Metadata: an Integral Part of Statistics Canada's Data Quality Framework[1]

Marcelle Dion[2]
Director General Agriculture, Technology and Transportation Statistics Branch
Statistics Canada
13th Floor Section B-7, Jean Talon Building
170 Tunney's Pasture Driveway
Ottawa (Ontario), Canada
K1A 0T6
Email: marcelle.dion@statcan.ca

Abstract

Statistics Canada's product is information. The management of quality must therefore play a central role within the overall management of the Agency. The Quality Assurance Framework describes the approaches that Statistics Canada takes to the management of quality. This Framework is based on six indicators: relevance, accuracy, timeliness, accessibility, interpretability and coherence. Metadata are at the heart of the interpretability indicator by providing the information necessary to interpret and utilize the statistics appropriately.

The first part of the paper will describe the Quality Assurance Framework and will show how metadata relates to the framework.

The second part of the paper will be devoted to metadata. It will include a description of the IMDB, its governance model, the mechanisms put in place to assist managers in loading information and ensuring its coherence, the monitoring of its quality and users' access.

Another section will outline the minimum set of metadata required to comply with the Policy on Informing Users of Data Quality and Methodology. A final section will focus on the agriculture statistics program.

[1] The opinion expressed in this paper are those of the author and do not necessarily reflect the official position of Statistics Canada

[2] The author wishes to acknowledge the contributions of Alice Born, Denis Chartrand, Cindy Ingalls and Philip Smith in the preparation of this paper.

Introduction

Statistics Canada's product is information. The management of quality therefore plays a central role within the overall management of the Agency. There are three key elements[3] of managing quality: the Quality Assurance Framework, the Policy on Informing Users of Data Quality and Methodology and the Integrated Metadata Base (IMDB).

The Quality Assurance Framework defines what is meant by data quality. The Policy on Informing Users of Data Quality and Methodology requires that users be provided with the information necessary to judge the quality of the information they require and to judge its fitness for their intended use. The IMDB[4] is the primary vehicle to adhere to the policy and to provide metadata or "information about information".

The first part of the paper will describe the Quality Assurance Framework and will show how metadata relates to the framework. The second part will be devoted to metadata. It will include a description of the IMDB, its governance model, the mechanisms put in place to assist managers in loading information and ensuring its coherence, the monitoring of its quality and users' access.

Another section will outline the minimum set of metadata required to comply with the Policy on Informing Users of Data Quality and Methodology. The final section will focus on the agriculture statistics program.

I- Quality Assurance Framework

The Quality Assurance Framework describes the approaches that Statistics Canada takes to the management of quality for all its programs including those related to agriculture. Statistics Canada has defined data quality in terms of "fitness for use". Quality is important, but it is a matter of degree. One needs very high standards of accuracy, timeliness, etc. for some statistical applications, but one can 'make do' with much less accuracy, timeliness, etc. for some other purposes. This is what the "fitness for use" concept is all about. Six dimensions of quality have been identified within the concept of "fitness for use".

1. The ***relevance*** of statistical information reflects the degree to which it meets the real needs of users. It is concerned with whether the available information sheds light on the issues of most importance to users.
2. The ***accuracy*** of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure. It is usually characterized in terms of error in statistical estimates and is traditionally decomposed into bias (systematic error) and variance (random error) components. It may also be described in terms of the major sources of error that potentially cause inaccuracy (e.g., coverage, sampling, nonresponse, response, etc.).
3. The ***timeliness*** of statistical information refers to the delay between the reference point (or the end of the reference period) to which the information pertains, and the date on which the information becomes available. It is typically involved in a trade-off against *accuracy*. The *timeliness* of information will influence its *relevance*.
4. The ***accessibility*** of statistical information refers to the ease with which it can be obtained by users. This includes the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which the information can be accessed. The cost of the information may also be an aspect of *accessibility* for some users.

[3] Statistics Canada's Intranet Site http://stdweb/standards/IMDB/IMDB-nutshell.htm
[4] Johanis, Paul, *Role of the Integrated Metadata Base at Statistics Canada*, Statistics Canada International Symposium Series – Proceeding, 2001

5. The ***interpretability*** of statistical information reflects the availability of the supplementary information and metadata necessary to interpret and utilize it appropriately. This information normally covers the underlying concepts, variables and classifications used, the methodology of collection, and indicators of the accuracy of the statistical information.
6. The ***coherence*** of statistical information reflects the degree to which it can be successfully brought together with other statistical information within a broad analytic framework and over time. The use of standard concepts, classifications and target populations promotes coherence, as does the use of common methodology across surveys. *Coherence* does not necessarily imply full numerical consistency.

Metadata are at the heart of the management of one of these indicators, the interpretability indicator, by informing users of the features that affect the quality of all data published by Statistics Canada. The information provides a better understanding of the strengths and limitations of data, and how they can be effectively used and analyzed. Metadata may be of particular importance when making comparisons with data across surveys or sources of information, and in drawing conclusions regarding change over time, differences between geographic areas and differences among sub-groups of the target populations of surveys[5].

II- Integrated Metadata Base

As mentioned above, Statistics Canada's responsibility in managing interpretability is primarily concerned with the provision of metadata or "information about information". It is important for statistical agencies to publish good metadata because by doing so they show openness and transparency, thereby increasing the confidence of users in the information they produce. If the methods and quality measures underpinning the statistics are impenetrable or difficult to access, it is difficult to have that trust. But if a statistical agency is completely open about how it produces its products and what their quality characteristics are, then users can have great confidence in the integrity of those statistical products. Openness and transparency about the information -- about its weaknesses just as much as about its strengths -- breeds users' trust in a statistical agency's products.

Metadata for Statistics Canada's active and inactive surveys (415 active surveys and 400 inactive surveys) are stored on the corporate repository of information, the Integrated Metadata Base (IMDB). It provides an effective vehicle for communicating metadata to data users. Its coverage of Statistics Canada's data holdings is exhaustive, the provided information on data quality complies with the Policy on Informing Users of Data Quality and Methodology, and it is presented in a consistent and systematic fashion[6].

For the purposes of the IMDB, the term "survey" refers to the collection, analysis and reporting of data concerning characteristics of a population. These data may be collected directly from survey respondents, derived from other Statistics Canada surveys and/or collected from administrative files. In the case of the agriculture program, these include data from the crop, livestock and financial surveys, data from marketing boards and other administrative sources, data compiled from tax records as well as from the Census of Agriculture.

The type of information provided covers the data sources and methods used to produce the data published from surveys and statistical programs, indicators of the quality of the data as well as the

[5] Statistics Canada's Intranet Site http://dissemination.statcan.ca/english/concepts/background.htm
[6] Johanis, Paul, *Role of the Integrated Metadata Base at Statistics Canada*, Statistics Canada International Symposium Series – Proceeding, 2001

names and definitions of the variables, and their related classifications. The IMDB also provides direct access to questionnaires.

The IMDB has been built to facilitate the maintenance of historical statistical metadata as well as providing a snapshot of the metadata for any survey instance as far back as November 2000 – the starting point of the IMDB.

The metadata supports all of the Agency's dissemination activities including its online data tables, *CANSIM* and *Summary Tables*, publications, analytical studies and *The Daily* (Statistics Canada's official release bulletin).

The metadata also supports data collection activities. The IMDB is the source for the survey information displayed on the *Information for Survey Participants* module on the Statistics Canada's website.

III- Governance Model

A metadata program will be successful if program managers comply and keep the metadata up-to-date. To ensure compliance a governance model, technical assistance and monitoring tools have to be in place.

i) The policy

Statistics Canada's Policy on Informing Users of Data Quality and Methodology ensures compliance as it requires that all statistical products include or refer to documentation on data quality and methodology. The underlying principle behind the policy is that data users first must be able to verify that the conceptual framework and definitions that would satisfy their particular data needs are the same as, or sufficiently close to those employed in collecting and processing the data. Users also need to be able to assess the degree to which the accuracy of the data and other quality factors are consistent with their intended use or interpretation.

The policy defines standards and guidelines that describe the kind of documentation that is expected. The **Standards** detail the mandatory requirements for documentation on data quality and methodology for all products under this policy. For certain programs and their products such as price indices, the system of national accounts, a broader and more detailed range of methodology and data quality documentation is desirable. The **Guidelines** outline the types of information to be included in such additional documentation.

The policy also identifies who in the organization is responsible for the implementation of the policy:

- Directors of program areas (for example agriculture, transportation, etc.) are accountable for providing quality survey metadata to the IMDB administration.
- The Methods and Standards Committee, a management committee composed of senior managers from across Statistics Canada, is responsible among other things for defining the standards and guidelines and producing reports on the state of compliance with the policy.

ii) Technical assistance

Managers' initial time investment in developing metadata information for each of their survey program is not insignificant and can result in a delay in the provision of the information and/or in the provision of information of a lesser quality if not managed properly.

In Statistics Canada, Standards Division has been assigned the responsibility and the resources for the management of the Integrated Metadata Database. Initially, Standards Division put together a team of five metadata officers who worked closely with survey managers to help them with the development and update of metadata. Over time, the team developed training materials and tools such as the *Guidelines for Authors* to facilitate managers' compliance with the policy requirements[7]. These guidelines can be accessed online and provide among other information, a template for completing or updating the IMDB and the IMDB Definitions and Business Rules[8].

In addition to the help provided by Standards Division's metadata officers, areas of Statistics Canada such as the Business and Trade Statistics Field (BTS) put in place a support mechanism to assist managers in developing the initial set of metadata to ensure compliance with the standards and the quality of the information stored on the IMDB. In the case of BTS, the support mechanism consisted of one person coordinating the work of all managers, providing them with information on the process and templates to facilitate the development and, reviewing the final product for completeness, readability and consitency. The end result was a set of good quality metadata information for business surveys being loaded on the IMDB.

iii) Monitoring[9]

Statistics Canada's IMDB includes information for surveys and statistical programs for which data have been collected/disseminated since November 2000 — i.e. the date when the IMDB became operational. While during the IMDB initial development phase, the trigger for the creation of a new IMDB record was a data release in *The Daily,* the creation of a new record and its documentation is now triggered when a new survey is authorized by the Chief Statistician.

Statistics Canada has a quality assurance process in place for monitoring the quality of IMDB records. The first stream of the quality assurance is to ensure the quality of the metadata as it is being created and stored in the IMDB by providing managers with a toolkit containing detailed guidelines and a template for completing the metadata for each survey record. The second stream is referred to as the "measurement process" i.e. how well the documentation for each of the surveys and statistical programs meets the spirit of the Policy on Informing Users of Data Quality and Methodology. The "measurement process" assesses the degree to which each record meets the conditions of the policy, as outlined in the IMDB Template. Records are rated on a four-point scale ranging from "fully compliant" to "has major deficiencies"[10]. In between exhaustive review of the IMDB information, metadata officers review the ratings[11] as and when updates are made to IMDB records by program areas[12]. These measures allow the identification of targets for corrective action in areas requiring more support from the IMDB team. As the result of this initiative, results have improved from 34% of the records being compliant in March 2003 to 93% of the records being compliant by December 2006[13].

---

[7] Statistics Canada's Intranet Site, http://stdsweb/standards

[8] Statistics Canada's Intranet Site, http://stdsweb/standards/imdb/template-update.doc

[9] Born, Alice, *Metadata as a Tool for Enhancing Data Quality in Statistical Agency*, European Conference on Quality and Methodology in Official Statistics, 2004

[10] More information on the rating process is available in the Appendix A of the report *Quality Evaluation of IMDB Records* (Statistics Canada's Intranet Site:http://stdsweb/standards/menu.htm)

[11] Standards Division, *Biennial Program Report - Standards Division FY 2002-03/2003-04*, October 2004, page 3

[12] The content of the IMDB records must accurately reflect the quality and methodology applying to the data for each individual reference period for which the data are disseminated. It follows that an IMDB record, once adequately completed, is not something done once for ever. Revisions to the content of an IMDB record are integrated in the regular survey process in order to insure that any changes in any steps of the survey and in any numerical measures of the accuracy of the resulting statistical estimates accompany the data as they are being disseminated.

[13] Statistics Canada, Standards Division, Qualitative Analysis of IMDB Records, 2003 and 2006

iv) Access

Users[14] can access individual IMDB records on the Agency's website through hyperlinks from *CANSIM* (Statistics Canada's on-line database), the on-line catalogue of products, *Summary Tables* and *The Daily*. Users can directly access the full list of Statistics Canada surveys, questionnaires and variables on the Definitions, Data Sources and Methods module that are organized alphabetically or by subject. Metadata can also be accessed through the *Information for Survey Participants* module on the website.

IV-Minimum Information Required

The information needed to understand statistical information falls under three broad headings:

- The concepts, variables and classifications that underlie the data
- The methodology used to collect and compile the data; and
- Indicators of the accuracy of the data

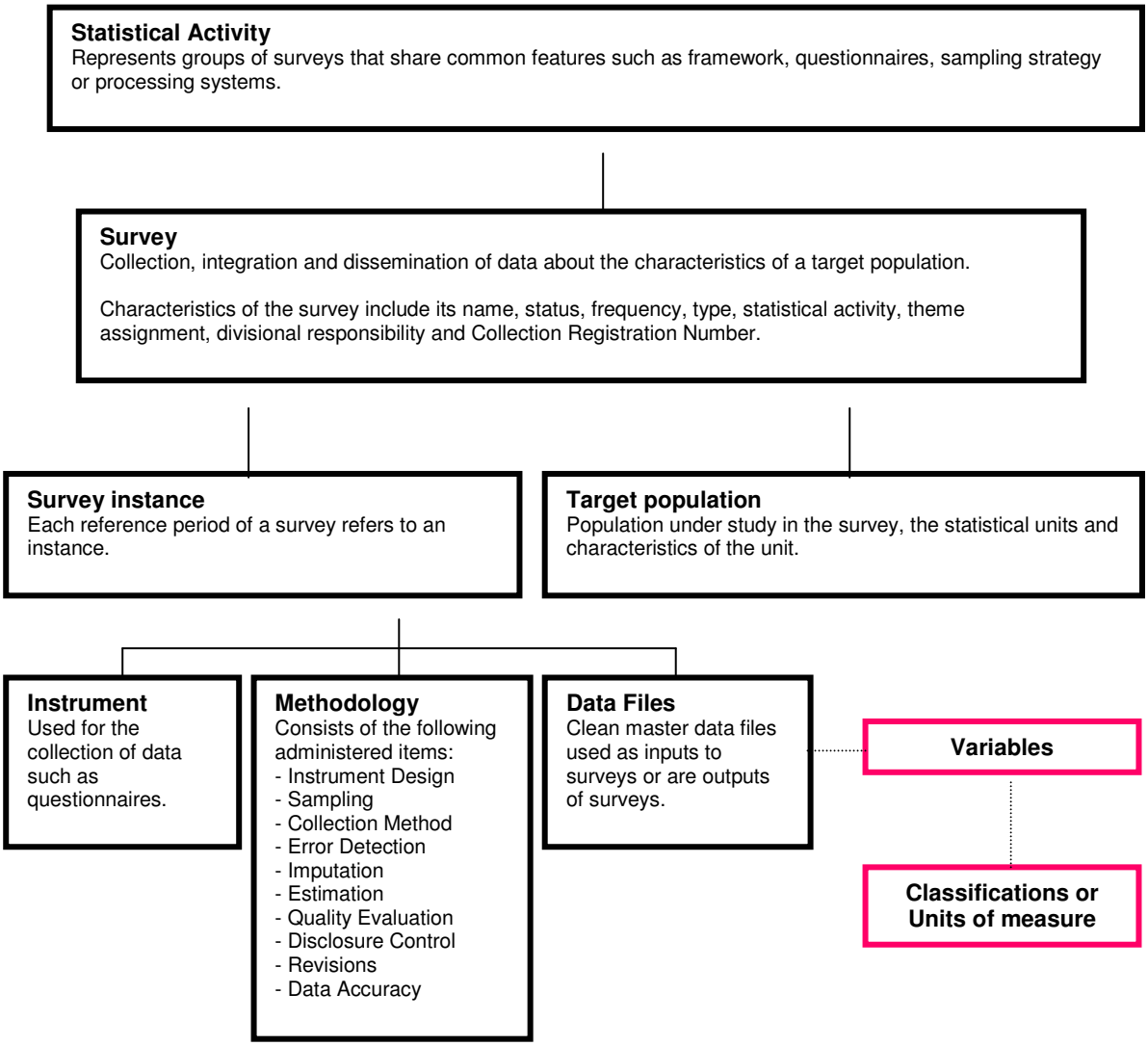
Each IMDB record is organized around the survey lifecycle. Figure I presents the minimum metatada information required under Statistics Canada's Policy on Informing Users of Data Quality and Methodology. Metadata are collected for each survey instance (each survey can have one or more survey instances, each representing one cycle of the survey) and made available in a standard format to data users on the Statistics Canada's website.

The information on the IMDB has 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. Metadata related to the Livestock Survey is a good example of the application of the structure displayed in Figure I[15].

[14] Born, Alice, *Metadata as a Tool for Enhancing Data Quality in Statistical Agency*, European Conference on Quality and Methodology in Official Statistics, 2004

[15] Born, Alice, *Metadata to Support the Survey Life Cycle*, Joint UNECE/EUROSTAT\OECD work session on statistical metadata, Geneva, 3-5 April 2006, pages 7 and 13

Figure I - Minimum Metatada Information Requred under Statistics Canada's Policy on Informing Users of Data Quality and Methodology



V- Agriculture Statistics

i) A centralized approach[16]

The production and dissemination of estimates on the agriculture sector is part of Statistics Canada's mandate. The statistical agency carries out monthly, quarterly, annual and/or seasonal data collection activities related to crop and livestock surveys and farm finances as needed, conducts the quinquennial Census of Agriculture in conjunction with the Census of Population, and produces and publishes economic series on the agriculture sector that flow to the System of National Accounts (SNA) to form the agriculture component of the Gross Domestic Product (GDP).

Administrative data supplement limited survey taking in supply-managed agriculture sectors such as dairy and poultry. Taxation data are mined extensively to tabulate annual disaggregated financial characteristics by major farm types, revenue classes and regions for Canadian farm operations.

The extensive cost-recovery program constantly evolves to meet the changing needs of clients, and provides valuable auxiliary statistical information to complement the core survey program on topics such as agro-environment, farm management, farm assets and liabilities, etc. The Farm Register maintains the agriculture survey frame and profiles large, complex agricultural operations via an Enterprise Portfolio Management style function[17] – essential to track increasing sectoral concentration.

Joint collection agreements are in force with most provincial and territorial departments of agriculture to minimize burden by reducing or eliminating duplicative local surveying. These agreements cover production surveys and, increasingly, provincially-administered stabilization program data. This latter source is important for verifying survey data and has the potential for replacing some survey data in the future.

ii) Agriculture Statistics Metadata

As can be seen in Figure II, all of the above collection activities are an integral part of the Agriculture Statistics Framework that feed into the SNA for the calculation of the GDP for the agriculture sector. They are part of the Income and Expenditures Accounts and are also used in the Input/Output Tables.

An integrated statistical framework plays two important roles: a quality assurance role and a "fitness for use" role facilitating the interpretability of the information

- Quality Assurance Role

The quality assurance process of the agriculture statistics program is carried out in two phases: on an annual basis, with the provision of information to SNA and on a quinquennial basis with the release of the Census of Agriculture data.

As can be seen in Figure II, the Farm Register provides the frame for most survey activities; the crop, livestock and financial data series serve as inputs to the derivation of the farm income series.

---

[16] Statistics Canada, Agriculture Division, *Quadrennial Program review Report 2001/02-2004/05*, page 1

[17] The Enterprise Portfolio Management function is responsible for managing all aspects of Statistics Canada's relationship with Canada's largest businesses, including organizational profiling, survey reporting arrangements, issue resolution, coherence analysis and, eventually, data collection.

In turn, the production of the farm income estimates is a coherence exercise that allows the validation and revision (if necessary) of the input data.

Every five years, the same holds true with the Census of Agriculture. It allows for an update of the Farm Register used to draw sample for the various crop, livestock and financial surveys and a base to revise estimates (including farm income estimates) produced between censuses. In turn, sample surveys and farm income series are useful tools to validate census data.

- A "fitness for use" role facilitating the interpretability of the information

As mentioned previously, the agriculture statistics program derives its information using various methodologies such as census, survey, administrative data and derived information such as the farm income indicators estimated using a combination of administrative data and survey estimates.

The IMDB includes 45 separate records covering each of the agriculture statistics program's current survey activities[18]: IMDB record 3438 related to the Census of Agriculture, 3450 to the Farm Financial Survey, 3432 to Milk Sold Off Farms and Cash Receipts from the Sale of Milk and 3437 to Farm Cash Receipts are examples of metadata information on each of the various methodologies used to produce estimates related to the agriculture sector.

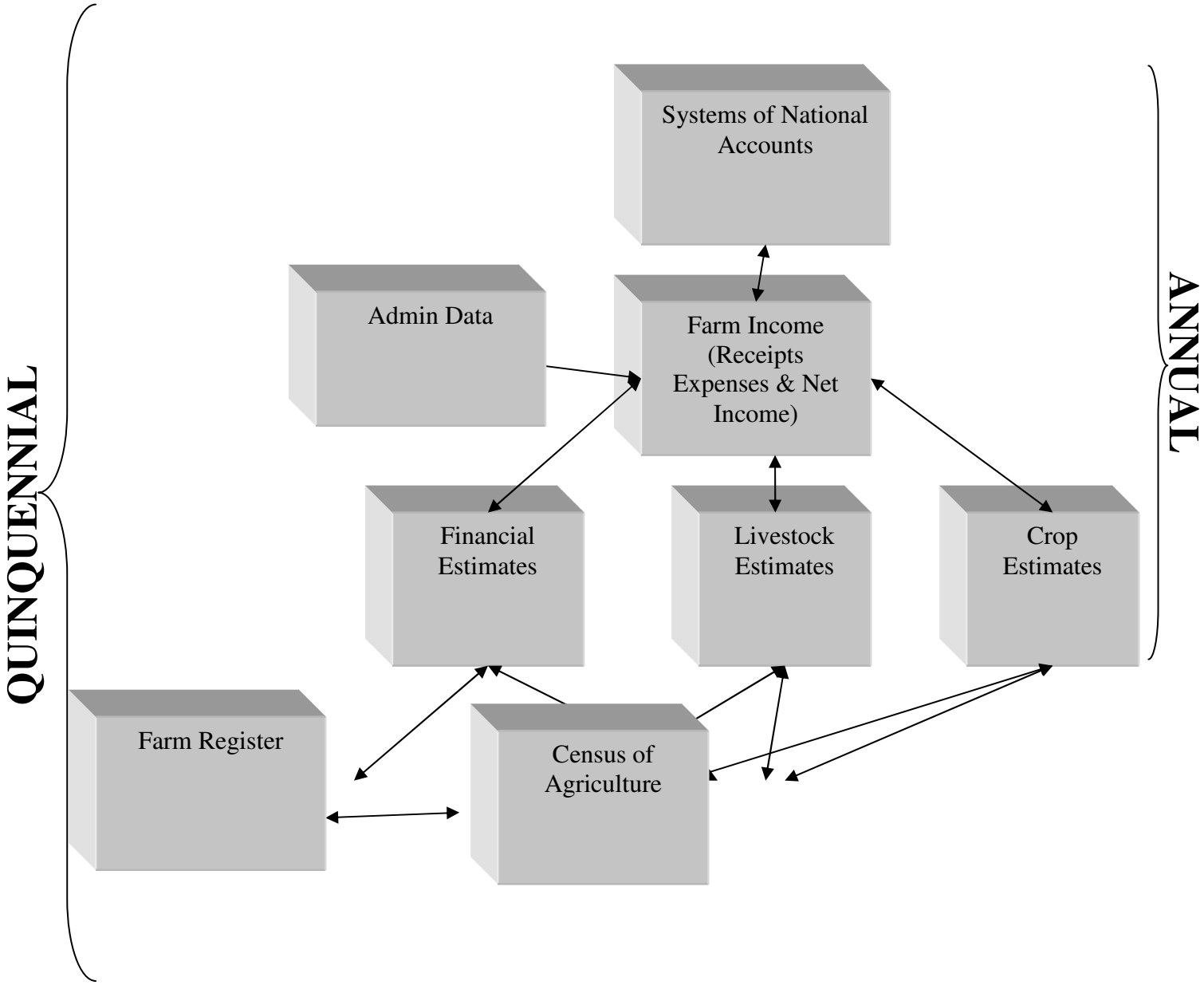
Statistics Canada's IMDB offers the possibility to go one step further in illuminating the "fitness for use" for data users by organizing the agriculture statistics program in a statistical activity. A statistical activity is another level of the IMDB structure that groups together surveys that share common processing system or conceptual framework[19]. Such an approach has been created for the Canadian System of National Accounts (CSNA) and the Unified Enterprise Survey. A similar approach could be applied to agriculture statistics: metadata organized around the hierarchical structure of the farm income series shown in Figure II would facilitate users' understanding of the interrelationships between the various statistical activities. It would also increase users' awareness that the "fitness for use" test of the farm income series should take into account the metadata information of all the farm income series' inputs. Such an approach could be used for other statistical agriculture series such as the Cash Flow Account, the Balance Sheet, etc.

Conclusion

Statistics Canada defines data quality in terms of "fitness for use". Six dimensions of quality have been identified within the concept of "fitness for use": relevance, accuracy, timeliness, accessibility interpretability and coherence. Metadata are at the heart of the management of the interpretability indicator. The Integrated Metadata Base is Statistics Canada's single source of metadata information describing surveys and programs. The quality of the IMDB information has to be monitored regularly to ensure completeness and accuracy. It is important for statistical agencies to publish good metadata because by doing so they show openness and transparency and breed trust with data users.

[18] Metadata for all agriculture surveys is available at http://www.statcan.ca/english/sdds/indext.htm
[19] Born, Alice, *Metadata to Support the Survey Life Cycle*, Joint UNECE/EUROSTAT\OECD work session on statistical metadata, Geneva, 3-5 April 2006, page 4

Figure II - Agriculture Statistics Framework



References:

Born, Alice, *Metadata as a Tool for Enhancing Data Quality in Statistical Agency*, European Conference on Quality and Methodology in Official Statistics, 2004

Born, Alice, *Metadata to Support the Survey Life Cycle*, Joint UNECE/EUROSTAT\OECD work session on statistical metadata, Geneva, 3-5 April 2006

Johanis, Paul, *Role of the Integrated Metadata Base at Statistics Canada*, Statistics Canada International Symposium Series – Proceeding, 2001

Statistics Canada, Agriculture Division, *Quadrennial Program Review Report 2001/02-2004/05*

Statistics Canada, Standards Division, *Biennial Program Report - FY 2002-03/2003-04*, October 2004

Statistics Canada, Standards Division, Qualitative Analysis of IMDB Records, 2003 and 2006