Metadata and Metainformation - Old Concepts and New Challenges

*by Dusan Soltes**

Introduction

Since the very beginning of computerized data processing there has been a tendency for ever growing amounts of data to be processed and stored by computers. Probably not by accident, the modern computerized data processing was also referred to as mass-data processing. Especially in the environment of the so-called large-scale information systems, as e.g. statistical ones, there was an ever growing necessity to find the ways and means how to handle these rapidly expanding amounts of statistical data. The technological advancement and users needs finally led not only to introduction of very-large data bases and their distribution to the data base networks but also to the necessity to invent and introduce the particular tools for handling especially their content, i.e. data and information in the form of data and information on another - source and/or object data and information, which started to be referred to as metadata and metainformation.

Brief History of Such "Old" Concepts as Metadata, Metainformation and Metainformation System

Since their introduction in the 1970s, metadata, metainformation and metainformation systems have been the object of systematic research and development at international as well as national levels: first as part of a cooperative network program of European statistical offices and later on in 1981-84 as an object of the joint work of an inter-country joint group of experts of national statistical offices under the Statistical Computing Project of the United Nations Economic Commission for Europe. The main results of this international joint work, in the form of a pilot Users Guide to Metainformation Systems in Statistical Offices (1) and Selected Chapters for Designing METIS in Statistical Offices, defined the following basic concepts: metadata as a physical representation of metainformation, metainformation as a semantical contents of metadata while metadata is a description of (statistical) data and metainformation informs about (statistical) information. The data and information as objects of "description" by metadata and metainformation are, for terminological clarity, also referred to in this paper as object data. Metadata and metainformation at the same time represent a content of metainformation system (METIS) in the form of its metadata base. Basic relations between these fundamental concepts of metainformation were defined for the first time by B. Sundgren of the Central Statistical Office of Sweden. His work became a basic theoretical framework of the particular international joint work.

If we also try to define the METIS itself then we come to conclusion that it is again a specific information system which as an object of its representation has an another i.e. object information system , e.g. in this particular case a statistical information system. Then on a more detailed level as the objects of METIS we can define various other individual components of the particular information system, not only its data and information but also:

- statistical surveys
- statistical forms and/or questionnaires
- statistical populations and/or files of statistical units
- statistical indicators
- data files/data bases
- publication tables and publications themselves
- programs, methods and procedures

All these objects are needed one way or another for any proper handling of the content of the particular object information system i.e. its data and information. If any user of these data or information wants to properly use, analyse, interpret them he/she always needs to know not only the quantitative values of the particular data or information but needs also to know many other accompanying information on this data or information i.e. metainformation. Without knowing which statistical survey it has produced these data and when and how, and how the objects of a survey i.e. statistical units and their populations have been defined, etc, it is almost impossible to utilize them properly or even to utilize them at all.

It is quite evident that not only for the convenience of users, but for any systematic handling of these specific accompanying data i.e. metadata they have to be organized as any other data into some organizational units - records, files, databases, etc. In general all these forms of metadata organization and storage are identified as a metadata base. The metadata base itself is organized as a system of mutually related individual metadata files or holdings

which can have different forms. The basic forms of metadata files and/or holdings in a metadata base are e.g. as follows:

- catalogues
- dictionnaries
- directories
- registers
- glossaries
- thesauri.

Some other authors include into the content of a metadata base also some other rather specific „metadata" and their files and/or holdings as e.g:

- classifications
- nomenclatures
- code-lists.

The inclusion of these specific data - also sometimes referred to as service or auxiliary data - into the metadata base is based on the fact that their primary function is not to „describe" objects of the real world as (statistical) data do but to assist in more precise specifications of object data/ information themselves. The content of the metadata base itself is created by formalized descriptions of particular objects of formalized descriptions such as e.g. (statistical) data i.e. indicators, but also surveys, units, files/ populations etc. The most important and at the same time also the most voluminous part of any metadata base is its part regarding the core part of the data component of the particular object information system i.e. its operational data which in the case of statistical information systems are represented by (statistical) indicators. In this case the formalized descriptions contain descriptions of such attributes which help to properly interpret quantitative values of statistical information as e.g.:

- indicator name
- type of indicator
- code of time characteristics
- periodicity of collection
- measurement unit
- origin
- semantics and/or definition
- cross-sectional classification characteristics
- acronym and/or code (identifier) of indicator
- etc. depending upon the specifics of the particular object information system.

On the basis of its metadata base and its contents of formalized descriptions, METIS is then able to fulfill and/ or at least assist or support such various functions regarding the object data and information as e.g.:

- information
- identification

- interpretation
- navigation
- localization
- retrieval
- etc. regarding possibly also some other functions depending upon the objectives of the particular information system and/or its metainformation system.

On the basis of that METIS then can serve in several possible operational and functional modes e.g. as a SILS i.e. a simple information/interpretation and location system which assists users in proper interpretation and localization of object data/information without any further functions towards accessing them directly. This function is approximately on the level of catalogue systems in libraries which inform users about the books and their basic characteristics and location but without possibility of any direct retrieval. The higher function of METIS is its function DIRS - detailed information and retrieval system which in combination with the particular database and retrieval system enables also a direct access to the object data.

Challenges of Contemporary Information Highways vis-a-vis Metadata, Metainformation and Metis
If we compare the above concepts, elements and functions of metadata, metainformation, metadata base, METIS with the challenges of the contemporary world wide web and in general with information sources on contemporary „information highways" we may see their almost absolute inevitability, relevance and direct utilization especially in a case when users worldwide have access to practically unlimited sources of various data. Under such conditions it is sometimes almost impossible to secure any kind of proper information, identification, interpretation, comparability, consistancy, etc. between these data coming from very different methodical environments if there is not at the same time available also some kind of accompanying metadata and/or metainformation. If we take as an example statistical data on education and only on the level of basic and/or elementary schools we have almost unlimited possibilities for various interpretations of this relatively easy and commonly very well known concept in case that the data are coming from different data/surveys sources. They are as follows:

- The first problem is with the proper interpretation what it is an elementary and/or basic school or education. In different parts of the world it varies from 5 years e.g. in many developing countries up to 8, 9, or even 10 years of by the law obligatory „basic" education.

- The second problem is with the structure of this kind of education. In some countries it is a continuous education through the above 5 to 10 years. But in some

other cases it has been divided into two subsequent levels as e.g. 1 through 4 and then 5 through 8 or 9, 10, etc. But in some other cases it is organized in several parallel options as e.g. 1 through 4 and then 5 through 9 but there is also an alternative that enables for the best pupils after completing 1 through 4 grades to continue in an uninterrupted education from 5 through 12 grades what is already automatically combined basic education with the 8 year high school education completed by leaving examinations. But there still exist also a possibility to complete first 8 or 9 year basic or elementary school and then to proceed to a four year high school.

- The third problem is with the interpretation of the age. In some countries the basic education starts at the age of 5 but in many other countries it is at 6 and there are also countries where this age has been defined as 7 years.

Even this simple example demonstrates that not having the proper meatadata to identify and interpret the data or information on elementary education could provide us with results covering a variety of populations including: 7 through 12, 5 through 17 or even 7 through 19 years of age. These differences or variations in expression of data are too great for some analyses and international comparisons. Therefore the existence of accompanying metadata and/or metainformation in the conditions of a direct access to these data by any user in the world in the conditions of an information highway becomes an absolutely inevitable and objective necessity.

Unfortunately, this is not always the case at present. The data and/or information are available, but their interpretation in many cases is left up to the users. Hence, analogically as in the case of large statistical information systems in the past also in the current www, it is possible to expect that one of their further main development trends will be towards some kind of normalization, standardization, unification and finally towards legally required necessity for providing some accompanying metadata and metainformation. Sooner or later we can expect that in addition to the existing information highway there will be necessity to have either its parallel metainformation (sub)highway and/or what seems to be more practical and less technically demanding, that all data and information will have to have a some kind of accompanying metadata sector which will contain basic (meta-)information describing all particular data/ information. As we have already demonstrated also in this paper, the basic methodology for such accompanying metadata and metainformation has already been at disposal for a while, the new and still newer metadata handling systems are becoming ever more common. What is „just"

missing for the time being and what is at the same time also the main challenge in this field for the future it is to find the ways and means how to standardize all this metadata, metainformation, METIS concepts and mainly tools into a world wide accepted standards. If it is too much and/or too demanding we will have to see.

Reference:

1) User Guide to Metainformation Systems in Statistical Offices, ECE/UNDP/SCP/H.4, United Nations Development Programme - Economic Commission for Europe - Statisitical Computing Project, Geneva 1984

Dusan Soltes is a Senior Lecturer for MIS at the Faculty of Management of the Comenius University of Bratislava. Before joining the faculty he was working as a researcher in the area of computerized statistical information systems not only in his home country but also under the cooperative and development programmes of the United Nations – Economic Commission for Europe and the Conference of European Statisticians. Among others, in 1981-84 he has led a particular international joint group of experts for METIS under the ECE/UNDP Statistical Computing Project which has produced two basic publications viz. Users Guide to Metainformation Systems in Statistical Offices and Selected Chapters for Designing METIS in Statisitcal Offices.