

How Useful Are Statistical Metadata  
in Assessing Data Quality?

by

Michael Colledge and Denis Ward



## Content of Presentation

- Concepts
  - quality components, types of metadata
- Assessment of data quality in terms of quality components
- Example of assessment using MEI data
- Conclusions

## Quality: Definition and Components

- Quality is fitness for use
- Components of quality
  - Fecso (1989): relevance, accuracy, timeliness and cost
  - Statistics Canada (1998): relevance, accuracy, timeliness, accessibility, coherence and interpretability
  - Eurostat (2000)
  - IMF (2000): data quality assessment framework

## Assessment of Data Quality by User

- Directly
  - through practical observation and experience
  - by accessing and using the data
- Using the metadata accompanying the data

## Types of Metadata

- Systems - used to drive automated operations
- Dataset
  - title, data item names, reference period(s), measurement units, cell row and column annotations
- Definitional - units, populations, classifications
- Procedural - descriptions of procedures
- Operational - implementation measurements
  - response rates, edit failure rates, imputation rates

## Quality of Metadata

- Fitness for use
  - by users in locating data, understanding data, assessing data quality
  - by producers in designing and operating data collection, processing and dissemination
- In this paper we focus only on use of metadata by users to assess data quality

## Quality Component: Relevance

- Definition
  - The relevance of data or of statistical information is a qualitative assessment of the value contributed by these data. Value is characterized 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 characterized 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

## Use of Metadata in Assessment of Relevance

- For simple use of data
  - example: consumer price index
  - use by journalist, student
  - dataset metadata are generally sufficient
- For sophisticated use of data
  - example: for detailed economic analysis, policy making
  - need definitional and procedural metadata, which are often not readily available

## Quality Component: Accuracy

- Definition
  - 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.

## Use of Metadata in Assessment of Accuracy

- For simple use of data
  - example: consumer price index
  - use by journalist, student
  - don't need metadata - rely on "agency label"
- For sophisticated use of data
  - example: for detailed economic analysis, policy making
  - need definitional and procedural metadata, often not available

## Quality Component: Timeliness

- Definition
  - 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
- Assessment
  - directly by user in accessing the data

## Quality Component: Accessibility

- Definition
  - 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
- Assessment
  - directly by user in making the access

## Quality Component: Coherence

- Definition
  - 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

## Assessment of consistency over time

- Assessment requires indications of:
  - changes in definitions, procedures, operations over time
  - how series breaks are defined
  - how changes are handled
- Dataset annotations sometimes indicate changes
- Changes in procedures or operations often not indicated - changes are “wedged in”

## Assessment of Consistency across Datasets

- Assessment requires detailed definitional and procedural data
  - often not available
  - difficult to make assessment due to differences in presentation of these metadata
- Clear case for international standard

## Quality Component: Interpretability

- Definition
  - 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



## Assessment of Interpretability

- Definitional, procedural and operational metadata
  - generally limited
  - often presented from producer perspective
  - not geared to user
- Clear case for international standard

## OECD MAIN ECONOMIC INDICATORS (MEI) PROGRAMME

- Is described briefly in our paper
- Also, in more detail in Room Document 1 for this meeting, particularly the metadata system

## MEI overview

- Provides extensive range of short-term economic indicators, plus aggregates for G7, EU12/EU15, OECD-total
- Extensive time series for 30 OECD and 11 non-member countries
- Data are disseminated monthly in paper publication, CD-ROM, csv file, SourceOECD (on-line), OLISnet
- Content primarily driven by needs of OECD Secretariat and Committees

## MEI METADATA - OVERVIEW

Comprises:

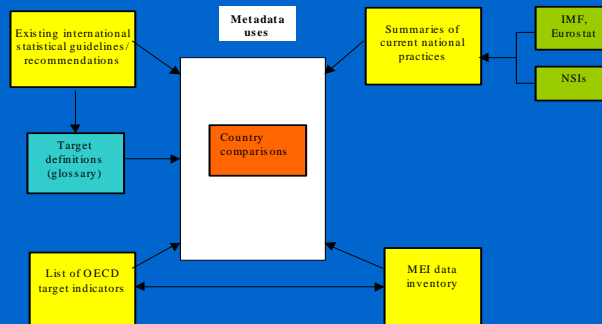
- summary metadata (S&D) using template based on five broad categories (definition, coverage, collection, calculation and source);
- located in database and is linked to data;
- more detailed sources and methods metadata a more detailed metadata template (prices, domestic finance, labour and wage indicators) - discontinued
- new MEI Methodological Analysis publications - compares national methodologies - against each other and with relevant standard

## MEI METADATA - OVERVIEW

MEI summary metadata:

- is focused on providing information to internal and external users;
- updated on on-going basis;
- disseminated in monthly CD-ROM with data, free of charge on internet; in a paper publication
- is compiled from national publications/websites + other I/O websites - rarely go directly to national agencies;
- recent facility to include URLs in database to more detailed metadata available at national agency and other I/O websites
- use a layered presentation approach - table headings/footnotes + explanatory notes + S&D + URLs









## MEI METADATA - OVERVIEW



## MEI METADATA - OVERVIEW

- Approach to metadata is highly pragmatic
- Linked to resources available to maintain it
- Deliberately aim to maintain minimal metadata in our databases
- Aim is to minimise reporting burden on member states for metadata through making use of existing sources and metadata from other I/Os - more use of URL links
- Want to devote more effort to analysing issues of data quality - e.g. comparability

## MEI METADATA QUALITY REPORT

Quality component	Assessment	Issues
relevance		Reasonable amount of metadata available to enable user to assess relevance, though not necessarily quality.
accuracy		Not really done at all.
timeliness	 	Not systematic/comprehensive. Review of Part 1 of MEI is one way of assessing relative timeliness of the data
accessibility		Linked to data. Available on website.
coherence		Cross country comparability difficult
interpretability	 	Only "descriptive" designed to give transparency to national practices -more required on appropriate use of data

•  
•  
•

## CONCLUSIONS AND RECOMMENDATIONS FOR FUTUREWORK

Existing standards should be more readily available in one location

Additional metadata standards required:

- standards for common terminology of metadata elements;
  - more metadata to assist users in appropriate use of data in specific problem areas;
  - development of standards for presentation on websites - navigation and search facilities, principle of free access, regular maintenance (Rauch 2000 a good starting point)
- •  
•  
•  
•  
•  
•  
•