Modeling of Statistical Data and Metadata

Silviu Florin TEODORU
Bucharest, România
silviu.teodoru@oracle.com

An increased need for the exchange of statistical information among users, as well as the recent technological development in communications and micro-computing, calls more and more for the effective design, development and implementation of statistical meta information systems. To design such systems correctly requires the harmonization of technological, subject matter oriented and organizational aspects of statistical information systems.

Keywords*: knowledge, statistical, metadata, modeling, process.*

1 Conceptual foundation

Statistical metadata are data which are needed for proper production and usage of statistical data. They describe statistical data and - to some extent – processes and tools involved in the production and usage of statistical data. Expressed briefly, statistical metadata are data about statistical data.

A statistical meta information system is a system, which uses and produces statistical metadata, informing about statistical data, and which fulfils its tasks by means of functions like "statistical metadata collection", "statistical metadata processing", "statistical metadata storage", and "statistical metadata dissemination".

Statistical data are the primary objects of the descriptions provided by statistical metadata. Thus in order to understand the meaning and contents of statistical metadata, we must have some understanding of what statistical data are, and what it is about them that may have to be described.

Statistical data may be microdata (observation data) or macrodata (estimations result).

Microdata, sometimes called observation data or measurement data are the result of observations or measurements of a set of object characteristics (states and events).

Macrodata, in daily talk simply referred to as "statistics", are the result of estimations of a set of statistical characteristics (statistical concepts). The estimations are made on the basis of a set of microdata, that is, a set of observations of a set of object characteristics.

2. Statistical Information Systems

Beside the statistical data themselves, the **processes** of the **information systems** associated with the statistical data are important description objects of statistical metadata.

A **statistical information system**:

- ✓ provides statistical information, that is, information about collectives of objects (rather than individual objects) in the object system;
- ✓ supports so-called directive actions, like general level planning, decision-making, and evaluation.

A statistical information system accomplishes its tasks by performing three major functions:

- ✓ an **input acquisition function**, which directly and/or indirectly observes (measures) certain object system characteristics, and which prepares and stores the observation data obtained as microdata in an observation register;
- ✓ an **aggregation function**, which transforms the microdata produced by the input acquisition function into macrodata, or "statistics", which are estimated values of statistical characteristics;
- ✓ an **output delivery function**, which makes macrodata (statistics) available to the users, and which assists the users to interpret and analyze the data further.

Modern technology permits a much more flexible organization of the processes for producing and disseminating statistics. The statistical system assumed to be:

- ✓ **database-oriented**: the microdata and macrodata, which are stored and processed, are communicated within and between the functions and subfunctions via a database;

✓ **self-describing**: the microdata and macrodata are described by means of accompanying **metadata**, which are stored in the database, and which are consistently transformed, whenever the described data are transformed.

The **statistical information system architecture** covers many different types of statistical information systems: survey processing systems, register management systems, user-driven retrieval systems.

A survey processing system focuses on a data collection process, resulting in a collection of microdata, which are aggregated into estimated values of certain statistical characteristics.

A user-driven retrieval system focuses on the needs of a particular category of statistics users, and aims at making available macrodata and microdata from different surveys (and other sources), which may be relevant for the particular category of users.

Register management is an important auxiliary process for statistics production. There are two kinds of registers, which are particularly important for statistical information systems: base registers and code registers. A **base register** establishes and maintains an authorized list of the objects belonging to a certain population. A **code register** establishes and maintains an authorized list of the values belonging to the value set of a certain variable or classification.

3. Architecture of Meta Information Systems

The desirable architecture of future meta information systems should consider:

✓ Metadata collection activities should be minimized in the sense that no metadata should have to be entered more than once, and derivable metadata should be automatically derived rather than manually entered.

✓ Huge retrospective metadata collection activities should be avoided. Instead as much as possible of be avoided. Instead as much as possible of the metadata input flow should be generated as a side-effect of other activities. For example, the more or less formalized descriptions that are typically generated by systems analysis and design activities should be automatically captured and organized as potential metadata for the information system under development;

✓ Some type of cost/benefit mechanism needs to be introduced into the architecture of a meta information system in order to relate users and producers of metadata in a constructive way. The mechanism needs to be relatively sophisticated, since there is a many-to-many relation between users and producers of metadata: the same metadata may be used by many different users, and the same user may need metadata concerning many several data collections from several producers.

Statistical information systems are valuable assets. However, without properly integrated meta information systems, the value of the information systems is drastically reduced. Since today's statistical information systems are by and large formalized and computerized, the meta information systems must also be formalized and automated, if the pace of the metadata flow is to keep up with the pace of the object data flow.

3.1. Metadata holdings

The statistical metadata managed by a statistical service need to be stored in **metadata holdings**, organized into one or more **statistical metadatabases**. A statistical metadatabase may be **active** (integrated with a statistical database) or passive (separate from a statistical database). It may contain local metadata concerning individual surveys and/or global metadata concerning a wide range of surveys and data collections. It may be physically stored and maintained locally or centrally in the organization.

There is a need to store a lot of different kinds of metadata in a statistical meta information system. The metadata can be categorized in several different dimensions, for example:

✓ by metaobject type;

✓ by being microdata-oriented or macrodata-oriented;

✓ by data type (quantitative, qualitative, textual);

✓ by type of formalism (fixed-format facts, logical expressions, mathematical expressions, algorithms, graphs, free text);

✓ by being data-oriented or process-oriented;

✓ by being procedural or declarative;

✓ by representing specific facts or general knowledge.

Since the metadata will be in many different forms, relatively advanced **database** management software will be needed for handling metadata holdings properly.

In many respects a statistical metadatabase can be designed in the same way as any other kind of database. For example, it is advisable to use a so-called object graph (or Entity Relationship graph) for modeling the contents of a statistical metadatabase. Such a **metaobject graph** contains metaobjects, metavariables, value set of one or more variables etc.

A three-layer model is one way of taking care of the user needs for comparability in time and space. The type layer should contain metadata, which are

"usually" the same, or at least "similar" for different members of the same type.

The type level metadata have the character of "general rules" or "typical descriptions"; exceptions to the rules can be given for subtypes and/or occurrences of the types.

Analogously, **the series layer** should contain metadata, which are "more or less" the same for different repetitions within a time series. Once again exceptions to

the typical descriptions can be given on the occurrence level.

The occurrence layer should primarily contain metadata, which are known to be different between different occurrences within the same series, or the same type, respectively. High variability in this sense is typical for most operation-based metavariables, like "measurement problems" and "non-response rate". Design-based

metavariables will not change their values between repetitions of "the same" survey to the same extent.

To summarize, many metavariables will have to be recorded on the occurrence level. However, if a metavariable is known to be relatively stable over time, it could be recorded on the series level, provided that there is an option to record **occurrence level exceptions** from the **series level rule**. The exceptions could result in **footnotes** in appropriate places, when the data are presented.

3.2. Metadata flows

Every statistical service function, which somehow manages data, should also manage the metadata, which is associated with the data.

In fact automation and computerization of survey management has up to recently implied disintegration of the natural relationships between statistical data and metadata, which existed in earlier manual systems. For example, in a questionnaire, when it has been completed, it contains both data (answers to questions) and the associated metadata (the questions themselves and accompanying instructions for answering the questions).

An essential feature of modern metadata management is that it is reintegrated with object data management, so that for example the metadata describing the figures in presented tables would in fact be the result of a chain of systematical, formally well-defined, and automated transformation processes, starting with the metadata in the questionnaire, or maybe even earlier, with the metadata generated by design decisions preceding the (computer aided) construction of the questionnaire.

During all activities of all phases of the life-cycle of a statistical system, the different actors produce decisions, documents, etc, which contain metadata. If the metadata are properly captured and organized, they may become very useful, when the same statistical information system, or other ones, require metadata input.

It should be a challenge for every statistical service to organize its metadata flows in such a way that:

✓ as many metadata as possible can be obtained from existing metadata holdings, whenever they are needed by a certain actor in a certain statistical system;

✓ as few metadata as possible have to be produced for its own sake, rather than as a side-effect of other (necessary) activities of the statistical systems monitored by the statistical office.

Sharing of metadata (as well as sharing of object data) within **and between systems** is essential for any statistical service aiming at rational, computer-supported planning and operation of its statistics production. Systematical, automated exchange of metadata between different activities promotes two good causes at the same time:

- ✓ it decreases the burden on metadata providers;
- ✓ it increases the benefits gained from metadata, which are already available.

International standards for the storage and exchange of statistical data and metadata should significantly facilitate the efforts of statistical services to systematize and automate exchange of data and metadata, both internally and externally.

4. Benefits

There are important interdependencies in the metadata flows between

- ✓ local and global metadata holdings
- ✓ different statistical information systems
- ✓ different phases of the life cycle of each statistical information system.

There can be an important feedback loop from the local metadatabases of a number of (different types of) statistical information systems to a common global metadatabase. The local databases contain detailed knowledge concerning specific systems and their data (observation registers and statistics). The local metadata have to be processed (as automatically as possible) in order to create extracts and summaries, which can be managed in the global metadatabase, from which metadata can be retrieved by local as well as global and external systems. In order to make the retrieval of global metadata as efficient and user-friendly as possible, the extracts and summaries have to be further processed (once again as automatically as possible) in order to create and maintain reference data like tables of contents and indexes.

The generation of extracts, summaries, and reference data is one part of the knowledge acquisition process for the global metadatabase. Another part is the acquisition of general knowledge: handbooks, encyclopedia, thesauri, standards, software, etc. The acquisition of general knowledge can be performed rather independently of the feedback loop just described. However, there is a potential for creating an "intelligent" inductive learning loop from the local and global specific knowledge to the global general knowledge. At the present state of the art this inductive learning loop will be highly dependent on human efforts, but artificial intelligence may contribute increasingly in the future.

Bibliography

- www.unece.org/stats/publications/metadata modeling.pdf - “Guidelines for the Modeling of Statistical Data and Metadata”
- http://www.insse.ro/cms/rw/pages/index.ro.do - “On-line Monthly Statistics”
- www.unece.org/stats/documents/ece/ces/ge.40/2006/wp.9.e.ppt; “Conceptual Modeling of Statistical Metadata and Metadata Data Model”
- www2.computer.org/portal/web/csdl/doi/10.1109/SSDM.1996.506059 - “Conceptual data model with structured objects for statistical databases”