

Work Session on Statistical Metadata
(Geneva, Switzerland, 18-20 February 1998)

Item 5 of the provisional agenda

**THE EVOLUTION OF METADATA AT STATISTICS CANADA:
AN INTEGRATIVE APPROACH**

Submitted by Statistics Canada ¹

¹ Prepared by E. Boyko.

I. INTRODUCTION

1. The rapid growth of computing and telecommunications technology is both fuelling the demand for electronic information and making it easier to publish information. Statistical agencies are responding by developing websites and databases for access via both the Internet and CD-ROM.

Electronic publishing brings new opportunities and challenges to both the producers and users of statistical information. One such challenge is the ability to find specific material on increasingly complex websites. A response to this challenge has been the construction of better tools for finding and using information. Such tools come under the general heading of meta-information or metadata.

2. Twenty-five or more years ago, metadata was a term used by computer system specialists in their work on statistical computer systems. Today the term is broadly used in a variety of circumstances where information about data and information is being discussed. The two terms (metadata and meta-information) are sometimes used interchangeably but metadata can also be used to refer to machine level information. Current usage seems to be tilted in favour of metadata as a general term. This paper will use the term metadata throughout.

3. The use of metadata within statistical agencies is only a small sampling of the metadata explosion that we are witnessing thanks to the Internet. In its early days, the Internet could be likened to having a truckload of books dumped into the main hall of a library. Yes, the information may have been there but it was very difficult to find because there were no indexes or catalogues. More recently, the situation has changed with the introduction of search engines that are hungry for metadata. As a result, what we used to know as indexes, catalogues and resource frameworks are now discussed as metadata. While the growth in the use of the term 'metadata' may in some sense be overstated, the problem that this sort of information attempts to overcome is by no means trivial.

4. The purpose of this paper is to highlight some of the work being undertaken at Statistics Canada in the area of metadata; in particular, a conceptual model that focuses on the objectives of metadata will be developed and discussed as a way of focussing work on metadata. As well, an innovative way of collecting metadata will be described. Since there have been various metadata projects undertaken at Statistics Canada over time, there will be an emphasis placed on approaches to integrating the collection and dissemination of metadata.

II. METADATA AT STATISTICS CANADA

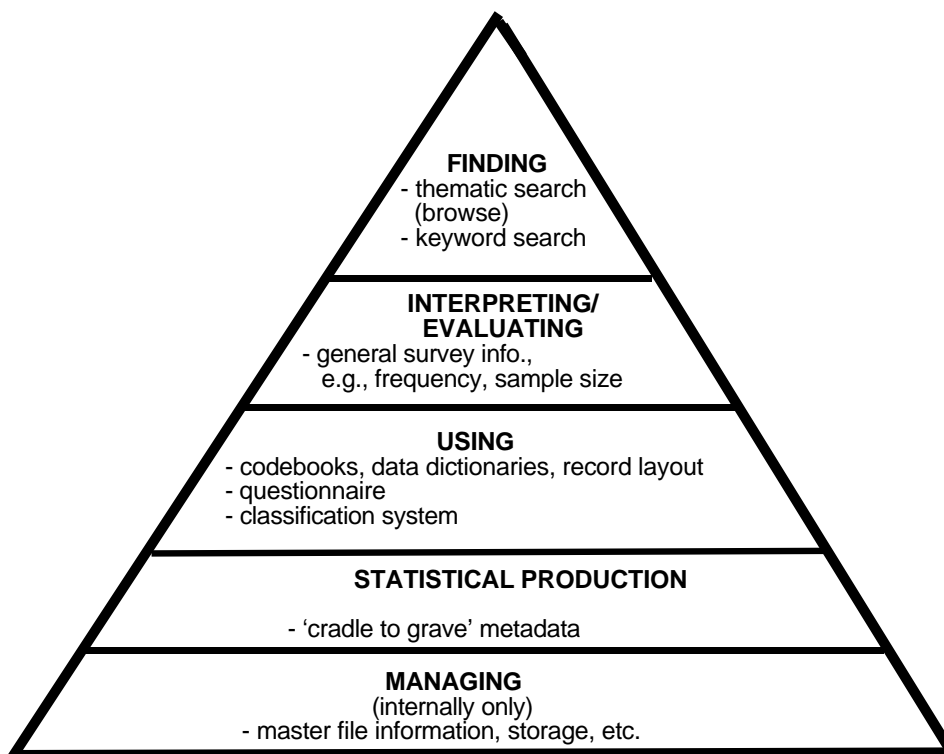
5. Metadata initiatives are by no means a new activity within the Agency. A system for documenting surveys (Statistical Data Documentation System-SDDS) has been in existence for nearly 20 years. A directory system to support the online dissemination of time series data via CANSIM (Statistics Canada's time series database) has been routinely produced for nearly 30 years. The dissemination of public use microdata files (PUMFs) has depended on file documentation of various types and more recently, the Thematic Search Tool (TST) was set up as a means to browse this type of documentation. The Statistics Canada Catalogue has been available electronically on the Statistics Canada website during the last few years. The

development of new computer systems such as the one that was developed for the Canadian census is metadata driven. In short, there is no shortage of metadata projects at Statistics Canada and there will undoubtedly be others as computer systems are redeveloped using new technology.

6. In April 1997, Statistics Canada's management approved a new metadata project to provide focus to this area and to emphasize its dissemination to the user community. This project draws on the work done by others in the Agency, in particular on the work done on the Thematic Search Tool and the SDDS. The TST was initially created as a tool to identify opportunities for harmonizing statistical concepts within Statistics Canada's social statistics program. It was made available to the public via the Statistics Canada website and has proved popular with researchers. One of the objectives of the new metadata project is to extend the concept of the TST more broadly within the Agency, to integrate it with other metadata initiatives and to make it available to the user community via the Statistics Canada website.

III. THE OBJECTIVES OF METADATA: A CONCEPTUAL MODEL

7. While statistical agencies are not strangers to metadata, the role of metadata appears to be shifting from being a means to document data and statistical activities to being a tool to support the dissemination, use and management of statistical information. After reviewing the existing work of Statistics Canada with respect to metadata, the project team has developed a simple model to describe its work objectives. The model is depicted below using a hierarchical presentation of metadata based on its role from a user point of view. Figure 1 below summarizes these functions.



8. The apex of the pyramid represents information that helps a user find an information resource. Resource discovery is usually the first step that users take while attempting to find information relevant to their needs. Does the information exist? If yes, where is it?
9. Assuming that the user finds information sources that appear to be relevant to his/her needs, the next question is one of assessing their relevance and adequacy. This function can encompass a variety of fields of information depending on the information resource. Some commonly included information would be abstracts, definitions, concepts, measures of quality and methodology.
10. The third level concerns the access and use of information. Once again this can take on various forms. It can include the labelling information attached to data elements in a database, information on the formats of files, classification information or codes, and the codebooks supporting the dissemination of public use files.
11. The fourth level of metadata can be used in statistical production. This is typically found in integrated survey systems such as those that support CATI/CAPI applications. The emerging systems in this area can provide 'cradle to grave' metadata support for statistical activities.
12. The base of the pyramid denotes metadata required for managing a statistical program. This type of metadata has been largely used internally in statistical agencies for the management of concepts, classifications and standards. It can also involve the information required for the long term preservation of information as required by most national archives.
13. It should by no means be suggested that each function or type of use identified above requires unique metadata. There is a considerable overlap with respect to the metadata content which can be viewed in different ways. Also, the metadata from level four can be used to generate most of the other metadata that are required for other purposes. Statistics Canada is presently assessing user priorities and will be developing a user guide to help users find appropriate metadata.
14. The next portion of this paper will focus on the issues and considerations that the project team at Statistics Canada is attempting to deal with in the development of metadata for the first three layers of metadata as depicted by the metadata model presented in Figure 1 above. The paper will conclude on our current thinking as to how a system for managing and collecting such information should be focussed.

IV. ISSUES AND CONSIDERATIONS IN DESIGNING A METADATA SYSTEM AT STATISTICS CANADA

4.1 Project Orientation

15. One of the major objectives of the Statistics Canada metadata project is to develop an appropriate metadata repository that will be kept up to date. In a paper publishing world, this information was updated publication by publication and there was little need to develop a corporate repository. Publishing via a website or other integrated approaches changes the way in which this work will be done inside a statistical agency. One of the objectives of the Statistics Canada project is to keep the amount of information collected to a minimum (so as not to create

an unrealistic workload for author areas) and to make the provision of this information by internal respondents as easy as possible. It is felt that if the metadata content is too elaborate and detailed, it will be very difficult to keep current but, if it is not complete it will fail to serve its purpose. The key is to find the appropriate balance. The guidelines provided to the metadata project by the Chief Statistician of Canada are quoted below.

"Two fundamental priorities are driving the structure of the meta-database: it must be comprehensive in coverage; and it must be driven by what clients are likely to want rather than by what we think they should know. . . .

"If we focus on the ultimate meta-database, the level of detail required would impose an enormous reporting burden on the subject-matter divisions that must provide and update the information. Frankly, the database would collapse under its own weight." (Fellegi, 1997).

16. The other challenge being faced by the Statistics Canada project involves the integration of the various existing metadata collection activities. There are currently five different metadata collection/production activities for the following metadata bases: the electronic catalogue (the IPS); the Statistics Canada library online public access catalogue (OPAC); the Statistical Data Documentation System (SDDS); the metadata base for managing (archiving) the statistical program (MIDAS); and the Thematic Search Tool (TST).

17. While it is not clear whether total integration is feasible or desirable, it is evident from the outset that some integration will be necessary in order to produce a viable system that will meet the needs of users.

4.2 Finding Statistical Information

4.2.1 Issues

18. Users start with problems or reasons for which they may choose to look for information. Going to a statistical agency may or may not be their first instinct. Very few problems come labelled as to which type of information or data will shed light on them. As often as not, they may go to their library to look for this information or, in today's situation, may use the Internet. It should also be mentioned at this point that the term library also includes data library and data archive, each of which have staff that are extremely knowledgeable about data from statistical agencies and other sources.

19. The first action of a librarian or a data specialist will be to see if the required information is available locally. This means that he/she will search a local catalogue. These are often referred to as OPACs or online public access catalogues. In today's world, OPACs may refer to Uniform Resource Locators (URLs), and websites as well as local electronic and hard copy information resources.

20. A typical inquiry will start at a broad level, perhaps with a review of existing studies and published research. The next step may be to try to find information that is specific to the particular geographic area under consideration. Again, the researcher will want to know whether there are existing studies or data compilations.

21. The tools being used at this stage of the study would likely be library catalogues and bibliographic search tools. These are likely to be accessed at research libraries and at larger public libraries. They are also available at Statistics Canada's library in Ottawa, but not in Regional Reference Centres. Increasingly, library OPACs are available for remote access and searching. However, if Statistics Canada's holdings are not catalogued by the library involved (until recently, this was often the case) the researcher has little chance of finding any reference to relevant information product from the Agency in the OPAC or catalogue. In such a case, his/her best ally would be a knowledgeable 'government documents' librarian who could refer the person to the area housing the Statistics Canada collection and to the Agency's catalogue either in the paper version available locally or on the Statistics Canada website.

4.2.2 Response

22. To help users find information available in existing products and services, Statistics Canada has developed a tool called the Information on Products and Services (IPS). It is essentially an electronic catalogue. An example of an IPS record is shown in Appendix 1. The IPS can be accessed on the Statistics Canada website (www.statcan.ca) by clicking on Products and Services on the Statistics Canada home page and choosing catalogue. The search engine underlying this file is Open Text. Fielded searching capabilities will be added in 1998.

23. Lets assume that the researcher has reached the statistical agency website. How do they want to search the website? A number of possibilities present themselves. The home page should give broad directions as to where to start. Search engines and searching for specific words (subject, title, author, keywords in full text) is common-place. A major challenge that most users will have in these situations are choosing the right words. Thesauri and glossaries may help in this regard but these take effort to develop and integrate into search tools. Other alternatives are to construct subject indexes showing key subjects and related areas and to index the information by geographic areas. The difficulty here is the effort required to construct the links and in keeping them up to date.

24. General products, such as the Statistics Canada Daily, can also be a useful way of directing a user's search. For example, The Daily, summarizes all major releases and provides a broad summary of the major findings. Users who search The Daily can be directed to the bibliographic record for the product, information for ordering it and to a database if this exists.

25. Last but not least, it should be mentioned that the website itself may be a source of the required information. Appropriate web designs and search engines are necessary, if not sufficient tools for finding this information.

4.2.3 Future Considerations

26. From a user's perspective, the objective is to get the information resource as quickly and as easily as possible. One way of helping to achieve this is to make statistical metadata as broadly available as possible. Well-indexed websites with good search engines are essential. This work is now underway at Statistics Canada. Additionally, in an effort to assist users doing remote searches, the Statistics Canada Library OPAC will be accessible from the website. Important features are that it uses a Z39.50 (ANSI) search engine and a standard record (MARC-Machine-Readable Cataloguing). This provides a means for remote searching and for

exporting records to other library systems. Its main drawback at this stage is that it only contains a limited number of records for data files but this is being addressed in the context of Statistics Canada's Data Liberation Initiative (DLI).

27. GILS is another important metadata initiative that is becoming increasingly important for searching for government information. GILS stands for Global Information Locator System and started in the US as a government information locator system. It contains fewer fields than a MARC record but can be generated from a MARC record. The IPS can also generate a GILS record.

28. A final global initiative, that is aimed at helping users find information, is often referred to as the Dublin Core Project. The project aims to identify a common set (currently about 15) of fields which can be used to tag information entities for the humanities and the social sciences generally.

29. Since researchers often do literature searches when they are doing their research, it is important that the use of statistical agency information is properly cited in published research. This is well understood and carried out in a paper environment but is more difficult in an electronic world. Take for example a researcher using a public use microdata file. Full citation would involve identifying the data set, the variables and the software and the algorithm used to manipulate them. Statistical agencies can improve the chances of having future researchers find their statistical material by suggesting citation standards appropriate to their databases and data sets.

4.3 Evaluating Information

4.3.1 Issues

30. Let's assume that our researcher has found a number of Statistics Canada references and must now evaluate their relevance. We are now talking about the second level metadata depicted in the model in Figure 1. Statistics Canada publications typically come with methodological notes. In an electronic world, users expect to see everything on the website.

31. At this level, the user is interested in concepts, definitions, sources and methods. What universe is covered? What questions were the respondents asked? What definitions have been used? How have standard classifications been applied?

32. Links to this level of information can be added to the bibliographic record where there is a clear link between the output under assessment and various inputs such as surveys. Users may expect surveys for which there are public use files to have the entire codebook up on the website, but this approach may not suit other types of data. Outputs which are summaries of surveys lend themselves to this linkage quite easily, but it is a whole other challenge for documenting the national accounts. This will provide a user with all the necessary information in evaluating the information resource.

4.3.2 Response

33. As was mentioned above, Statistics Canada has been systematically collecting basic information about its survey program via the SDDS project for nearly 20 years. More recently, the Thematic Search Tool project has expanded the amount of information available via a pilot project. In particular, it added variable level information for the public use microdata files. For the next phase of metadata development at Statistics Canada, variable level information will be added for as many surveys as possible. Where applicable, the codebooks and data dictionaries will be part of the documentation.

34. Documenting surveys at the variable level can have important benefits for finding information. The discussion of searching above assumes that the user can be led to the information because the subject sought is mentioned in the abstract or is somehow part of the standard record. If it is not found, does it mean that the information is not available? Not necessarily. It may be part of an internal database or a statistical master file, i.e. the question was asked on a survey but was not tabulated as part of a standard output. It may also be part of a public use file but was not indexed in the bibliographic record.

35. Linkages from bibliographic records to survey level information can enrich the search results and give the user more options. If the information cannot be accessed from a standard product, then it may be possible to ask for a custom tabulation. For example, the custom tabulation program for the Canadian census is large and goes on for a number of years after the census results are released.

36. The main challenge for Statistics Canada in the upcoming year will be to integrate the various sources of metadata into an integrated record with a common user interface and with linkages to other databases.

4.4 Accessing and Using Information

4.4.1 Issues

37. Many would consider data use to be beyond the scope of a discussion of metadata. Metadata have traditionally been considered to be of relevance to the humans that use data. But it is also useful to think of machine-level metadata. This concept goes beyond machine-readable metadata to metadata that are used in computer processing. In the case of CANSIM data, this would include the formats, series numbers and titles that allow the user to import and keep track of the series.

38. Having found a data source and decided that it is relevant, users want access. If the user is in a library, he/she hopes that the information resource is available locally or can be obtained quickly. In a 'point and click world', users expect to be able to access information directly from websites and databases. Ideally it should be in a format that allows it to be used in a variety of analytical softwares. One way in which this can be done is to have linkages from the metadata used to find and evaluate data to the actual data.

39. The users of public use microdata files absolutely depend on the codebooks that accompany these files. One difficulty that they often encounter is an underdeveloped standard for

the production of these codebooks and a lack of adherence to the standards that do exist. In addition, there is often a lack of appropriate formats to enable the data to be used directly without considerable effort on the part of the user.

4.4.2 Response

40. The response to the issues of access and use take on a number of characteristics at Statistics Canada. For data that can be accessed directly via the website, links are (or will be) made from the metadata (The Daily, the IPS) to the database, e.g., CANSIM.

41. For aggregate data not on the website, users are presented with the option of formatted data or data wrapped in a PC-based software. Initially, Statistics Canada used its own in-house, DOS-based software. For the past few years, the Agency has moved to a commercial product called Beyond 20/20. It is a Windows-based software developed by a Canadian firm called Ivation (www.ivation.com). Although it uses a proprietary format, the main advantage that such a software presents for Statistics Canada and for the user is that it ensures that all the relevant metadata required to access and use the data are part of the package. Currently, such data packages are offered on CD-ROM but in the future the same convenience will be offered in client server mode via the Internet.

42. The same software, Beyond 20/20, can also be used to disseminate microdata files. The experience with the DLI project has shown Statistics Canada that the software command files for the commonly used statistical packages used in academia (SPSS and SAS) and ASCII versions of the data should also be part of the packages for microdata files. The DLI project team creates such command files if necessary but an automated solution is expected for the future.

43. The Thematic Search Tool and its successor will contain all of the metadata at the level of the codebook for microdata and relevant classification standards for macrodata. The major challenge in the dissemination of such information is its production. In the experience of Statistics Canada, the best documentation is produced by the CATI/CAPI survey systems where the metadata are an integral part of the production process. These are the 'cradle to grave' systems that are mentioned above and correspond to level 4 of the conceptual model presented in figure 1. It has also been the experience of the Agency that well documented files require less user support than those with poorer documentation.

4.4.3 Future Considerations

44. The future design of database systems (such as the new CANSIM base) will take metadata requirements of users into account as part of the design. They will be fed by the corporate metadata repository wherever possible and the repository will also hold the metadata generated specifically for the CANSIM II. There are indications that an electronic glossary would be a welcome addition on the Statistics Canada website to support the dissemination process.

45. Statistics Canada is currently working with an international group focussing on the data documentation. The Data Documentation Initiative (DDI) was formed in 1995 at the annual conference of the International Association for Social Science Information, Service and Technology (IASSIST) under the auspices of that group and the Inter-University Consortium for Political and Social Research (ICPSR) based in Ann Arbor, Michigan. Current membership is

drawn from the U.S., U.K., Denmark, and Canada. The DDI is currently focussing on standards for codebooks for microdata files and more recently has started to work on aggregate data. They are hoping to facilitate the production of standard documentation through the use of an SGML coding structure. This should lay the basis for the automatic generation of software command files and more efficient remote searching.

V. METADATA COLLECTION AND CENTRAL MANAGEMENT

46. It is a premise of the metadata project that an efficient and effective collection tool will encourage author areas to respond to calls for supplying metadata. To meet this challenge, an electronic template with response categories for all fields of desired information has been developed and deployed to each author area. The template is basically an electronic form into which the authors can cut and paste their information from their various production systems. The form is linked directly to a database so that the information entered in the form is stored directly without manipulation or re-entry. This process was successfully tested in several divisions before full implementation.

47. The database into which the information is entered (Paradox) is SQL compliant so that the information can easily be moved to other softwares if necessary. Minimal manipulation will be performed on the information in the database (spell-check, English/French consistency/comparison and checks against the Statistics Canada thesaurus). The accuracy of the content will be left to the author areas (at least initially).

48. It is possible that certain metadata may be aimed at internal audiences only and should not make up part of the public database. This will certainly be an issue for Statistics Canada. Accordingly, there will have to be two repositories; an internal one and an external one. Since Statistics Canada's security policy requires the Agency to have two networks, this should not be a problem.

VI. FUTURE DIRECTIONS

49. Metadata are an essential part of disseminating and managing statistical information both nationally and internationally. The widespread use of the Internet will increasingly bring data from various countries together in much the same way that it has within countries. Variations in metadata content, quality and presentation will be obvious to users. This in turn will raise questions as to standards. It has been suggested that important improvements could be derived from working together on metadata standards. Standards pertaining to minimum information and standardized fields would aid in the international exchange and integration of information.

50. The work of METIS is important in this regard. The GILS initiative is also an important step in this direction as is the DDI work being undertaken by IASSIST and ICPSR. These initiatives, joined with the work of METIS, should be supported by statistical offices to ensure more effective use and management of statistical information.

Information on Products and Services (IPS) Catalogue

Catalogue No. 15-203-XPB[Paper][[New search](#) | [Back to Statistics Canada's home page](#)]**Provincial gross domestic product by industry****Author(s):**

RICHARD MARTEL

Frequency: Annual**Medium:** Paper**First Issue:** 1971**Latest release:** May 16 1997 for the 1984-1996 edition**Available in:** Bilingual (English/French) edition**Status:** Ongoing**International Standard Serial Number:** ISSN 0712-8762**Abstract:**

The publication presents current price estimates of provincial gross domestic product at factor cost (GDP) for all major industries of the Canadian economy, including aggregates and special industry groupings. A brief text provides general highlights and charts for each province. 15-203-XPB continues 61-202.

Pricing:(Prices do not include sales tax)

	Issue
Canada:	\$52.00
Outside Canada:	US \$52.00

Last revised date:

1998-01-03