IMDB will bring it closer to end-to-end support of the statistical life cycle at Statistics Canada.

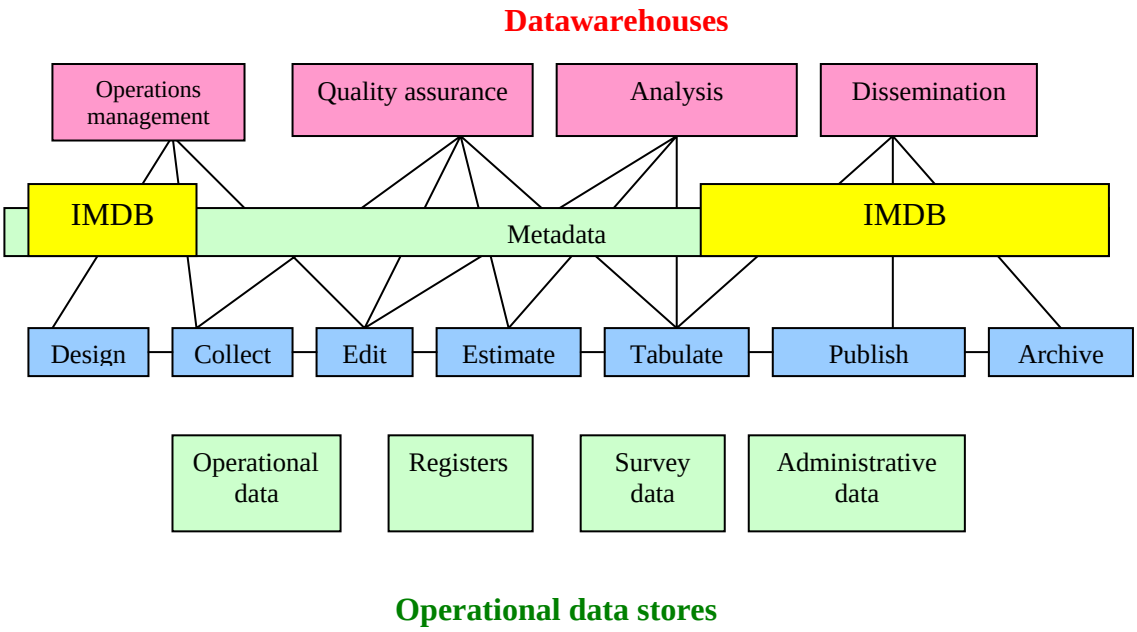
[1] Prepared by Alice Born (alice.born@statcan.ca) and Tim Dunstan (tim.dunstan@statcan.ca) of Statistics Canada.

II. SURVEY LIFE CYCLE AND THE IMDB

5. Figure 1 shows where the metadata in the IMDB 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 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 and 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 fully established.

6. The focus of this paper is on the role of the IMDB in meeting the metadata requirements for archiving structured data. Structured data are generally microdata datafiles that are outputs of survey processes or files from derived statistics such as national accounts or price indexes. Preparation for data archiving should begin early in the survey life cycle, and incorporate a schedule for deposition products and the creation and preservation of accurate metadata, ensuring the usability of the data itself at the end of the process.[2]

Figure 1. The role of the IMDB in the survey life cycle



III. PROPOSAL FOR THE ARCHIVE AND DISPOSAL PROCESSES AND SUB-PROCESSES

7. During the METIS workshop held in July 2007, it was proposed that a generic statistical business process model be adopted to provide a better basis for the generic Common Metadata Framework (CMF). As result of the meeting, it was agreed to add an "archive" phase in order to emphasize the importance of preservation of statistical data and related statistical metadata in many statistical agencies today. The archive phase has been expanded to include disposal in the statistical business process model since the business rules often include both retention and disposal of data and associated metadata.

[2] Jacobs, James A. and Charles Humphrey, 2004, *Preserving Research Data*. Communication of the ACM. 47, 9 (2004): 27-29.

8. The processes and sub-processes presented in Figure 2 are based on the generic statistical business process model presented by the UNECE and Statistics New Zealand (see Working Paper 17, METIS 2008). The three tiers developed for the archive and disposal phase follow the structure of the generic model: the first tier is simply the archive and disposal process, the second tier identifies the sub-processes within the phase, and the third tier identifies the sub-processes within the second level. In this case, some to the sub-processes identified in the dissemination phase of the generic model have been re-assigned to the archive and disposal phase; in particular, the sub-processes 7.1.4 Manage the destruction of data and associated metadata and 7.1.5 Preserve data and associate metadata found in Process 7. Disseminate.

9. In addition to the metadata required for describing the datafile and the metadata associated with survey methodology, data sources, data quality and definitional metadata, metadata required for maintaining and keeping track of the archived datafile has been added as sub-process 8.1. Sub-process 8.2 Preserve data and associated metadata relates to the retention of the archived datafile while sub-process 8.3 Dispose the data and associated metadata relates to the destruction of the archived datafile.

10. There are a number of data life cycle models that include archiving at the end of the life cycle or a part of repurposing the data. The actual process may not be as linear as diagrams suggest but it is important to plan to address the archival considerations since the data and metadata come from many parts of the data life cycle. During the data collection and datafile creation, it is important to follow best practices to facilitate archiving the structured data at the end of the data life cycle. For the data, the integrity of the dataset needs to be created at the beginning with references to variable names (and their definitions), labels, coding and missing data. For documentation, the use of metadata standards such as DDI, ISO 11179 and others are useful for ensuring that all the relevant metadata are there.

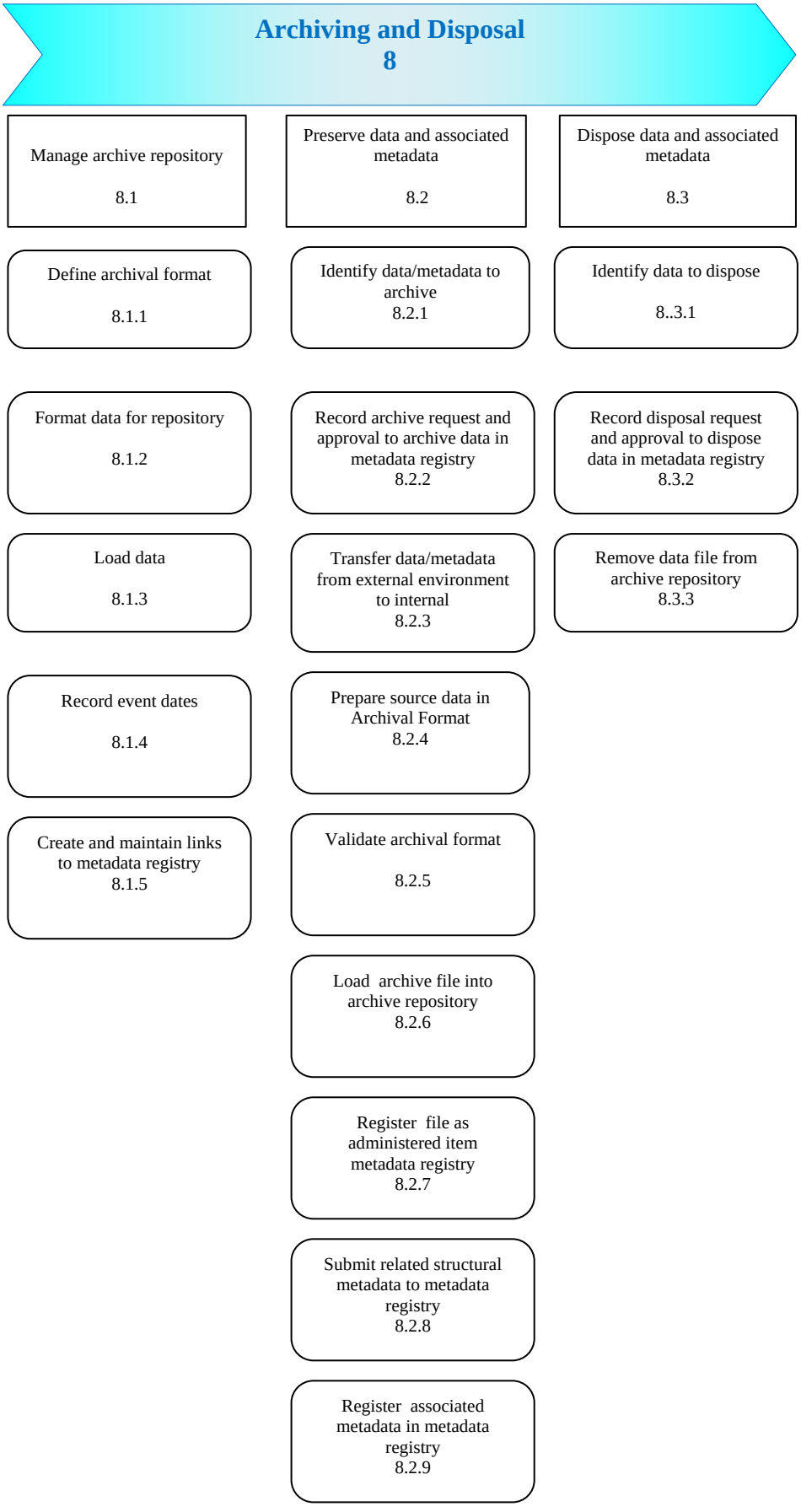
11. At Statistics Canada, we are beginning to formulate the sub-processes and the business rules around the retention and disposal of structured datafiles and their associated metadata. One option is to leverage the metadata already in the IMDB when implementing an archive solution, making the profiling of data less confusing to the end-user as well as being more closely aligned with the existing metadata produced during the life cycle of microdata, and supporting the eventual integration of different information systems. These metadata could be stored with the microdata files that are to be archived. We are looking at the possible use of the IMDB interface for storing the metadata necessary to support archived datafiles. Administered items in the IMDB data model such as information on the retention period and other metadata fields could be uploaded into the Archive repository. The metadata in the IMDB and archived datafiles could eventually be linked allowing employees to search by survey number, survey name, subject and other fields in order to obtain a list of relevant archived files, limited by privacy and security concerns. Therefore, an archive solution should build on existing metadata frameworks such as the IMDB as well as developing a set of metadata that relates to datafile itself.

12. Registration plays a major part in the support of archiving and disposal of data and metadata in the IMDB. All datafiles, archived or not, have a registration status. All metadata linked to this file also have a Registration status. Presently the registration status includes two values for metadata and datafiles, which can be used once the datafile or metadata have been withdrawn from common use. These values are superseded and retired.[3] We are proposing to add two more codes to the registration status code list called "archived" and "disposed" to support the archiving and disposal of datafiles and associated metadata.

13. For sub-processes 8.2.1 and 8.3.1 in the process diagram, the datafile and associated metadata would be withdrawn from common use and assigned a "retired" registration status. Subsequently, once the datafile is loaded in the archive repository, the registration status would change to "archived". The metadata follows the datafile into "retired" and then "archived" status. A datafile and its metadata can be tracked over time using version control and registration in the IMDB. For the sub-process 8.3, the registration status of "disposed" can be attributed to the datafile and structural metadata once removed from the Archive repository.

[3] For more information on registration status in the IMDB, refer to Born, Alice, 2007: Integrated Metadatabase (IMDB) – A Metadata Repository to Support the Survey Life Cycle, UNECE Workshop on the Common Metadata Framework (METIS), Vienna, July 4-6, 2007.

Figure 2. Archiving and disposal processes and subprocesses



IV. METADATA REQUIREMENTS FOR ARCHIVING STRUCTURED DATA

14. At Statistics Canada, we are considering storing the clean master datafiles and other important file types (Appendix I) that are to be archived for extended periods or permanently with all metadata and documentation necessary for their interpretation and use (e.g., record layouts, data dictionaries, software, program files etc.). Additional metadata specifically related to the archiving of structured data may include workflows and procedures, classification scheme for organizing the files, retention schedule, new data format and media extinction procedures and other information management business rules. One option is to store the metadata and documentation in a metadata registry, such as the IMDB, and the microdata files in an archive repository. A cross-reference field is created allowing the file submitter to associate supporting metadata with a microdata file, and store them together in the archive repository.

15. The types of metadata for archiving structured data include: 1. metadata needed for maintaining the archived file and keeping track of the files (i.e., administrative metadata); 2. metadata used to describe a file and its contents (i.e., structural metadata); and 3. information related to the file that is already available in the IMDB such as the data source of the datafile, description of survey methodology, measures of data quality, variables including their definitions and related classifications (i.e., survey and definitional metadata).

16. Below are some examples of the administrative metadata elements required for archiving structured data, based on the Statistics Canada experience:

(a) Link to an IMDB record (through the Statistical Data Documentation System (SDDS) number). If applicable, version identifiers are required to allow every file to be uniquely identified.

(b) Responsible manager for the datafile to be archived. Different roles are played by different persons at different times. It is important to identify who in each role.
- Creator – This is the division or work unit responsible for the creation of the file. This could be a support division, such as methodology or systems, creating a file on behalf of a client subject-matter division.
- Owner – This is the division for whom the file was created or which provided the specifications for its creation. The owner is solely responsible for authorizing access to the file.
- Custodian – This is the division responsible for the security and safekeeping of the file, but not for its contents. This would usually be the division with responsibility for the location where the file is stored, for example, a divisional share drive.
- Other governance roles – If there are any other roles related to the creation, storage and destruction of survey microdata files, they must be explicitly described.

(c) Retention (or destruction) period for the datafile - the retention period may be a maximum or minimum retention period with decision to destroy prior to maximum or later than minimum rests with the owner of the file. Exceptions to the retention requirements must be approved.

(d) Registration status

17. Table 1 presents a proposed list of administrative metadata elements to support archiving at Statistics Canada.

18. Currently the IMDB is mandated to contain information on the location, format and content of the confidential master datafiles for all of Statistics Canada surveys. The following are examples of the structural metadata elements that are currently in the IMDB:

(a) Title: refers to the name of the datafile. Each division has its own naming convention. In most cases it is the acronym of the survey with a “yy” variant for each reference year.

(b) File format is a basic requirement. Sufficient detail is provided to allow the datafile to be read including the record layout, variable names, variable labels and value labels (i.e., software syntax, such as SPSS and Stata).

(c) Software used for the datafile formats and includes the name and version number of the software.

(d) Storage media refers to the physical storage medium used to store the datafile including the storage level (i.e., primary and backup files), media type (e.g., hard disk), location (i.e., actual locations of the storage media – server) and the computer environment (i.e., a unique identifier for the computer on which the datafiles are stored – server and its operating system).

19. Finally, archived datafiles should be linked to or be stored with associated survey and definitional metadata. These metadata already exist in the IMDB and have been already presented in Part B and Part C of the Common Metadata Framework.[4] They also include Users' guides, data dictionaries, taxonomies and classification schemes. Table 2 provides examples of the metadata elements to support archiving as proposed in the Guide to Social Science Data Preparation and Archiving.[5]

Table 1. Examples of administrative metadata elements to support archived structured data.

Field Name	Syntax
Survey or Statistical Program Name	Drop-down list of valid names (from IMDB)
SDDS Number	Drop-down list of valid surveys (from IMDB)
Survey or Statistical Program Description	Populated from the IMDB
Survey or Program Classification	Populated from the IMDB
Survey or Program Reference Period	Validated date range syntax
Survey or Program Stage	Drop-down list of valid stages
Timestamp	System field
File Comment	Free-form text
File Name	Auto-populates when file is uploaded
ItemType	Drop-down list of accepted types, including the adopted list of microdata file types
File Extension	Automatically abstracted when file is uploaded
File Creation Date	Entered by the File Submitter (the date when the file was considered final, as opposed to the date when it is actually being declared)
Office of Primary Interest	As indicated in the Divisional PAAs (from IMDB)
File Creator	Identifies the program area staff member most involved in the creation of the file
File Submitter	Identifies the authorized submitter from the program area
File Approver	Identifies the Branch-level IM contact
Cross reference	Allows the user to identify, with links to files displayed in a directory or unique ids, all supporting or related files needed for the interpetation of the primary Item.
Financial Responsibility Code (FRC)	FRC of the OPI is automatically applied but may be over-ridden by the Branch-level OPI.
Program Element (PE) Code	From the IMDB

[4] Born, Alice, 2007: Integrated Metadatabase (IMDB) – A Metadata Repository to Support the Survey Life Cycle, UNECE Workshop on the Common Metadata Framework (METIS), Vienna, July 4-6, 2007; and Johanis, Paul and Daniel W. Gillman. 2006. Metadata Standards and Their Support of Data Management Needs. Joint UNECE/Eurostat/OECD Work Session on Stastistical Metadata (METIS), Geneva, April 3-7, 2006.

[5] Inter-university Consortium for Political and Social Research. 2005. *Guide to Social Science Data and Preparation and Archiving – Best Practice throughout the Data Life Cycle*. Ann Arbor, Michigan: University of Michigan.

Table 2. Metadata elements for archived data

Type	Description	
Question	Exact wording of the question or exact meaning the datum; links to questionnaire number	
Universe information	Who was actually asked the questions; skip patterns to determine which respondents were asked the question	
Summary statistics for the variable	Unweighted frequency distributions, missing data, etc.	
Missing data codes	Codes assigned to missing data	
Imputation and editing information		
Constructed variables	Derived variables from data collection; program statements used to construct variables	
Exact meaning of codes	Variables supported by classifications with definitions for each class to aid in interpretation	
Location in the datafile	For raw data, column location and record number (position number); order of variables in dataset	
Variable groupings	For large datasets, documentation should categorize variables into conceptual groupings.	
Data collection instruments	Copies of original data collection forms or instruments (e.g., CATI/CAPI programs);	
Flowchart of data collection instrument	For complex questionnaires showing which respondents were asked which questions and their linkages.	
Index of variables		
Interviewer guide		
Coding instrument	Rules and definitions used for coding data	
Sample and sampling procedures	Description of the population being measured and sampling methods; discussion on standard errors based on simple random sample are appropriate or if more complex methods are required; if weights are required, they should be described; indication of the response rate	
Weighting	Information on weights and how they should be used	
Units of analysis/ observations	Unit of analysis	
Variables	For each variable, the following information should be provided: • Exact wording of the question or exact meaning of the datum • Universe information (i.e., who was actually asked the question • Unweighted frequency distributions or summary statistics for each item • Missing data codes • Imputation and editing information • Details on derived variables • Codes, value meanings and definitions • Variable groupings (i.e., conceptual groupings)	
Technical information on files	Information on file formats, file linking	
Data collection instruments	Copies of original data collection forms and instruments (e.g., CATI/CAPI)	
Flowchart of data collection instrument	For complex questionnaires.	
Index	For large datasets, list of variables	
Interviewer guide		
Coding instrument	Rules and definitions used for coding the data.	

V. CONCLUSION

20. Statistics Canada is at the early stages of defining its metadata requirements and business rules for archiving and disposing its structured data. Early findings indicate that there is a need to create metadata that are at the datafile level in order to render these data independently understandable at some future date. This means expanding our current metadata framework to include administrative and structural metadata necessary to achieve this goal.

VI. REFERENCES

Berridge, Scott, C., John J. Bosley, and Daniel W. Gillman. 2006. Research-based Metadata Requirements for a BLS Reports Archive. UNECE Work Session on Statistical Dissemination and Communication. Washington, D.C., September 12-14, 2006.

Born, Alice, 2007: Integrated Metadatabase (IMDB) – A Metadata Repository to Support the Survey Life Cycle, UNECE Workshop on the Common Metadata Framework (METIS), Vienna, July 4-6, 2007.

Data Documentation Initiative Structural Reform Group. 2004. *DDI Version 3.0 Conceptual Model*. accessed on internet

Inter-university Consortium for Political and Social Research. 2005. *Guide to Social Science Data and Preparation and Archiving – Best Practice throughout the Data Life Cycle*. Ann Arbor, Michigan: University of Michigan.

Jacobs, James A. and Charles Humphrey, 2004, *Preserving Research Data*. Communication of the ACM. 47, 9 (2004): 27-29.

Johanis, Paul and Daniel W. Gillman. 2006. Metadata Standards and Their Support of Data Management Needs. Joint UNECE/Eurostat/OECD Work Session on Stastistical Metadata (METIS), Geneva, April 3-7, 2006.

Lavoie, Brian. 2004. *The Open Archival Information System Reference Model: Introductory Guide*. Dublin, Ohio: Online Computer Library Center Inc. and Digital Preservation Coalition.

Statistics Canada. 2008. *Archiving Business Rules – A Summary of Environmental Scan, Analysis and Associate Recommendation*. Report prepared by Systemscope Information Management and Transformation Partnerships. Ottawa: Statistics Canada.

Annex I. Types of microdata files

Microdata Files	In most cases, this refers to electronic files. In all cases this refers to information collected and retained under the Statistics Act at the microdata level (i.e., at the level of an individual person, business or organization). Most of these files are confidential. Business rules are currently being developed for determining which datafiles should be preserved (archived) or disposed.
Disseminated Files	Online statistical databases (e.g., CANSIM), datasets for publications and papers. These files are not confidential and are available in the public domain. These are not considered for archiving and disposal.
Microdata File Types	
Census of Population files	Required for transfer to Library and Archives Canada after 92 years.
Current internal master (analytical) files	An internal master file is a final survey file after completion of all data processing from which survey outputs are produced. Any analysis could be conducted from this file. It contains all survey analytical variables. Current internal master files may be subject to revision.
Historical internal master (analytical) files	Historical internal master files are generally not subject to revision. Should be considered for the archived repository.
Analytical work files	Files that support the analytical activities for a survey. They may be subsets of internal master files or preliminary versions of master files.
Sample files	Microdata files that identify the units selected for the survey (for example, addresses, names, telephone numbers). Usually, the file is created by the subject-matter division or the methodologists working on the survey and then passed to a division responsible for data collection.
Completed paper questionnaires	
Collection data files	(Interviewers / Regional Offices, Collecting division in HO, Subject matter division): Files that represent the information as collected. They include data-capture files from paper questionnaires, Blaise data files, and collection files transferred from one format to another.
Processing files	Files where the data have been modified in any way from that which was collected.
BTH (Blaise Transaction History) files	Files produced by Blaise giving information on the collection operations for a survey. To emphasize, these are always microdata files.
Public use microdata files	Files approved for public release by the Microdata Release Committee, as not containing information that can identify an individual person, business or organization.
Data share files	Files produced as a result of a signed discretionary disclosure order. Files approved for release to specific organizations that meet the legal requirements of the Statistics Act
Research data centre files	Analysis files used for approved research projects, usually in Research Data Centres but occasionally in subject-matter offices in Ottawa. These are similar to internal master files, but do not contain direct identifiers or other information that serve no direct analytical purpose.
Paradata files	Paradata files contain information related to a statistical data collection or production process that is linked to an identifiable person, business or organization. It is distinct from the information that is the objective of the statistical data collection or production process.
Temporary or work files	These are created by a single person for that person's sole use. These are ad hoc files created for very specific and short-term purposes.
Other microdata files	Anything not falling into one of the above categories. The number of files in this category should be very low. If not, then additional categories should be created.