STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 13
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

HUNGARIAN SOLUTION FOR METADATA MODEL

Submitted by Hungarian Central Statistical Office [1]

Contributed paper

I. ABSTRACT

1. The present paper is closely connected to the paper called "Metadata management" presented at the last METIS Work Session (I. Györki, M. Rónai, 1999.)

2. This previous paper introduced the goals of the metadata system of the Hungarian Central Statistical Office (HCSO), where we use metadata in a wide sense. Information characterising the statistical data and information describing control of the design, collection, production and dissemination functions of the statistical system belong to the metadata system.

- The metadata system can be divided into chapters according to the aim of user groups of the metadata and according to the place of metadata in the statistical processing flow;
- There will still be types of metadata which will differ.

"Metadata management" summarises the content of the chapters on the metadata system, the participants and the rules of maintenance for the different chapters.

3. This paper attempts to clarify the various approaches to the formulation of the present metadata model in HCSO. It is not possible to deal with all chapters of the metadata system in the same depth and, therefore, this paper focuses on the metadata topics of interest to external and internal users of statistical information.

II. INTRODUCTION

4. The basic theory of metadata models was given at the beginning of the seventies by B. Sundgren, J. Dörnyei and J. Olenski, and since then a great number of papers have dealt with the topic.

5. To support user inquiries regarding statistical data, it is generally agreed that it is necessary to describe the statistical information available for the users. This information can be characterised by the following properties:

- What is the subject of the observation? What is the so-called statistical indicator or variable (i.e. production value or number of employees, etc.)?

---

[1] Prepared by Ildikó Györki.

- For who and to what scope does the observation extend? What is the observation unit and what is the population of the observation (e.g. enterprises in the industrial branches where the number of employees is greater than four)?
- To what period does the observation refer (year 1999 or January 2000)?
- To what object does the information refer? What is the aggregation level of the given information (given county or county and branch, etc.).

6. If we know the properties of statistical information, why is it so difficult to form an operational metadata structure? Why do the list of stored metadata and the metadata structure vary so in different countries? Why must we amend or modify the metadata structure, which has worked for decades?

7. The reason is that, besides the general concept mentioned above, there are a lot of questions to be answered:

- The metadata do not describe unique items but rather a given set of statistical data. What groups or levels of metadata should be described?
- Statistical data exist in different states within the system according to production phase, aggregation, homogeneity, etc. What states of statistical data should be described?
- The scope and properties of statistical data change from time to time. How should we manage the time-series and the time variants of properties?
- The scope of properties of statistical data could differ according to the users. Which properties should be described and to what extent should they be included in the description of statistical information, e.g. their sources, quality, processing method, etc.?
- The properties can be described using either free text, codes or algorithms. Which method should be chosen for the different properties?
- There are different links between statistical data. How should the metadata support those links?

8. Of course the ideal state would be if each item of statistical data is described with precise metadata, but in the design of the metadata system we have to consider other points of view as well. Some of the important ones are listed below:

- Decisions concerning the questions listed above will define the quantity of metadata. An overly precise description will increase the burden on statisticians. Therefore, any labour-intensive description of statistical data should be proportion to its userfulness;
- The description must be clear to users; the structure of metadata and the way it is described should not discourage users from using them!
- The method of description determine the type of specialists responsible for the maintenance of metadata.

9. To maintain a clear metadata structure, the statistical system has to work with clear statistical terms and standards. Any deviation from these standards will either complicate the metadata system (to modify and manage the different connections) or the system will lose the precision of descriptions. When designing metadata contents and structure, the above requirements and conditions must be taken into account.

III. THE METADATA STRUCTURE

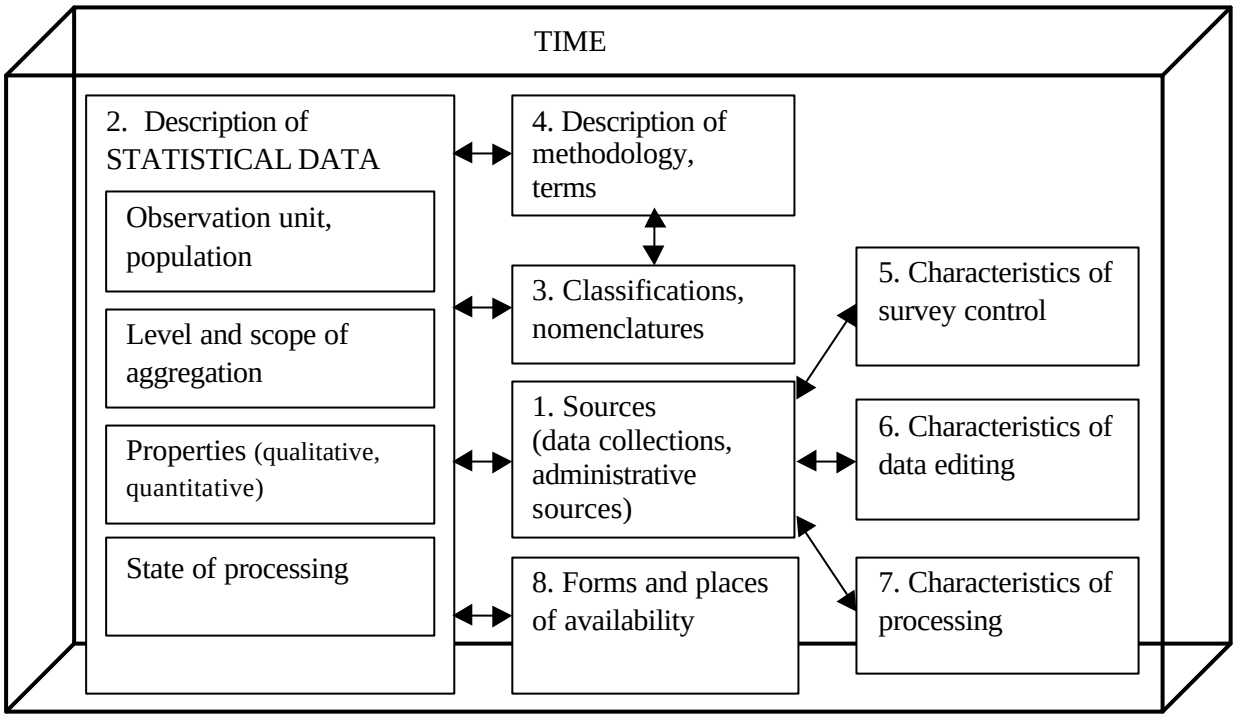

Figure 1: The structure of metadata system

10. The content of the metadata system is derived from the following:

- The metadata system should be expanded to all fields encountered by users of the statistical information system;
- The metadata system has to meet the requirements of statisticians in the planning and production phase;
- The general processes in statistical production require descriptions of the characteristics for different statistics.

11. These requirements form the chapters of metadata system. The starting point is the description of the statistical data collection, questionnaires (1). These descriptions are necessary for all activities of the statistical data collection: they are used by the government to conduct the data collections and data preparation. Data collection via Internet increases the importance of descriptions for the data suppliers; these descriptions appear as sources of statistical information for the users. This chapter encompasses the collection from administrative sources as well.

12. The core of a metadata system is the description of statistical data (2). These can be micro or macro data, obtained directly through data collection or data for dissemination. The description of statistical data must involve

- the statistical unit or population of observation, the source of the information;
- the level of statistical data: this may be the same as for the observation or may be aggregated by certain properties of the statistical unit;
- the observed or computed properties of the statistical unit;
- the characteristics related to the processing;
- information about the relations between statistical data (in time, in common population, in common aggregating level, etc.).

13. The most frequently used part of a metadata system is the description of nomenclatures and classifications (3). This description includes three subsystems:

- the simple nomenclatures, the coded domains of properties (dictionary of codes) (3.1);
- the classifications, registers, where in addition to code interpretation it is necessary to describe other properties, history , etc. for the elements of classification.

This part of the metainformation system is maintained in a special database.

14. The methodology chapter (4) of a metadata system serves two main purposes: (i) the description of statistical concepts and terms aids the users of the statistical system to interpret its content and the statisticians in planning the data collections. This section should be as complete as possible so as to cover all indicators available for the users; (ii) a textual documentation of data collection, according to given regulations, which follows each step of the process.

15. The following three chapters - description of survey control (5), data preparation (6) and processing (7) primarily serve as support for the statistical production process. They contain metadata to assign the population of data collections, print the questionnaires, conduct their mailing and monitor the return. The metadata on data editing describe the connections of data and data collections, editing criteria, types and messages of error, etc. The metadata on processing contain information about the steps involved such as grossing parameters, etc. These chapters contain more in-depth information than is usually necessary for users, and a smaller, selected part is devoted to dissemination for qualified users.

16. To support access to and processing of statistical data, the metadata-base contains information about the characteristics of statistical data base, table and file and the method of access (application software) (8).

17. All information about statistical data, their source, interpretation and processing, change over time and these changes should be managed in the metadata-base, either through descriptions of different versions of information with validity time, or repetition of the descriptions year by year with identical or different content according to the nature of metadata.

18. The chapters are linked to each other. The description of connections enables networking inside the metadata system.

IV. CLASSIFICATIONS AND NOMENCLATURES

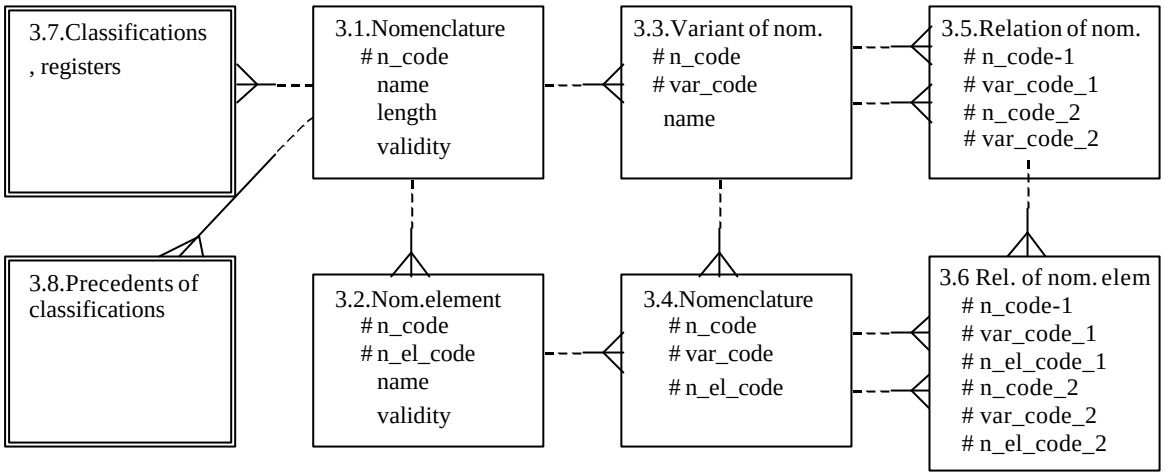

Figure 2: Data model of classifications and nomenclatures

19. All variables in the statistical system possessing a discrete value set (domain) are given an identification, and are written as nomenclatures in the metadata-base. These nomenclatures describe:

- objects of data collection (observation units, data suppliers, etc);
- all properties of objects where the values of properties can be coded (activity, county, sex);
- value sets of qualitative indicators (e.g. type of education);
- variables which aid system operation (methods of data collection, strength of error, etc.).

20. The table of nomenclature (3.1) contains the name, length of codes, validity and the type of value set for the nomenclature. We differentiate between three types of nomenclatures:

- nomenclatures with few elements when we code lists in a common table (3.2) for different nomenclatures;
- the same code lists can be applied to different tasks. These tasks are regarded as different nomenclatures. They receive unique identification, but a primary nomenclature is assigned and the value set is stored only for that primary nomenclature; the others refer to that primary value set;
- the nomenclatures containing thousands of elements characterized by their many properties (not only by name) are identified and described in the table of nomenclature (3.1); the elements and their history are stored in separate subsystems and tables. The table of nomenclature contains only reference to the application. Classifications and registers such as product code systems, country registers, settlement registers, business registers, etc. belong to this group.

21. The table of elements of nomenclature (3.2) contains all elements of nomenclatures belonging to them during their lifetime. Therefore, as well as the element code and name, the validity interval is an important property of elements.

22. The table of variants of nomenclature (3.3) describes the names of the variants. The element of variant (3.4) is a copulative table between the variant (3.3) and its element (3.2). Only the name of the variant can be stored.

23. The third part of the nomenclature subsystem describes the relation between two-two variants of different nomenclatures (3.5, 3.6). These links describe hierarchies of nomenclatures and are frequently used in decoding, aggregation, etc.

V. STATISTICAL DATA

24. The structure of the chapter on statistical data in the metadata base follows the theoretical structure of statistical data. It should describe the observation unit, population, property and time information as parts of statistical data (Figure 3).

25. Indicators are divided into groups (2.1). The indicators in a group belong to the same topic of observation, they have a common observation unit, common population and the periodicity of observation is the same. The groups assist better researching information. They are classified under one or more themes. These themes are the starting points in the metadata application for researching information.

26. Indicators in a group (2.2) can be quantitative or qualitative properties of the observed unit. The description of an indicator contains coded fields rather than free text in order to guarantee a unified understanding of the content. Free text is applied to the name of the indicator without the name of the observation unit, periodicity, reference time or unit of measure, which is coded information. For a qualitative indicator, there is a reference to the nomenclature describing the domain and the value set of the indicator. For quantitative information, the description optionally contains the size and number of decimals for the indicator. Furthermore, it is important to know how the indicator is computed and/or aggregated.

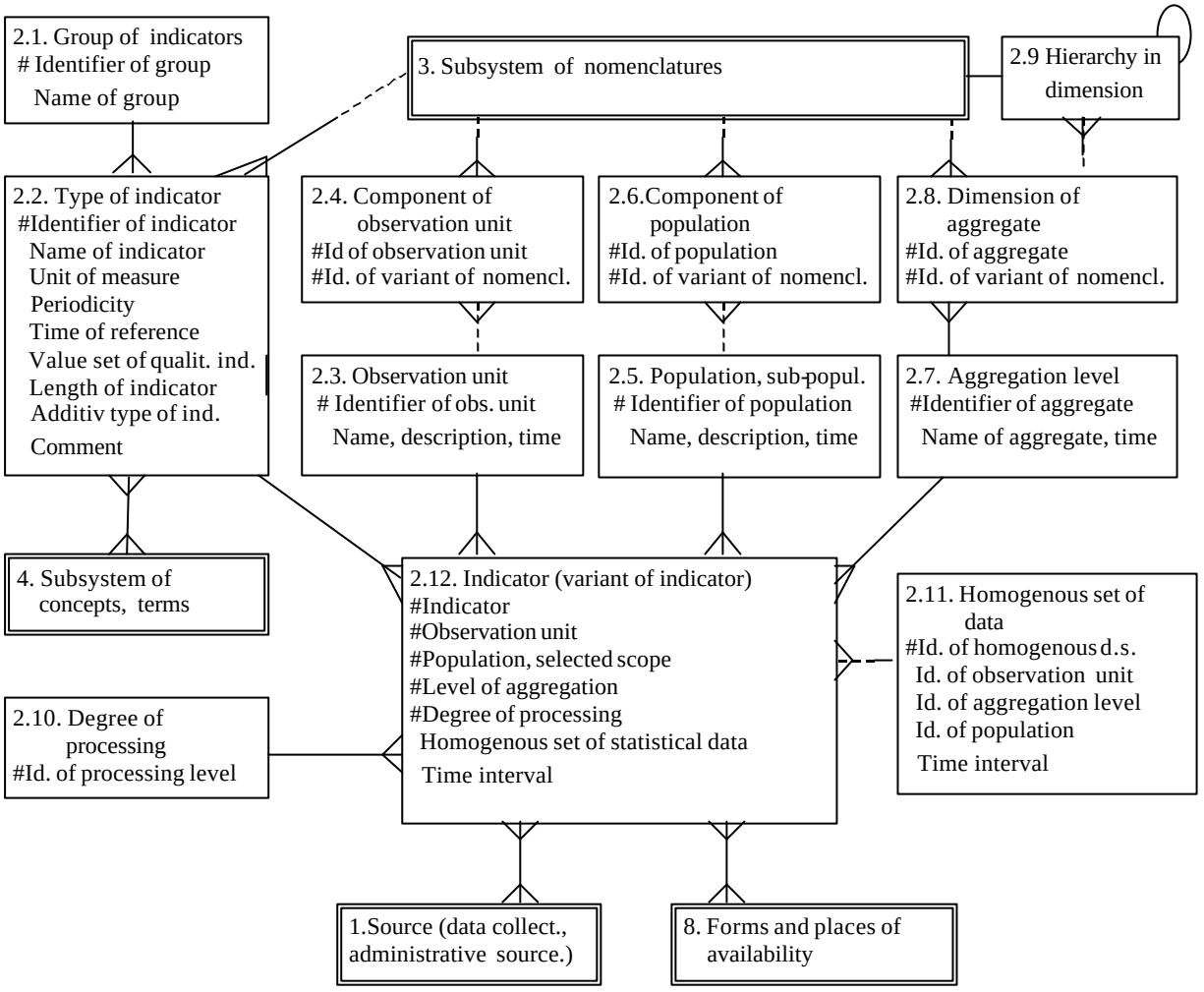

Figure 3. Data model of statistical data

27. The observation unit, population and aggregation level are the most critical aspects of the description. In previous years the system described only the archived micro data. It was therefore sufficient to provide only the observation unit and the observed population, made possible by variants of nomenclatures ordered to the elements of observation unit. For example, the industrial products produced in industry can be classified using two variants of nomenclatures: the enterprises in industry and the industrial subset of the given product classification system.

28. Nowadays, when users have direct access to statistical data, it is necessary to support their inquiries, to describe the derived, aggregated indicators and important subsets of population. So we must separate and differentiate between the descriptions of the observation unit, population, and aggregates. Concerning population, the same indicator (e.g. employee number) can be described in the system both for the whole national economy as total population and for its industrial part as sub-population (2.5, 2.6). This assists easier search since the connection to other industrial indicators can be observed.

29. For similar reasons, the same indicator can be stored in the database and described in the metadata base for different aggregates (2.7, 2.8) such as county and branch or size-categories of enterprises.

30. The necessity to distinguish between aggregate and observation unit (2.3, 2.4) arises as a practical question. It is rare for the same indicators to be observed on different observation unit and aggregated at the same level. But there are examples of this happening: the production value observed

either for economic units or local units and both aggregated for industrial branches. The result is not the same, so both types of statistical data must be described with special observation units and aggregates.

31. The indicators in the system are described in different degrees of processing (2.10.) like “ready for external information”, “for internal use”, or “in the state of data preparation”.

32. All the above-mentioned elements, i.e. type of indicator, observation unit, population, level of aggregation and degree of processing, serve to define the different occurrences, variants of statistical data in the system.

33. The aim of the system is not only to catalogue the available statistical data, but also to support users’ queries. To do this, the system describes homogenous sets of indicators (2.11). These indicators include the statistical data with common observation unit, aggregation level and population or sub-population for the same time interval. As well as being of general interest, the homogenous set of data is the main component of the data warehouse system. The queries refer to indicators only within a homogenous set.

34. There exists further support for queries in the system. Descriptions of observation units, population and aggregate are related to the chapter of nomenclatures. Much effort is being made to identify and describe all types as nomenclatures and variants of nomenclatures in order to ensure their unified appearance. The relation to the nomenclatures provides not only the definition but also:

- the possible aggregate criteria to an observation unit (e.g. county, activity, size-category, etc. to the economic unit);
- the hierarchies within a nomenclature or among nomenclatures (e.g. the levels of activity code, or settlement, county, region, etc.).

35. The interpretation of statistical data is aided by concepts and terms from the methodology chapter (4). The source of indicators is described in the system (1). There is demand for research on both sides: to see the indicators, the results of a data collection and also to see the source of the indicators. The most important external connection is the availability of indicators (8). The metadata base has to show the database, software, table name, column name, physical characteristics and where and how the indicator can be accessed.

VI. METHODOLOGY, CONCEPTS

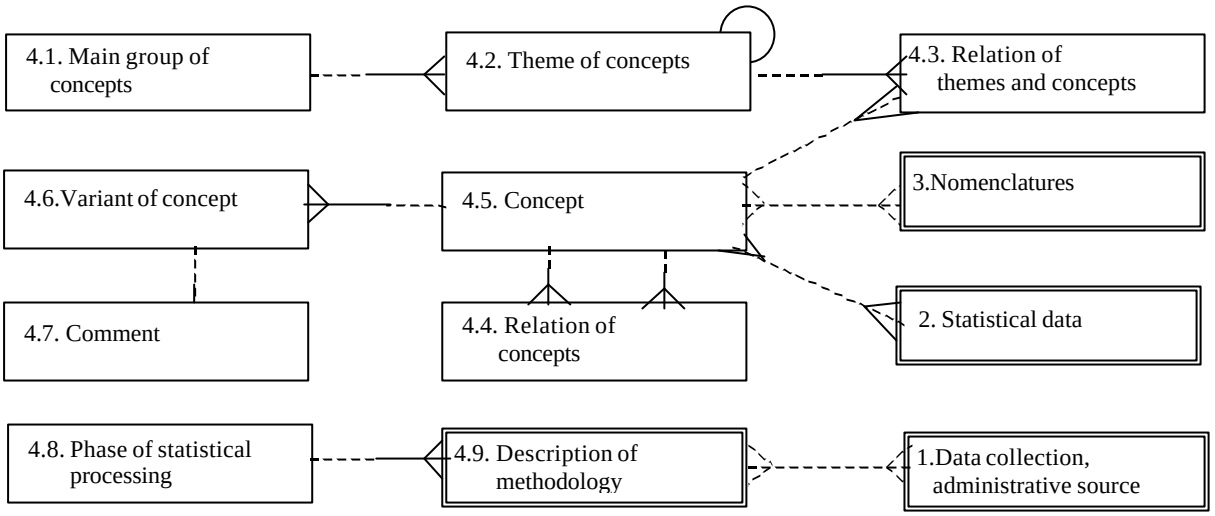

Figure 4. Data model of methodology and concepts

36. The aim of the chapter on methodology and concepts is:

- to explain the statistical data and nomenclatures available in the system (statistical concepts);
- to explain the general concepts of the system (the so-called system or meta-concepts); and
- to document the aim, method and technology of the statistical data collections.

37. The two types of concepts (statistical and system concepts) are managed in the same structure under different themes. The system stores not only the Hungarian but also some international systems of concepts. Therefore, the concepts are collected into main groups (4.1). To facilitate searching, the concepts are ordered into multilevel themes (4.2). One concept may belong to more than one theme (4.3). The interpretation of concept for different time intervals, in different languages, can be found in 4.6, 4.7. The concepts can be related to statistical data, nomenclatures and statistical data collection according to the type of concept.

38. The methodology description is related primarily to data collections. This is a free text, which should follow a proposed layout. The recommendation covers all steps of statistical processing, from planning to dissemination (4.8,4.9).

VII. METADATA SUPPORTING DATA COLLECTION AND PROCESSING

39. The other chapters of metadata system do not directly serve the users of statistical data. So far, we have looked at the structured description of data collections, survey control of data collections, and data editing of data collections as the documentation and control data of the given steps of statistical processing. But in the near future, there will be increasing demand to support, control and check the data suppliers using direct completion of the questionnaire on Internet. These systems will emphasise the role of metadata, which will in consequence become much more important than has been the case previously.