HOW USEFUL ARE STATISTICAL METADATA IN ASSESSING DATA QUALITY?

Submitted by OECD[1]

Invited paper

SUMMARY

That statistical metadata are needed to assess quality is a motherhood statement. International national organisations expend a great deal of effort in disseminating metadata to accompany the statistical data they produce. Such metadata are made available through a variety of media such as paper publications, electronic products and websites. They are usually compiled without reference to international metadata content standards, of which there are few. This paper assesses how much use these metadata are to users in obtaining an assessment of the quality of statistical data.

In the absence of any universally accepted definition of the term "quality", the paper utilises the six elements of quality defined in Statistics Canada's Quality Guidelines, namely: relevance, accuracy, timeliness, accessibility, interpretability, and coherence. The paper draws extensively on OECD experience in the dissemination and use of statistical metadata for short-term economic indicators, including the preparation of reports that compare the methodologies used by OECD Member countries.
The paper concludes with the authors' views on areas where international standards for metadata need to be developed to enhance the usefulness of metadata in assessing data quality. Groups such as METIS and Metanet could undertake such developments and promote use of the resulting standards by international and national organisations.

I.   INTRODUCTION

1.   The starting point in discussing how useful users find statistical metadata in assessing quality is a common understanding of the terms quality, statistical metadata and user.

2.   The quality of data in general, and of any dataset in particular, is interpreted in this paper in the broad sense of fitness for use. To translate this general concept into one that can be applied practically in a statistical office, quality is viewed as comprising a set of quality components. Various sets of components have been defined by various individuals and agencies, for example, by Fecso (1989), Statistics Sweden (1995), Statistics Canada (1998), Eurostat (2000a), and the IMF (2000). There is no international standard nor universally accepted quality components model. The Eurostat and Statistics Canada models are quite similar. The IMF quality model, defined in the context of the Data Quality Assessment Framework (Carson, 2000) introduces the notion of the quality of the statistical system generating the data – i.e., of the statistical production process and of the institutional infrastructure supporting that process - as well as that of the data product per se.

---

[1]   Prepared by Michael Colledge and Denis Ward.

3.   It is not the purpose of this paper to compare the alternative quality models nor to attempt to identify the circumstances in which they are most appropriate. We have simply and somewhat arbitrarily selected the quality model in Statistics Canada's Quality Guidelines as the basis on which we discuss assessment of data quality. The corresponding six quality components (relevance, accuracy, timeliness, accessibility, coherence and interpretability) are elaborated below in Section 3.

4.   Statistical metadata are data describing statistical data (UN/ECE, 2000). The qualifier statistical simply implies that the metadata are referring to statistical data rather than to some other type of object, such as library books or museum artifacts. The qualifier is mostly omitted in this paper as the context is evident.

5.   This definition of statistical metadata, like that of quality, is rather too general for direct practical application. Statistical metadata come in all shapes and sizes and it is helpful to categorise them. Following Colledge (1999), for the purposes of this paper they are classified into five types, namely:

??   definitional - describing statistical units, populations, classifications, data items, questions and question modules and other statistical terms;
??   procedural - describing the procedures by which data are collected, processed and disseminated - for example, survey frame construction, follow-up, adjustment for non-response, and estimation procedures;
??   operational - describing the results of implementing the data collection, processes and dissemination procedures - for example, measures of respondent burden, response rates, edit failure rates, and costs;
??   systems - metadata used to drive automated operations for collection, processing and dissemination - for example, file layouts and access paths;
??   dataset - the minimal metadata required to present a dataset, including title, data item names, units of measurement, reference periods and footnotes.

I.1   Use of metadata

6.   There are two general types of reasons why a statistical office collects, maintains and disseminates metadata:

??   for process oriented reasons - to inform the suppliers of the data - for design, redesign, and operations;
??   for product oriented reasons - to inform the users of the data - for quality assessment, i.e., determining fitness for use.

7.   This paper is concerned with the second of these groups of reasons, i.e., use of metadata by users for quality assessment. User here refers to the person using the data. A user thus defined is interested in the metadata only as a means to an end (namely assessment of fitness for use) in contrast to the supplier of data who is interested in the metadata per se.

8.   Users can assess data quality, i.e., determine the characteristics of data in sufficient detail to determine their fitness for use, by two basic mechanisms:

??   through experience and evaluation, i.e., by analysing the results of using data and by seeing whether the data have led to useful or misleading conclusions, whether it is useful in practice - this requires considerable time and effort on the part of the user;
??   by using metadata that accompany the data - i.e., by assessing data quality through the metadata that accompanies them - this can be done quickly with much less effort on the part of the user.

9.   This paper focuses on only on the latter, i.e., the actual use of metadata to assess the components of quality.

10.   For any particular dataset, two aspects of quality may be examined from a user perspective:

??   the quality of the dataset, i.e., the extent to which the dataset actually satisfies user data needs;
??   the quality of the metadata accompanying the dataset, i.e., the extent to which the metadata satisfies user needs in allowing users to assess the quality of the dataset, i.e., fitness for purpose.

11.   This paper focuses on the second aspect, i.e., the extent to which the metadata meet the needs of users in allowing them to assess the quality of the data. Users must be able to assess quality; they must be convinced that the data are fit for use. As noted by Fellegi (1996) the reputation of a statistical office depends not only upon the quality of its data but also upon users' perceptions of that quality - on the credibilty of the data. General concerns that users may have regarding quality of data include whether the data are:

??   accurately reported by the original respondents;
??   are manipulated by the statistical office;
??   collected and processed using a good methodologies.

I.2   Metadata quality components and presentation

12.   This is a paper about the quality of metadata and, just as for data, the quality of metadata may be defined in terms of fitness for use. Thus, as exemplified by Colledge and Boyko (2000), the quality of metadata may be measured in terms of an equivalent quality model with six components. Dippo et al (2000) describes assessment of metadata in terms of two components - relevancy and accuracy. However, in both papers assessment is from a supplier and user perspective whilst in this paper metadata quality is considered only from a user viewpoint.

13.   There are many types of user and uses for any given dataset. The wide range of possible users and uses means that a broad spectrum of metadata requirements has to be addressed. In particular the data supplier must make sufficient metadata available to enable the least and the most sophisticated users to assess quality readily. This implies segmentation of users into groups and a layered approach to metadata presentation, in which each successive layer provides more detail.

14.   Typically metadata are made available at two levels and in two distinct forms:

??   dataset metadata presented as an integral part of the data tables; and
??   additional (definitional, procedural and operational) metadata that may accompany the tables or be presented separately via the Internet or in occasional publications.

15.   A study of the metadata available from national statistical office (NSO) and national bank websites in OECD member countries was conducted by the OECD (2000a). According to the study, about one half of the NSOs disseminate detailed metadata for at least some of their major datasets, typically, labour force, consumer prices, production, domestic and external trade. Of the remaining NSOs, about two thirds disseminate summary metadata and the remainder do not disseminate any additional metadata via the web. Of the national banks, about 30% produce detailed metadata, about 30% summary metadata, and the remainder no metadata. Metadata availability can also be assessed by observing the countries that have subscribed in accordance with the Special Data Dissemination Standard (SDDS) (IMF, 1997) and the General Data Dissemination Standard (GDDS) (IMF, 1998).

I.3 Outline of this paper

16.   Section 2 of this paper provides a brief description of the OECD's Main Economic Indicators (MEI) programme (including a description of the programme's metadata strategy). Section 3 discusses metadata quality issues in the context of Statistics Canada's six quality components. It defines what is meant by the each component and indicates how a user might use the metadata typically available to assess quality in terms of that component. Each case is exemplified using the MEI dataset. Section 4 presents general conclusions and recommendations where further work on international metadata standards by international organisations and groups such as METIS would enhance the usefulness of metadata in user assessment of data quality.

II.   OECD MAIN ECONOMIC INDICATORS PROGRAMME

17.   This section describes features of the OECD Main Economic Indicators (MEI) programme pertinent to the use of its metadata for quality assessment by users. A more detailed description of the programme is provided in OECD (2002).

18.   First, we make a remark on the general role of the OECD in terms of data collection. Typically, data are not acquired directly from the original respondents - persons, households, businesses, etc. Rather, they are obtained from national statistical agencies that have conducted the primary collection work. Data are "second hand". The OECD adds to them and repackages them, as is also the case for other international organisations. The value added by the OECD in performing this function is:

?? synergy - bringing together data from different sources;
?? verifying and editing the data - reasonableness checks sometimes reveal errors in the original data received from the donor countries that are subsequently corrected by those countries;
?? providing metadata to enable users to assess quality, metadata which may not be published by the donor countries;
?? in particular providing metadata to enable users to assess comparability of each indicator across countries, even providing comparability analyses;
?? producing supra-national data, i.e. aggregations of country data such as Total-OECD, Major seven, euro area, EU-15.

II.1 MEI data collection and dissemination

19.   MEI data are supplied by the 30 OECD Member countries and 11 non-member countries. The data items collected and disseminated, referred to as indicators, are those thought to be most relevant for short-term economic analyses.   The dissemination media include a monthly printed publication, a monthly CD-ROM, through SourceOECD, an on-line database accessible by subscription by the public, and monthly updates to OLISnet (a private network for OECD members).   The monthly printed publication contains current data for the most recent 12 months, 9 quarters and 4 years. Historical data covering longer periods of time are available from the CD-ROM, on-line via SourceOECD, OLISnet and on request from the MEI database (internal OECD users only).

II.2 Users and uses of MEI data

20.   As for many datasets produced by international and national agencies, the MEI dataset services a wide range of users and uses. Users may be classified into three groups:

?? internal to OECD - well known, with whom MEI staff are in constant touch;
?? national and international statistical organisations – whose needs are reasonably well known;
?? other external users – whose needs are not well known.

21.   One sense in which the MEI programme is not a typical OECD data collection is that was not set up specifically to satisfy the data needs of a particular OECD committee. The absence of a single dedicated user and the wide range of users, especially the body of external users, emphasises the need to provide access to comprehensive metadata. Given that each user's purpose cannot be anticipated, the MEI has to provide a broad range of metadata so that each user can evaluate quality from his/her perspective.

22.   As an indication of the significance of the external users, the OECD Statistics Portal has 40,000 hits per month. It is the most heavily accessed web page other than the OECD home page itself. As noted by Giovannini (2001), statistics are a very significant OECD output. User impressions of the OECD are determined by the perceived quality of the data they obtain. In this respect, users are influenced by the attention they see being given by the OECD to quality considerations, for example by the existence of a corporate quality framework.

II.3   Metadata collection and storage

23.   A peculiarity of the MEI dataset is that it is not a dataset collected by a single process from a single set of respondents within each country. Rather it contains data from many different sources within each country as well as across countries. The quality of the metadata thus varies significantly by indicator as well as by country. Notwithstanding this, the metadata for each indicator in the MEI database are held in a structure, referred to as the MEI metadata template, that is common to all indicators. There are actually two MEI metadata templates, namely:

??   a summary sources and definitions template containing five broad categories (definition, coverage, collection, calculation and source); and
??   and a more detailed sources and methods template, described by Petit et al (1996).

24.   MEI experience has shown that it is very costly to collect and maintain all the metadata specified by the more detailed template and the only metadata actually stored and maintained in the MEI database are those specified in the sources and definitions template. A detailed description of the current MEI metadata system is given in OECD (2002).

II.4   Metadata dissemination

25.   MEI metadata are disseminated:

??   via the monthly printed publication;
??   electronically via the OECD website (freely available at: http://www1.oecd.org/scripts/mei_sd/) CD-ROM, and SourceOECD;
??   in occasional sources and methods publications produced for a number of short-term economic indicators, namely: CPI, PPI, construction price indices, labour and wage statistics and domestic finance statistics (available at http://www1.oecd.org/std/meta.htm); and
??   in comparative studies that provide comparisons of current national practice for selected metadata elements for specific indicators against existing international statistical standards.

26.   The monthly printed publication metadata comprise:

??   the cover page title of the publication, back cover list of countries and general topics, index of indicators and introductory pages summarizing main features of the data tables and drawing attention to the issues of across country comparability;
??   dataset metadata, i.e., metadata directly presented with the data tables themselves, comprising names of indicators and countries, units of measurement, reference periods and footnotes describing selected unusual features;

??   additional dataset metadata in the form of data presentation notes, explaining symbols and conventions used in the tables; and

??   explanatory notes, describing selected aspects of the data collection and processing procedures, for example, construction of composite leading indicators and standardized employment rates.

27.   MEI metadata disseminated electronically are presented in accordance with the sources and definitions template, comprising five basic categories, the precise contents of which vary by indicator or group of indicators:

??   definition, including definitions of data items, inclusions and exclusions, indications whether value or volume, index or level;

??   coverage, including statistical units, geographic coverage, industrial coverage and relationship to ISIC Rev 3, inclusions and exclusions;

??   collection, including how the data are collected, frequency of collection, sample size, actual coverage of sample as proportion of population and of data item total, number of data items, response rates and treatment of non-response, use of administrative data;

??   calculation, including methods for computing estimates, treatment of trading days; and

??   source, meaning the agency or organization from which the data and metadata are acquired.

28.   It is doubtful whether MEI sources and methods publications will be produced in future. The intention instead is to provide users with universal record locator (URL) links to detailed metadata maintained on the websites of other international organisations and national agencies, and to focus OECD resources on the preparation of comparative studies of the indicators. The first in the series of such studies (for industrial production indices, retail and construction indicators) was published in January 2002 (available at www.oecd.org). Other studies in the series covering employment and unemployment and prices (CPI, PPI, and construction) will be produced in the first half of 2002.

III.   ASSESSMENT OF QUALITY USING METADATA

III.1   Introductory Remarks

29.   This section discusses the use of metadata in the assessment of quality by users. Quality is defined in terms of the six Statistics Canada components. The discussion draws on examples from the MEI metadata available to users in the printed monthly publication and electronically. Problems experienced by the OECD in MEI compilation and maintenance and in the preparation of the MEI cross country data comparison studies are outlined. General conclusions from these experiences and recommendations for future work on international metadata standards are provided in Section 4.

III.2   Assessment of Relevance

III.2.1   Definition

30.   "The relevance of data or of statistical information is a qualitative assessment of the value contributed by these data. Value is characterised by the degree to which data or information serve to address the purposes for which they are produced or sought by users. Value is further characterised by the merit of these purposes, in terms of the mandate of the agency, legislated requirements and the opportunity cost to produce the data or information." (Statistics Canada, 1998 page 4)

31.   From a user perspective, the merit of the purposes to which the data are to be put is not an aspect of data quality. Thus the following remarks focus solely upon the extent to which the users are able to assess the extent to which the data items address their needs.

III.2.2 MEI Example

32.   MEI dataset metadata, i.e., metadata embedded in the data tables, including footnotes and explanatory notes, provide a fairly good general indication of relevance.

33.   MEI metadata available from the OECD website are sufficient for the user to determine relevance in most circumstances. Examples with MEI metadata field indicated in parentheses are:

??   (definition) "Data are compiled according to SNA 93."
??   (definition) "Gross output covers the value of products manufactured by establishments during the accounting period, including the value of semi-finished products sold, or intended for sale, to other entities."
??   (definition) "Data measure the value of sales at current prices."
??   (coverage) "Indices are calculated for total industry based on an aggregation of 16 main branches…"
??   (coverage) "All types of outlets are covered, irrespective of the type of administration or ownership, including estimates for hidden trade and trade by private persons in informal markets."
??   (calculation) "Monthly figures refer to the average number of persons employed during the month."

III.2.3 General comments

34.   Dataset metadata are likely to be sufficient to indicate to a user whether the dataset has the potential to address the user's needs. For simple data uses, these metadata may be all that are required. For more complex data uses, the assessment depends upon access to more detailed definitional metadata (for definitions of statistical units, data items, classifications) and procedural metadata (for indication of sampled population and computation procedures). As mentioned above, the OECD is currently inserting URL links on the MEI database to more detailed metadata maintained by other international organisations and national agencies on their websites. Such detailed metadata are frequently not readily accessible on websites for countries for all MEI short-term economic indicators for any individual country.

35.   In summary, relevance can often be adequately assessed by unsophisticated users using only the dataset metadata that are embedded in the data tables. Relevance can be properly assessed by sophisticated users if and only if sufficient definitional and procedural metadata also accompany the data. This is often not the case. For the MEI, additional definitional and procedural metadata are available, but only in summary form. Some users may require access to more detailed metadata from the source national agencies, and, as previously noted, these are available from only 50% or so of the OECD member countries' statistical offices and national banks.

III.3   Assessment of Accuracy

III.3.1 Definition

36.   "Accuracy of data or statistical information is the degree to which those data correctly estimate or describe the quantities or characteristics that the statistical activity was required to measure. Accuracy has many attributes, and in practical terms there is no single aggregate or overall measure of it. Of necessity, these attributes are typically measured or described in terms of error, or the potential significance of error, introduced through individual major sources of error - e.g., coverage, sampling, non-response, processing and dissemination." (Statistics Canada, 1998 page 5)

37.   Accuracy is the quality component best understood by survey statisticians and least understood by users. As noted in the definition above, and elaborated by Groves (1989), Lessler and Kalsbeek

(1992), and Sarndal et al (1992), it comprises two basic aspects: sampling errors and non-sampling errors, with the latter typically being subdivided into coverage, measurement, processing, non-response and model assumption errors. Biemer and Fecso (1995) describe the assessment of measurement errors in business surveys.

III.3.2   MEI example

38.   The MEI data tables do not contain measures of accuracy, apart from an occasional indication of a data problem or limitation in the footnotes, for example:

??   "Series derived by OECD from index series with Previous Period=100".

39.   Whilst the MEI website metadata do not systematically present any measures of accuracy, they do contain some indications of the likelihood of certain types of errors scattered around the various metadata fields, for example:

??   (definition) "Subsidised goods are accounted for according to their full price."
??   (coverage) "Military reservations and compounds are not surveyed but are assumed to have the same unemployment rates as elsewhere."
??   (collection) "In 1998, about 27,000 units were surveyed, representing 83% of total industrial output".
??   (collection) "The non-response rate is 2 or 3% for large and medium enterprises and around 30% for small enterprises".
??   (calculation) "There is no specific methodology for the treatment of seasonal items"

III.3.3   General comments

40.   The non-systematic description of the measures of accuracy in MEI metadata is, by and large, a reflection of the wide range of national practices in this area. There are often insufficient metadata for a user to be able to assess the accuracy of a national dataset. Sampling errors are occasionally provided. Other types of errors are even less frequently reported. In such circumstances, users have no choice but to rely heavily on the reputation of the data producers.
Much of the metadata available on websites and in paper publications is unstructured and users frequently have to sift through a lot of text to find measures of accuracy of interest to them. Sampling errors and response rates are the most frequently reported operational metadata that reflect upon accuracy.

41.   Users might reasonably conjecture that the existence of procedural and operational metadata is correlated with good quality in terms of accuracy.

III.4   Timeliness

III.4.1   Definition

42.   "Timeliness of information reflects the length of time between its availability and the event or phenomenon it describes, but considered in the context of the time period that permits the information to be of value and still acted upon." (Statistics Canada, 1998 page 5)

III.4.2   MEI example

43.   The reference periods to which the data refer are clearly evident as the column headings in most tables. Comparisons in the timeliness of the various short-term indicators between Member countries can readily be made through examination of the subject tables in Part One of the MEI. These tables provide a general indication of relative timeliness amongst countries, though some differences could be due to

OECD database update procedures. The vision for the future is for updates to be made in "real-time", i.e. the inclusion of data in the MEI database on the day of their release by national agencies.
General comments

44. Only very limited metadata are needed to assess timeliness and they are almost invariably available. The user gets direct measure of timeliness from the existence of the data themselves and dataset metadata stating the reference period to which the data refer.

III.5 Assessment of Accessibility

III.5.1 Definition

45. "Accessibility reflects the availability of information from the holdings of the agency, also taking into account the suitability of the form in which the information is available, the media of dissemination, the availability of the metadata, and whether the user has reasonable opportunity to know it is available and how to access it. The affordability of that information to users in relation to its value to them is also an aspect of this characteristic." (Statistics Canada, 1998 page 5)

III.5.2 MEI example

46. The accessibility of MEI metadata is described in Section 2. Two practices worth reiterating are the linkage of MEI sources and definitions metadata to actual data and the practice of systematically making such metadata readily accessible by placing them on the Internet, accessible free of charge.

III.5.3 General comments

47. Access takes place through metadata, thus metadata and the organisation of metadata determine accessibility. Factors determining accessibility include the range of media on which the data are available, cost of access, speed of response and search software in the case of electronic access etc. OECD experience in the compilation of sources and methods metadata for inclusion in the MEI database and in the preparation of the comparison publications highlighted the diversity of national practices with regards to metadata accessibility. National agencies frequently cite the availability of metadata in publications that are not readily available outside the country or are out of print. The systematic adoption of a policy of inserting such metadata onto websites would constitute good practice.

48. However, notwithstanding the pre-eminent part that metadata play in access, users do not need metadata to assess the quality of access. They can assess accessibility directly themselves simply by accessing.

III.6 Assessment of Coherence

III.6.1 Definition

49. "Coherence of data and information reflect the degree to which the data and information from a single statistical programme, and data brought together across datasets or statistical programmes are logically connected and complete. Fully coherent data are logically consistent - internally over time and across products and programmes. Where applicable the concepts and target populations used or presented are logically indistinguishable from similar, but not identical, concepts and target populations of other statistical programmes, or from commonly used notions or terminology." (Statistics Canada, 1998 page 5)

50. This definition includes two aspects that are identified as separate quality components in the model defined by Eurostat (2000a), namely:

?? consistency of data over time (coherence in Eurostat model);

?? consistency of across data items and datasets (comparability in Eurostat model).

III.6.2 Consistency over time

51.   Key points in assessing consistency over time are:

?? whether the metadata reveal changes in definitions, procedures and operations over time;

?? how changes in definition or classification are handled;

?? how "series breaks" are defined, in particular, the criteria that determine the choice between (1) the termination of an indicator series and the start of a new one (2) a break being flagged in an ongoing single series, (3) no indication being given of a change;

?? how series breaks are handled, e.g., by "wedging in" the changes over time, by providing adjustment factors, or by backcasting the series.

III.6.3 MEI example

52.   of the strengths of the MEI is the presentation of long time series for most of the indicators included in the database. The decision as to whether or not changes in national definitions, etc, constitutes a series break is made on a subjective basis. MEI practice is to terminate the old series and to present the data as a new series when a significant change is judged to have taken place. Each series is accompanied by its own metadata. In this respect, the OECD is reliant on national agencies providing information on changes in methodology and, unfortunately, national practices across Member countries vary enormously. Some systematically provide information about all changes, some provide no information whatsoever. In the case of the latter the OECD depends upon data verification systems to detect the more obvious breaks. It is fair to say that there is still room for improvement in the compilation and dissemination of metadata describing series breaks and in presentation practices in this area in some MEI tables and graphs.

III.6.4 Consistency across data items and datasets

53.   This is a particularly significant quality attribute in the case of data produced by an international organization, given that a good portion of the value added by an international organisation arises from bringing data together to enable comparisons and that the data are obtained from different countries with potentially different definitions, procedures and operations.
Users need to know how comparable data are across countries. In this respect they are utterly dependent upon metadata. Comparability cannot be observed or deduced from the data themselves. The same data item may be presented under different names, conversely, different data items may have a common name.

III.6.5 MEI example

54.   Use of MEI metadata to make data comparisons between countries indicates the main deficiency of metadata currently stored in the MEI database. Recent experience in preparing the MEI data comparison publications highlighted two major problem areas:

?? National agency metadata (when available) frequently do not describe the same aspects of methodology, in other words the metadata required for comparing specific aspects of methodology are simply not available for all 30 Member countries.

?? Differences in terminology used by different countries. Individual national agencies often use different terms to describe the same phenomenon or the same term to describe different phenomena.

55.   Both aspects significantly increase the resources required to finalise the comparison publications.

III.6.6 General comments

56.   The absence of any universally adopted metadata content standard is a major area for future work by international organisations and groups such as METIS. The issue of terminological differences in metadata used for international comparisons is covered in detail by Ward and Pellegrino (2001). The approach described in the paper is the adoption by metadata authors of standard definitions of metadata elements drawn from existing international statistical standards and the availability of such definitions in a readily accessible glossary.

III.7   Assessment of Interpretability

III.7.1 Definition

57.   "Interpretability reflects the ease with which the user may understand and properly use and analyse the data or information. The adequacy of the definitions of concepts, target populations, variables and terminology underlying the data, and the information on any limitations of the data largely determines their degree of interpretability." (Statistics Canada, 1998 page 5)

III.7.2 MEI example

58.   Most MEI metadata could simplistically be described as "descriptive", providing limited information on current national definitions, collection and compilation practices. They are of very limited use in helping users make appropriate use of data or in developing an understanding of the adequacy of concepts, data items, etc.

III.7.3 General comments

59.   This is an area in which metadata seem to be generally weak. It would help if metadata were accompanied by information outlining the appropriate use of the data. For example, the uses of the different measures of employment could be explained, in particular of those derived directly from household surveys and those compiled for the national accounts.

IV.   CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

60.   The aim of this paper is to review the statistical metadata that typically accompanies datasets from the perspective of users wishing to determine the quality of the data to which the metadata refer. The assessment is in terms of six quality components - relevance, accuracy, timeliness, accessibility, coherence, interpretability. Examples are drawn from the OECD Main Economic Indicators (MEI) dataset. In rather general terms, the conclusions are as follows.

61.   Although in this paper quality is defined in terms of six components, there is actually no common vocabulary for describing quality. Thus users' perception of the utility of metadata in assessing quality is likely to depend upon the framework within which they are viewing quality.

62.   Many datasets are published with minimal metadata, i.e., dataset metadata comprising the title row and column headings and possibly table footnotes. This is never sufficient for a comprehensive quality assessment. In particular accuracy, coherence and interpretability cannot be assessed without additional definitional, procedural, and operational metadata. Even relevance is difficult to assess properly.

63.   Where more detailed metadata are made available they are more often suited to describing data and procedures from the perspective of a producer than that of a user. In the case of the MEI, for example, the metadata describe current national practices quite effectively, but are difficult to use for in depth analyses of cross country comparability.

64.   Providing sufficient metadata is a major challenge for most data producers, especially for international agencies that are disseminating data that they themselves have not collected directly. The only solution is to share metadata and metadata standards. As an illustration, the MEI metadata strategy is now to maintain fairly minimal sources and definitions metadata in the MEI database using a very simple model and to link these metadata to more detailed metadata maintained by the national agencies that supplied the data and by other international organisations. Necessary conditions for effective operation of this (virtual) extended MEI metadata repository are co-operation between the agencies and organisations and their universal adoption of the practice of placing metadata on websites. In order for such detailed metadata to be efficiently used for international comparisons, the development of metadata content standards is also required.

65.   There are few widely accepted metadata standards. In fact there is only one international metadata standard, namely ISO 11179. The focus of this standard is on metadata describing data elements. However, parts of the standards, in particular that dealing with metadata registration can be applied more generally. The standard has not yet been widely adopted.

66.   Additional metadata standards for operational and procedural metadata would be invaluable. Such standards should address both data producer needs and data user needs, which are different. Furthermore, international agencies are in a special situation because the data they disseminate are usually collected by others, i.e., national statistical offices, banks, etc. Thus, it is not simple to define operational and procedural metadata standards in a way that suits all needs. For example, even in the simple MEI sources and definitions metadata template there is a lot of crossover between the metadata categories.

67.   Some specific suggestions are as follows:
?? Widespread development and adoption of a common terminology for procedural and operational metadata, based on existing international standards, facilitated through the development of a common glossary. Work in this area would, for example, facilitate the efficient use of metadata for national comparisons;
?? Formulation of supplements to such a glossary that would assist users to make appropriate use of data in specific problem areas, for example, helping them in their choice of employment measure, depending upon the use to which they are intending to put the data;
?? Development of standards for the presentation of metadata on websites. Issues that need to be considered here include navigation and search facilities, procedures for regular maintenance, guidelines on best practice to overcome the problem of unstable URLs. Rauch (2000) provides a very useful starting point for the preparation of such standards.

68.   Finally, it must be said that metadata alone will never be sufficient for the major users. In addition, to providing comprehensive metadata, data producers must engage in face to face consultation with major users to discuss their needs and quality requirements and the strengths and limitations of the data and metadata being provided.

REFERENCES

Biemer, P.P. and R.S. Fecso (1995) *Evaluating and Controlling Measurement Errors in Business Surveys*, in Cox et al (eds) *Business Survey Methods*, John Wiley, New York pp 257-281

Colledge, M.J., and E. Boyko (2000) UN/ECE Work Session on Statistical Metadata (METIS), Washington November 28-30;

Colledge, M.J. (1999) *Statistical Integration through Metadata Management*, International Statistical Institute, Voorburg: 67, 1, pp 79-98.

Dippo, C.S., F.G.Conrad, and D.W.Gillman (2000) *Metadata and Data Quality*, UN/ECE Work Session on Statistical Metadata (METIS), Washington November 28-30;

Eurostat (2000a) *Definition of Quality in Statistics*, doc. Eurostat/A4/Quality/00/General/Definition, presented to Working Group 5/4/2000, Eurostat, Luxembourg.

Eurostat (2000b) *Standard Quality Report*, working paper Eurostat, Luxembourg.

Fecso R.S. (1989) *Back to the Future*, Proceedings of the ASA, Washington

Fellegi, I.P. (1995) *Characteristics of an Effective Statistical System*, Morris Hansen Lecture presented to the Washington Statistical Society October 25, available from Statistics Canada, Ottawa.

Giovannini (2002) *The OECD Statistics Strategy: Introduction of a Corporate Quality Framework*, internal working paper, OECD, Paris

Groves, R.M. (1989) *Survey Errors and Survey Costs*, John Wiley, New York.

IMF (1999a) *Special Data Dissemination Standard,* International Monetary Fund, Washington, most recent version 1999; electronic copy available at dsbb.imf.org

IMF (1999b) *General Data Dissemination Standard,* International Monetary Fund (2000) International Monetary Fund, most recent version 1999; electronic copy available at dsbb.imf.org

IMF (2000) *Data Quality Assessment Framework*, International Monetary Fund, Washington: electronic copy available at dsbb.imf.org/glossary.pdf

Lessler, J.T. and W.D.Kalsbeek (1992) *Non-Sampling Errors in Surveys*, John Wiley, New York.

OECD (2000a) *Summary of Metadata Available on National Statistical Agency and Central Bank Websites*, working paper, OECD, Paris.

OECD (2000b) *The Role of Metadata in Promoting International Comparisons and Adherence to International Statistical Standards*, Ward, D.,OECD, Paris; available at www.oecd.org

OECD (2002) *Metadata System for the OECD's Main Economic Indicators*, Ward, D., OECD, Paris, UN/ECE Work Session on Statistical Metadata (METIS), Luxembourg, March.

Petit G., P.Beziz and R.van Eck (1996) *List of Metadata Items for OECD's Main Economic Indicators*, UN/ECE Work Session on Statistical Metadata (METIS), Berlin October.

Rauch, L. (2000) *Best Practices in Designing Websites for Dissemination of Statistics*, presented at METIS, Washington DC, November 2000; copy available at http://www.unece.org/stats/documents/2000/11/metis/21.e.pdf

Sarndal C.E., B.Swensson and J.Wretman (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Statistics Canada (1998) *Quality Guidelines*, Third Edition, Statistics Canada, Ottawa; electronic copy available at www.statcan.ca

Statistics Canada (1992) *Policy on Informing Users of Data Quality and Methodology*, Policy Manual, Statistics Canada, Ottawa.

UN/ECE (2000) *Terminology on Statistical Metadata*, Conference of European Statisticians Statistical Standards and Studies – No. 53

Ward. D., and Pellegrino (2001) *Developing a Common Understanding of Standard Metadata Components: A Statistical Glossary*, OECD and Eurostat, presented at METIS, Luxembourg, March 2002.