STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 9
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Washington, D.C., United States, 28-30 November 2000)

Topic (ii): Metadata modelling and terminology issues

METADATA QUALITY FROM A BUSINESS PERSPECTIVE

Submitted by Statistical Office of the Republic of Slovenia [1]

Invited paper

I. METADATA QUALITY/THE QUALITY OF OFFICIAL STATISTICS

I.1 Introduction

1. Although the concept of quality has proved to be one of the most powerful production tools in industry in the last 20 years, it is not deployed to the same degree in all businesses. However, even though there is no widely recognised definition of the concept of quality in official statistics, some of the dimensions of quality are agreed upon.

2. The three concepts: contents (meaning); accuracy (precision, reliability); and availability of statistical data are often regarded as three major dimensions of quality of statistical data, and a description of the contents, accuracy and availability of a set of statistical data may be termed a quality declaration of the data. The term "quality" has actually two different meanings. According to one interpretation, "quality" is identical with "good quality". This interpretation assumes that there is general agreement on what "good quality" means in terms of properties of the entity whose quality is being measured. In the case of statistical data it is not easy to establish absolute quality criteria. The most fundamental quality requirement of statistical data is that the quality relevant properties of the statistical data should be known, i.e., the statistics should have known quality, and this quality should be well documented in some kind of quality declaration.

3. Some of the dimensions were already built into the fundamental principles of official statistics. No NSI would ever declare that it systematically violates the fundamental principles of official statistics. However, there seems to be a strong belief on the part of some (power) users (e.g. research institutions, international organizations) that it is their responsibility to verify whether fundamental principles are applied and respected in practice or not.

4. To refresh the memory: **Fundamental principle No. 3:** To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics[2].

5. In terms of so-called metadata (information about the data, i.e. definition of the population covered, definition of the variables, description of the data sources used, description of survey

[1] Prepared by Jozica Klep.

[2] UN/ECE, C(47) The fundamental principles of official statistics in the region of the Economic Commission for Europe.

methodology, etc.), there is general agreement that it is essential for the users of statistics to have as complete a set of metadata as possible. Therefore, national statistical agencies should ensure that full descriptions of the complete methodology for all their collections are documented and kept up-to-date.[3]

6. **Fundamental principle No. 5:** Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies should choose the source with regard to quality, timeliness, costs and the burden on respondents.

7. How effective and efficient is the data throughput in statistical offices, in terms of organization, methodology and technology? It is difficult to address the subject of quality of official statistics without considering the necessary metadata involved, and perhaps even more difficult to address the subject of assessment of quality of metadata without considering what metadata are needed for such a judgement. Therefore, the question implies the consideration of both sides of a single coin at the same time.

I.2 The IMF view

8. IMF has recently begun pushing forward with the establishment of the Dissemination Standards Bulletin Board and with a reference site for data quality. They say that the quality dimensions call for data users to be provided with information with which to assess the quality of the data; to help them verify whether the data meet their requirements, as well as to pressure the move towards international guidelines.

9. Furthermore, IMF has engaged Statistics Sweden, building on that organisation's long experience in the field, to help develop a framework. As the work proceeds, IMF expects to engage a spectrum of data producers and users in a dialogue about this framework and how it might be used.[4]

10. The IMF Data Quality Reference Site provides an introduction to the topic of data quality by referencing contributions in the field. The site also includes a selection of articles and other sources on data quality issues.[5]

11. The Special Data Dissemination Standard (SDDS) identifies best practices in the dissemination of economic and financial data in four areas—the so-called four dimensions: data coverage, periodicity and timeliness; public access to the data; integrity of the data; and data quality. In this context, quality refers to characteristics such as accuracy, adherence to international statistical guidelines, and consistency. The approach taken in the SDDS for the quality dimension is to call for the provision of information that would facilitate data users' assessment of quality according to their own needs. This information consists of methodological statements (covering the analytical framework, concepts and definitions, accounting conventions, the nature of basic data, and compilation practices) and information that permits cross checks for reasonableness.

12. In the wake of the recent financial crises, questions about data quality continue to arise. For example, what assistance can be provided to data users, including those in financial markets, to help them evaluate the quality of the data available to them? In the environment of increased access to data on the Internet, is there a way to focus more attention on data quality issues? How can national statistical authorities be assisted in assessing the quality of their data, and what incentives can be provided to encourage cost-effective improvements?

---

[3] De Vries, W.F.M., Performance indicators for national statistical systems, Netherlands Official Statistics, Volume 13, Spring 1998, pp. 5-13.

[4] UN/Economic and Social Council, report of the IMF on the Special Data Dissemination Standard and the General Data Dissemination System, and issues of data quality, E/CN.3/2000/8.

[5] http://dsbb.imf.org/dqrs_intro.htm

13. The IMF welcomes a dialogue about ways in which both the SDDS and its companion the General Data Dissemination System (GDDS) can be used to increase the understanding of data quality issues and to encourage improvement in data quality. The establishment of this site is part of the ongoing work, initiated to further that dialogue. Its main objective is to foster a common understanding of data quality. Drawing on contributions from the statistical community, the site introduces definitions of data quality, describes trade-offs among aspects of data quality, and gives examples of evaluations of data quality. In parallel, the IMF continues to work on elaborating a framework for assessing the quality of data used for macroeconomic analysis. The aim is to design an integrated and flexible framework in which data quality can be assessed by a broad range of interested users.

14. For decades, national and international statistical offices have made efforts to ensure that quality standards were continuously incorporated in their statistical activities and that users were well informed about the quality of their statistical products. New points of view continue to emerge, but a discernible consensus is being formed around a multidimensional concept of data quality.

I.3 Statistical quality in market perspective [6]

15. The quality related to a statistical product is determined by a number of factors, including product relevance (correspondence between the concept measured and the concept required by the application), timeliness (the period between the time of observations and the time to which the application refers), and accuracy (the deviation between the target size determined by a perfect process and the product size determined by an imperfect process). Wider quality concepts, as used for example by Statistics Canada, also include accessibility, interpretability and coherence.

I.4 The database approach

16. The introduction of a successful database approach for the statistics production system requires well-organized metadata. Metadata have to fulfil a number of functions. The UN ECE' s "Standards for Statistical Metadata on the Internet" identify three main types of metadata:

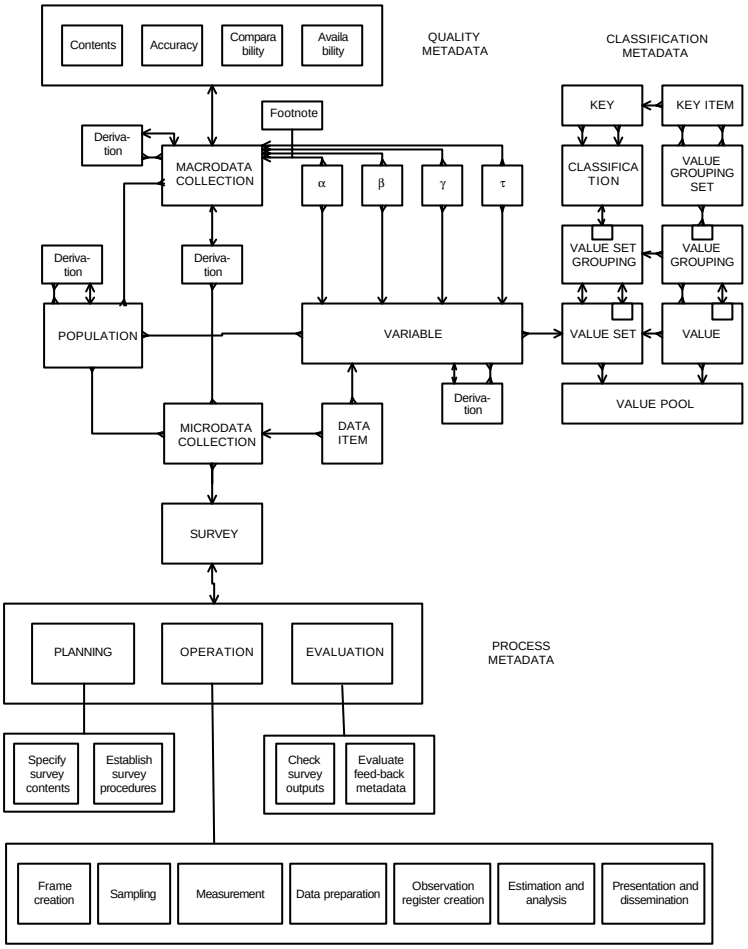
- Metadata designed to help users find the required information (descriptions, which data are available at various levels of aggregation, searches by key words and logical search menus, interfaces between various parts of the information system, etc.);
- Metadata which interpret data (description of the statistical units, units of measurement of the classifications used, the periodicity of data collection and the time-lag between collection and publication, etc.);
- Metadata which describe the quality of the statistical data (information source, the level at which the data were obtained, whether they are accounting data, survey data, etc., the comparability of dynamic series, including a description of the reasons for breaks in series, differences with international standards, etc.).

17. But this does not cover all aspects of metadata. Metadata are used during the whole statistics production process, beginning with the design of an application right through to the dissemination of the products but are still not really used in the computing environment.

---

[6] United Nations, Evaluating efficiency of statistical data editing: General framework, Geneva, 2000.

Figure 1: Metadata model for a statistical data warehouse (Sundgren 1997[7] , pp 32)



18. However, metadata are not only used to access data. They should be used to rationalize the whole production process. The different processes should do both: use the metadata stored in the metadata base and provide new metadata for the system.

19. For a statistical office it is recommended to develop and implement an integrated system of metadata sources around a centralized metadata repository. Within a statistical production system, the same metadata may be used for different purposes. It is often necessary and more efficient to create special metadata architectures for different purposes. A macrodata base may use the same metadata as the tabulation process of a survey, but both systems need metadata in a different way.

20. In most statistical offices, metadata were employed for special purposes, mainly to support database access. Meanwhile, it has been recognized that metadata are useful for nearly all kinds of statistical processes. To avoid redundant collection of metadata, a central repository that is able to feed other "local" metadata systems is one way to be more efficient in metadata management.

21. Another aspect that has to be taken into account is that metadata structures are more complex than the statistical data structures.

---

[7] Sundgren, B., An information systems architecture for national and international statistical organisations, Methodological report, draft, April 1997.

22. The central metadata repository need not necessarily be just one system. It may also consist of a number of subsystems that are linked together to a central metadata repository. In such a context, a documentation system for statistical products (surveys, registers, etc.) would be vital. As an example: Statistics Sweden has developed such a documentation system called SCBDOK.[8] Another dimension can therefore be added to the concept of quality - completeness.

II. ASSESSING THE QUALITY OF THE DATA - ASSESSMENT MONITORING

23. The question of quality assessment seems to be complicated as well. There might be four broad categories involved:

- users
- producers of statistics
- national statistical institutes themselves
- international organizations (IMF, Eurostat, etc).

24. The best way might be some kind of 360° Feedback.[9] It is clear that the assessment criteria may vary according to the needs of different user groups and that they are highly dependent on the level of expertise of different users. Therefore, it could be useful to distinguish between producers of statistics: NSIs as officially responsible for official statistics (on the inside) and between users and official international organisations (on the outside).

II.1 Users

25. "Statistical metadata are descriptive information or documentation about statistical data, i.e. microdata, macrodata or other metadata. Statistical metadata facilitates sharing, querying and understanding of statistical data over the lifetime of the data".[10] The most interesting aspect is the question of common understanding which was already asked: "But is it sufficient to convey meaning to a diverse set of users such that their comprehension of the term is equivalent? Probably not", argued Dippo and Gilmann (1999). Is it possible to convey meaning to a diverse set of users at all? Probably not, according to Wenger (1999).[11] Namely, he says that "the experience of meaning is not produced out of thin air, but neither is it simply a mechanical realization of a routine or a procedure..." He also says that the meaning:

- is located in a process which he calls the negotiation of meaning;
- the negotiation of meaning involves the interaction of two constituent processes, which he calls participation and reification;
- participation and reification form a duality that is fundamental to the human experience of meaning and thus to the nature of practice (Wenger 1999, p.p.52).

---

8 Tacis Task Force on Standardization and Training in Information Technology, Handbook on Information technologies for a national statistical office, Draft version 4.0, 1999-04-03.

9 Edwards M.R., and Ewen A.J., 360° Feedback, The Powerful New Model for Employee Assessment & Performance Improvement, Amacom, 1996.

10 Dippo C., Gillman D., The role of metadata in statistics, UN/ECE METIS, Geneva, 22-24 September 1999.

11 Wenger E., Communities of Practice, Learning, Meaning and Identity, Cambridge University Press 1999.

Figure 2: The duality of participation and reification (Wenger 1999, pp 63)

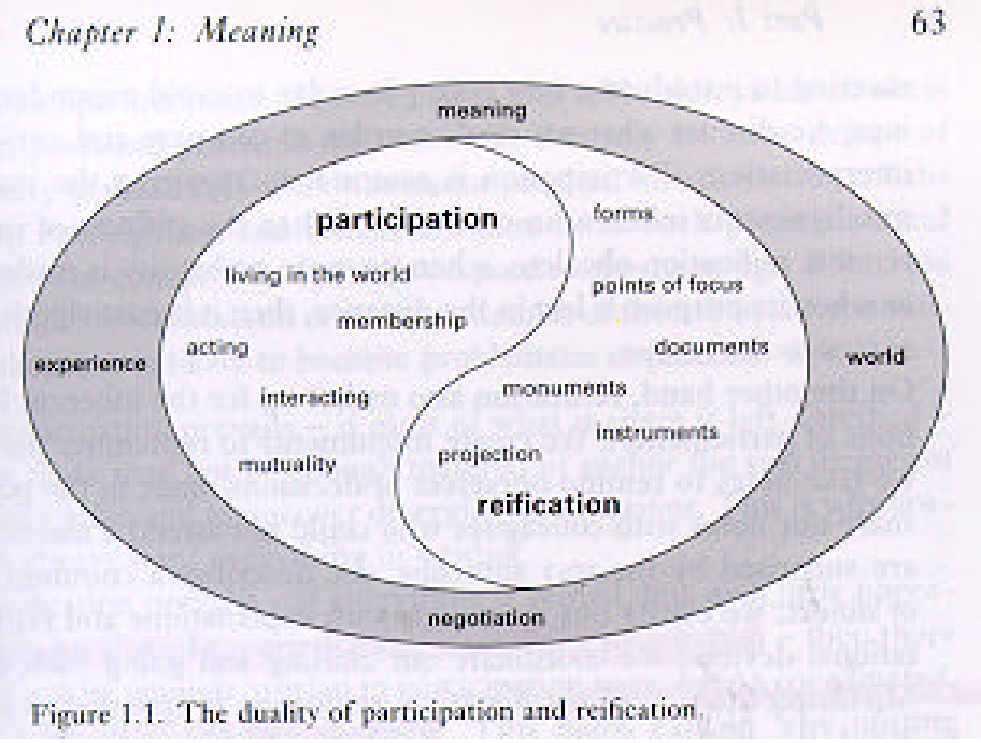


Figure 1.1. The duality of participation and reification.

26. ″Generally, there are discrepancies between what users really want to do and what they actually do. So, with this knowledge in mind, we should start designing systems that are better adapted to the factors that govern the process of information seeking and retrieval. In most cases, today's systems have not been designed to be adaptive, either to tasks or to users. We need to design tools which, among other things, support different information seeking and retrieval processes as well as tools which handle information in different ways. Communication agents, recommendations systems and text extractions are ways that may support some activities, as well as furthering knowledge on how people interact with and utilise information.

27. But most users still fail to find what they are looking for. A recent survey of search engines on the Internet concluded that the best only have a 17% successful hit rate. [12]

28. "An effective knowledge management strategy always proposes the following combination of questions:

- What we know we know;
- What we know we do not know;
- What we do not know we know;
- What we do not know we do not know."

29. Most of the time, companies are aware of what is happening in their sector, but as soon as something unusual floats to the surface, there is a real need to give priority to outside information: by

[12] Kalseth K.: Limiting the Gap Between Information Need and Satisfactory Result, Interview with Preben Hansen, FID Review, Vol. 1, No. 4/5, 1999, pp. 91-94.

closely following the patterns of changes there is a responsibility to continually scan the information in the environment...”

30. The aim of the above citations was to prove that the asseement based solely on user satisfaction surveys does not necessarily reveal an unbiased picture of the quality of official statistics. Some other methods, i.e. user studies, usability testing, cognitive studies, transaction log analysis and HCI might be of interest.

II.2 Producers of statistics

31. In a broad sense, "production" covers the whole life cycle of a statistical survey or a statistical information system, including design, implementation, operation, monitoring, maintenance and evaluation. Producers of statistical data therefore include: designers, input data providers, statisticians... All these categories of producers of statistical data have their typical medatada needs and they might be used as a source in assessment monitoring that is at least as important as the one of (outside) users of statistical products.

II.3 NSI - quality declarations in official statistics

32. The main aim of the list was to propose an instrument for systematic "self-evaluation".

Figure 3: Template for Statistical Quality Declarations[13]

1 CONTENTS 1.1 Statistical characteristics 1.1.1 Objects and population 1.1.2 Variables 1.1.3 Statistical measure 1.1.4 Presentation groups 1.2 Comparability with other statistics	2 TIME 2.1 Reference period 2.2 Production time 2.3 Punctuality 2.4 Periodicity 2.5 Comparability over time
3 RELIABILITY 3.1 Total reliability 3.2 Sources of uncertainty 3.2.1 Coverage 3.2.2 Sample 3.2.3 Measurement 3.2.4 Nonresponse 3.2.5 Processing 3.2.6 Mode assumptions 3.3 Presentation of measures of uncertainty	4 ACCESSIBILITY 4.1 Dissemination forms 4.2 Presentation 4.3 Documentation 4.4 Primary material 4.5 Information

II.4 International organisations

33. There is no doubt that international organisations play the most significant role in building common concepts in global statistical community and are therefore also the crucial moment in fostering worldwide common understanding. Public assessment lists of NSIs, based on previously known criteria, are still rare and argued when elaborated, but even more difficult for NSIs to cope with are "off-the-record" ratings. A combination with "peer review"[14] might be considered as a possibility to assess the strengths and weaknesses of a statistical system (in this case "peers" are qualified representatives of the same branch from another country). Cooperation between the international organisations has very often proved to be fruitful and it would be of tremendous benefit for the statistical community if they were willing to further elaborate the concept of quality in official statistics.

---

13 Statistics Sweden, User Handbook for the Documentation System SCBDOK with the Computer Support PCDOK, 1994-10-17, Translated October 1995.

14 Malaguerra C., Ryten J., Peer review as an essential part of the restructuring of national statistical services – Switzerland's experience, UN, Paris, CES/2000/6.

III. CAN AN APPROACH SUCH AS A FEDERATION OF UNIQUE BUT RELATED MODELS PRESENT A WORKABLE SOLUTION?

34. The answer to this question may be simple - only a set of related models can present a workable solution. Common templates, based on the minimum metadata requirements for statistical aggregates for assessing international comparability of statistical data, and promoting and using standards (ISO 11179) will achieve both goals: to foster common understanding and give the worldwide users the opportunity to become acquainted with, to "negotiate" the meaning of, the concepts applied in official statistics on the one hand, and to give the NSIs the opportunity to deploy sound and comprehensive metainformation systems on the other hand.