

A structured approach to a statistical metadata base.

Michalis Petrakos, Gregory E. Farmakis, George Petrakos, Panagiota Bisiela, George Koumanakos

LIAISON SYSTEMS S.A., Academias 77, Athens 10677GR,

e-mail: {FirstName.LastName}@liaison.gr.

Abstract: During much of the last decade statistical metadata have attracted a lot of attention in the Official Statistics literature and practice. There is now a battery of guidelines and standards, a set of established and operating statistical metadatabases in National Statistical Institutes (NSI) as well as ongoing efforts to set common standards and practises at the European and International level. Adding to this discussion this paper examines the possibility of modelling statistical metainformation using structured metadata, that is, metadata that correspond to measurements (at the nominal or higher levels of measurement) that can be exploited algorithmically rather than free text. The advantages are twofold, firstly, structured metadata accompanying statistical tables from different NSI's can be compared and their level of harmonisation assessed automatically. Secondly, metadata can be used to asses costs based on the response burden and the time it takes for several stages of the statistical process to complete, identify bottlenecks and inefficiencies and, therefore, facilitate managerial decisions. The metadata model described in this paper has been designed and is being implemented in the framework of the Integrated Information System for the National Statistical Service of Greece (NSSG).

Keywords: Statistical Metadata, Structured Metadata, Data Quality, Active Metadata

1. Introduction

Metadata have spilled over from Information Technology to a number of disciplines with great impetus in the past ten years. In statistics, there has always been more respect to numbers coming out of well-documented surveys rather than simple numbers. Statistical metadata of some form has always been available to users of statistics. This availability, however, has been uneven and incidental. As a result, statistics were allowed to be misused, sometimes in outrageous cases of factual abuse. Metadata are one hope of the statistical profession to avoid the slander that statistics are even worse than “damned lies”. A set of data accompanied by appropriate metadata although will probably not stop a determined data abuser, will, nevertheless, leave less room for manoeuvre. Any knowledgeable person will be able to question inferences based on wrong population assumptions if is aware of the statistical population and the master list of the survey.

There has been a great multidisciplinary effort dedicated to the modelling of statistical metadata which has led to several models for storing and processing metadata. There are now in place several approaches to metadata modelling, most of them coming out of projects funded under the EU's 4th and the 5th framework programs for research technology and development (RTD); existing metadata bases are also in place in National Statistical Institutes' (NSI) Information Systems.

Surprisingly enough, what constitutes appropriate statistical metadata is not defined in a uniform way among organisations that are involved in the production of official statistics. Different organisations, with diverse priorities, put emphasis on different types of metadata. One well-known example is that of “integrity assessment” which is high among the priorities of the IMF's Standards, but non-existent as a primary quality dimension in Eurostat's approach.

The metadata model we present here was developed for the needs of the Greek National Statistical Service (NSSG) and is currently in the implementation phase covering all surveys conducted by NSSG. The meta-database is designed as part of the architecture of the Integrated Information System of NSSG, which is a major re-engineering process currently under way.

1.1 Metadata Users

The needs of users of statistical data are one important pointer of appropriateness. We have identified three types of users of statistical metadata:

- ε Data providers, people involved in survey design and implementation, data collection, compilation and dissemination.
- ε Data users, including policy makers, the media, academics etc.
- ε Secondary data providers, organisations that receive statistical data and then recompile and disseminate them.

Data providers use metadata to design, implement, monitor and evaluate surveys as well as disseminate results. They use the available information on units, concepts, populations etc to design a survey, they follow the process of data collection and preparation by retrieving relevant metadata, and they finally disseminate the data, estimates etc. as a final

product accompanied with metadata

From the data users' perspective, metadata helps them identify the data set they actually want, and assess its quality and reliability[9]. There is a growing awareness about the quality of statistical data. This has led to better delineation of the quality components and how specific methodological aspects affect them. Additionally, there are now better ways to assess the accuracy of the data since research in non-sampling errors has made headway [4][5].

Some users collect similar data from different sources and recompile and disseminate the combined data sets. These users of multisource data include international organisations like the EU, IMF, OECD, UN, WTO, and others. They need to be able to determine if data coming from different countries are comparable or not. These organisations have increased their influence in what data an NSI produces and, in some cases, how it does so. This is generally true in the era of globalisation. Particularly in Europe, the *acquis communautaire* covers a large part of the data NSIs produce (and it is being extended to cover in more detail, more domains, in more countries). These secondary providers have setup a number of metadata standards and templates that even if they have not, so far, provided a universal standard they have pretty much outlined what should not be omitted and what should be elaborated upon [6],[7],[12],[13],[14].

2. What metadata to model

2.1 Types of metadata

There are several ways to categorise statistical metadata. For our purposes we use two types of distinction. The first is active and passive metadata.

- (i) Active metadata are the ones physically integrated with the information system containing the data that the metadata informs about. They are entered and updated without human intervention.
- (ii) Passive metadata refer to data but are not connected with them and are entered separately typically using forms.

The second distinction is in three layers,

1. metadata related to a variable including multidimensional variables, indicators, characteristics (definition, classification etc).
2. metadata related to the survey (population, frame, data collection etc)
3. metadata related to the organisation and are common in all surveys (general statistical legislation, policy documents etc.)

For layer-1 metadata we decided to follow the ISO/IEC 11179 standard for obvious reasons of standardisation in interoperability. This paper refers mostly to survey metadata (layer-2).

2.2 User needs

The approach taken is to include every metadata item useful for users of statistical data, so

as to cover the needs of international organisations. Data quality reporting is the most prominent of user needs. Quality is usually defined in a number (varying depending on the implementing agency) of frequently overlapping components. According to the Eurostat approach these are Accuracy, Comparability, Coherence, Relevance, Timeliness, Accessibility and clarity, and Completeness [2]. Hence, metadata useful for reporting in anyone of these components were included.

The needs of international organisations are also well documented and were included in the metadata base design. The OECD collects and disseminates large amounts of data and therefore, has an interest in metadata for both the use of individual data sets and for comparisons between them [6]. In 1995, it adopted a standard and comprehensive list of metadata items [7] covering the whole statistical production cycle for primary statistics. The main purpose of the list was to help achieve some consistency in the metadata collected by different parts of the organisation from national sources. This standard list is being used to collect detailed metadata from national sources and is completely covered by the developed metadata model. The IMF's SDDS is another metadata standard that is well established and is actually gaining ground. In 1999, Eurostat's Committee of Directors decided to adopt it for Euro-SICS indicators covering a number of statistical domains [9].

Another category of metadata is the one useful to the Statistical Institute for decision-making inside the organisation. This includes methodologies, systems used, as well as resources and associated costs, both internal and external, like the response burden.

One problem with current approaches to metadata modelling is the large size and enormous complexity of most of them. This makes their maintenance and use difficult and costly, and it sometimes results in not keeping metadata up-to-date. We, therefore, opted for a scalable model that still covers the user requirements for assessing the "fitness for use" of statistical data.

An additional guiding principle was to have as little human intervention in the metadata entry and updating process as possible and let the information system update most metadata during the various phases of the survey design, implementation and compilation of results. This enhances the quality of metadata by keeping it in phase with data and, thus, eliminate data-metadata mismatches that can become a serious quality issue. [10]

2.3 The need for structured metadata

Metadata is being available at a global scale. Its usefulness, however, is restricted by the fact that most of it is stored as free text, i.e. the computer knows what it is about but not what exactly it is.

One problem that is apparent when one tries to compare national practises is that template-based meta-information has to be read, evaluated and compared by a human expert every time one wants to make a comparison between indicators coming from different countries. For the 29 member countries of the OECD this means 406 comparisons in a number of metadata items, and, for the 49 subscribers to the IMF's Special Data Dissemination Standard (SDDS), an overwhelming 1176 comparisons. This looks like a quite costly

operation and might lead to little use of the appropriate metadata. It is, therefore, equally important to be able to assign numerical values or at least codify metadata so they can be algorithmically exploited.

This has led us to try to model as many metadata items as possible in at least the nominal level of measurement. Some items like time periods and sample size are numeric in nature, some others, like data collection and sampling methods, can be codified in a straightforward manner. In other cases, the item can be broken down in several parts that can be codified on. For example, Seasonal Adjustment can be broken down into whether data are seasonally adjusted and by which method. That still leaves a lot of metadata items that cannot be codified; however parts of them can. A report on breaks in time series should contain some text explaining the circumstances that created the break; nevertheless, the time the break occurred, the cause of the break (based on a list of causes of breaks) and its severity, can all be structured and they contain most of the valuable information. Definitions are probably the most difficult case to deal with, they are meant to be textual and human-not computer-understandable. Be that as it may, there can be some success in many cases: the definition of unemployment is based on whether someone is available to work in a certain amount of time, and, has been actively looking for a job for another amount of time and therefore the definition is based on two time parameters. However, this was beyond the scope of our work and the only structured information on definitions was whether they deviate from international standards.

3. The metadata model

A fundamental principle, developed over many years of relational database technology, is that metadata, being data themselves, should be treated, stored and managed in exactly the same way as ordinary data. For instance, active metadata of any relational database (i.e. information on tables, columns and references) are kept into ordinary tables within the same database schema, which can be queried upon, joined with other tables and operated upon, using the same tools. While this principle mainly referred to system maintained active metadata, needed by the query engine only, it has to be extended to all types of metadata.

Taking this concept of seamless integration of the various classes of data and metadata with actual data even further, we have opted to integrate all types of data, not only in the same database schema, but also even on the same physical tables, when they refer to the same logical entity. For example, for the logical entities “Survey” or “Classification Value”, we have “informative” data concerning the survey, operational data needed by the system, active metadata and passive metadata concerning the survey, in the same database table.

The second main concept underlying our design is that, in a world of increasingly interconnected information systems, we cannot know in advance the path followed by information towards the final information consumer: metadata can be potentially extracted by several different applications having access to our schema, processed, transferred, formatted and even translated, before it reaches the final recipient, which may be either human or another system. This has many implications:

a) Metadata should be analysed and meticulously decomposed into component elements, at a logical level as elementary as possible; thus the required meta-information, whose exact structure and content cannot be known in advance, may be easily reconstructed via selecting recombining these elementary components at will, in a way that fits optimally the ad hoc user requirements.

b) In order to allow this two stage decomposition – re-composition process, these elementary metadata elements, should be abstracted, classified and codified into a machine understandable format, i.e. the use of “free-text” information should be extremely limited and replaced by the use of codes; when machine-understandable, metadata become useful for the automation of statistical procedures, beyond the basic requirement of user information. Note that, in this way, passive metadata, as defined above, may actually become active ones.

c) Even when confined to human interpretation only, the decomposition of metadata into elementary machine-understandable data elements, allows their flexible delivery into different formats, structures and even different human languages, a feature extremely useful in an interconnected world.

Finally because we do believe that there is still room for improvement in the codification of metadata, the proposed architecture, while exploiting these concepts at a level feasible enough, allows easy future extensions of the system.

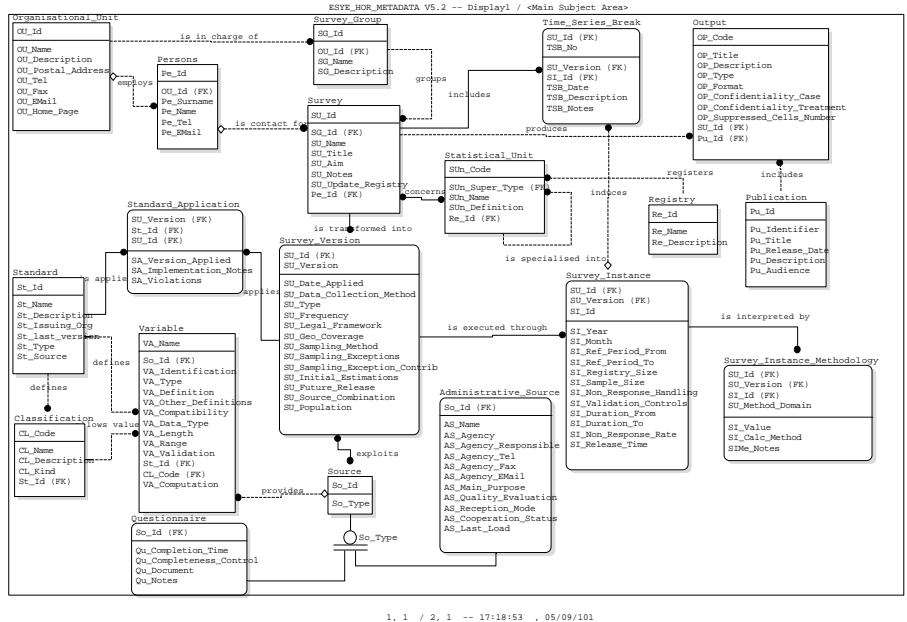


Figure 1. Entity relationship diagram for survey metadata.

The design is presented (in a simplified form) in the accompanying IDEF1X schema. Note that, while most of the entities represented hold also active metadata or even data, only passive metadata related attributes are represented here.

4. Conclusions

We have described a metadatabase model for a general survey (statistical survey, census, administrative source) that is being currently implemented for the Greek National Statistical Service (NSSG). It is an example of an active metainformation system that is related with the processes of the Information System and requires limited human intervention. Furthermore, whenever possible, it contains metadata in a structured form that can be algorithmically exploited in order to perform comparisons (time-, space- and domain-wise).

Future actions include the extension of the relation of the metadatabase with the database in order to support metadata-based statistical processing and the development and implementation of metadata-based quality indicators that will accompany data in the dissemination phase.

References

- [1] **Colledge J.M.**, (1999), Statistical Integration through Metadata Management , International Statistical Review, 67, 1, pp.79-98.
- [2] **Eurostat**, (2000). Definition of quality in statistics. Available in electronic form at: <http://www.forum.europa.eu.int/Public/irc/dsis/qis/library?!=public&vm=detailed&sb=Title>.
- [3] **Grossmann, W.**, (1999), 'Metadata', Encyclopaedia of Statistical Sciences, update Volume 3, pp. 811-815, 1999, S. Kotz, Editor-in-Chief, John Wiley and Sons, New York.
- [4] **Lessler, J. T. and Kalsbeek, W. D.** , (1992). *Nonsampling error in surveys*. New York: John Wiley and Sons.
- [5] **Lyberg, L. E., Biemer, P. P., Collins, M., de Leeuw, E., Dippo, C. S., Schwarz, N. and Trewin, D.** (eds), (1997). *Survey measurement and process quality*. New York: John Wiley and Sons.
- [6] **OECD** ,The role of metadata in promoting international comparisons and adherence to international statistical standards", <http://www.oecd.org/std/metarole.htm>
- [7] **OECD** , 997), Main Economic Indicators, Sources and Methods, Labour and Wage Statistics, OECD, Statistics Directorate, April 1997.
- [8] **Papageorgiou, H., Vardaki, M. & Pentaris, F.**, (2000a), 'Recent advances on metadata', Computational Statistics, 15(1), pp. 89-97.
- [9] **Pellegrino, M.**, (2000), 'The harmonisation of statistical metadata for the European Union: Eurostat's needs and responsibilities', CES, UN/ECE Work session on Statistical Metadata, Washington D.C., USA.
- [10] **Papageorgiou, H. Vardaki, M. & Pentaris, F.** , (2000b), 'Quality of Statistical Metadata', Research in Official Statistics, 2(1), pp. 45-57.
- [11] **Sundgren B.** , (1996), 'Making Statistical Data More Available', International Statistical Review, 64, pp. 23-38.
- [12] **United Nations**, (1995), *Guidelines for the modelling of statistical data and metadata*, Conference of European Statisticians. Methodological material, Geneva
- [13] **United Nations**, (1999), *Information systems architecture foe national and international statistical offices, Guidelines and recommendations*, Conference for European Statisticians, Statistical standards and studies, No 51, Geneva.
- [14] **Vale S. and Pellegrino M.**, (2000), *The Metadata Problem in a European Context*, Eurostat, Workshop on Statistical Metadata, Feb 2000, Working Paper No 11, Luxembourg.