STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint UNECE/Eurostat Work Session on Statistical Metadata
(6 - 8 March 2002, Luxembourg)

Working Paper No. 24
English only

Topic (iii): Metadata and quality

METADATA AND QUALITY – THE EXPERIENCE OF THE CZECH STATISTICAL OFFICE

Submitted by Czech Statistical Office[1]

Contributed paper

I. METADATA AND QUALITY – LINKS BETWEEN THE CONCEPTS

I.1 Which are the key links between the concept of metadata and the concept of quality?

1. There are no "good" statistics without metadata describing the data (figures) and their properties.

Statistics	
Metadata	Data

2. Information about data quality forms part of metadata.

Metadata	
Other metadata	Metadata on data quality

3. The quality of metadata is an integral part of the evaluation of the quality of statistics.

Metadata on quality of metadata	
Metadata	
Other metadata	Metadata on data quality

I.2 The criteria for the evaluation of metadata quality

4. According to the definitions of Eurostat on data quality, metadata should respect the criteria listed below.

i) relevance for the users:
- internal users as producers of statistics: producers of statistics need a complete documentation accompanying surveys or statistics (sources of data – questionnaires or administrative sources, data processing specifications, all the information linked to databases, quality reports etc.) – internal metadata;
- external users as "readers" of statistic: there are different groups of external users, the metadata should answer their needs – the solution is to organize metadata in several levels to help the users to find by steps all the information according to their demands – external metadata.

[1] Prepared by Miroslava Brchanova and Hana Slégrová.

ii) accuracy:
- clear description of the reality;
- explanation of the data content, methodology used;
- prevention of misunderstandings in the use of statistical data.

iii) timeliness and punctuality:
- existence of internal metadata before and during the data processing allowing the communication of participants to the process and their mutual understanding;
- dissemination of external metadata together with statistical data (not later).

iv) comparability and coherence:
- maintenance and development of the common vocabulary of definitions of the terms used in statistics respecting such rules as "one term for one definition";
- the development of statistical metadata and their structure as a system and at international level.

v) completeness:
- the content and the structure of metadata to cover the needs of internal and external users

5. Several general rules should be adopted for the proper application of the criteria of quality for statistical metadata and their evaluation, e.g.:

i) complexity: all the above-mentioned criteria should be taken into account;

ii) relativity: variety of the content and the extent of the information according to users;

iii) necessity of priorities: some criteria are "antagonistic", e.g. timeliness and completeness - less information on time is more helpful than none at all;

iv) definition of quality demands ex ante, measurement of the quality ex post:
- 1st step: "decision: "what we need"- the elaboration of concrete measures for specific metadata;
- 2nd step: "what we have, in comparison to the needs" - the evaluation of the results or of the existing status;

v) the co-existence of metadata with data (if applicable), e.g. active metadata explaining changes in methodology in time series.

II. THE USE OF METADATA FOR IMPROVING DATA QUALITY – THE EXPERIENCE OF THE CZECH STATISTICAL OFFICE (CZSO)

II .1 Statistical vocabulary

6. The aim of the vocabulary: the vocabulary represents an important tool for the unification of the terms used by all producers and users of statistics, for the improvement of clarity, coherence and comparability of statistics, and for instructing the public in understanding the main concepts of statistics and statistical data.

7. The content of the vocabulary = terms used in statistics covered by the following chapters:
- statistics (general);
- statistical surveys;
- statistical methods of analysis and prognosis in social and economic statistics;
- presentation of statistical data;
- social and economic statistics (indicators – chapter 3 or the presentation);
- organization of the state statistical service;
- abbreviations, acronyms;

– symbols used in statistics.

8. The idea was inspired by the need to improve the quality of statistics and by the similar vocabulary of Eurostat (CODED). The vocabulary is currently under preparation and will be available in electronic form.

II. 2 Subsystem "Indicators" of Metainformation System (METIS) of CZSO

9. This subsystem has been created from original specifications for programming that were accepted in 1995. The aims of the specifitions were:

– to identify fields of variables in a statistical questionnaire;
– to describe the content of these fields (i.e. name and definition);
– to connect this content with corresponding data in database through a back-to-back SW application "PROJEKTMAN";
– to create time series of indicators;
– to correspond values with their content.

10. Other benefits of the CZSO indicators description system are:
– unification of statistical terminology;
– set-up of an electronic version of „Methodological pages";
– stock-take of detected indicators in order to not duplicate terminology;
– control of questionnaire creation;
– to inform both experts and laymen about size (volume, amount) and content of detected indicators.

The initial main goals to identify data through the description process were abandoned due to technical and organisational reasons. By contrast, the potential benefits have advanced quite well. At the present time other SW applications are either fully functional or better described semi- functional:

11. Intranet already contains a list of valid indicators. You can find methodology (name, definitions.…) and identification attributes here. Advantages: terminology unification, set-up of an electronic version of „Methodological pages", set-up of indicators, control of questionnaires and informing CZSO experts about size (volume, amount) and content of detected indicators.

12. Internet also already contains a list of valid indicators. Here, you can find names and definitions. Advantages: terminology unification, set-up of an electronic version of "Methodological pages" and informing both experts and laymen about size (volume, amount) and content of detected indicators.

13. The dictionary (PODES) is currently under development. The aim of the first stage of this project is to develop a database of the most frequently used terms and definitions used by CZSO staff. The dictionary will be completed and available for use by external users at a later date. PODES will probably be connected with CODED.

14. The main goals of the original project are still available. SW application UPROP enables the connection of the described indicators with the application PROJEKTMAN. An updated application is also currently under preparation. It should also be able to describe CZSO published, derived indicators.

II.3 Metadata in connection with DataBase of Published data (PDB)

15. It is strongly recommended that the PDB data should be closely connected with subsystems of METIS. The PDB is currently under preparation and should use the same classifications and nomenclatures as subsystems of METIS. Our proposal could be illustrated using the following scheme:
Relationship between different levels of CZSO Public Database

level 1)	level 2)	level 3)
DataBase of Indicators (concepts + definitions) *an example of a concept:* *Economically active population*	**Harmonisation** (attributes which conform to EU directives) *English, French and German versions of concepts and English version of definition*	**Interface to public**
Specifications (clasifications + nomenclatures) *some examples of specifications:* *workers* *females on maternity leave* *females on child – care leave* *job applicants registered*		
DataBase of Published data (+ Quality Attribute – both textual and scale evaluation)		

CZSO guarantees the harmonisation of database concepts and definitions with EU directives.

III. CONCLUSIONS

16. There are very close links between the concepts of quality and metadata. There are no useful statistics without metadata. The information about data quality is a part of metadata. The evaluation of metadata quality should respect attributes and rules similar to those for data, while specific measures should be used for assessment of specific metadata quality.

REFERENCES

Eurostat, Definition of quality in statistics, 2000
Hana Šlégrová, Data quality assessment principles and their application in statistics, Statistika 2001