Working Paper No: 1.1

Workshop on Statistical Metadata

(Luxembourg, 14-15 February 2000, Bech Building, Room Ampère)

The Metadata Problem in a European Context

Prepared by S. Vale and M. Pellegrino
Eurostat

1. Introduction

The question of a clear definition of metadata has been discussed at length over many years in several international fora. The basic definition that metadata are "data about data" still seems to be a good starting point.

The "Guidelines for the Modelling of Statistical Data and Metadata", validated by the Conference of European Statisticians in 1995 makes the following general statements:

1. Metadata are physical representation of metainformation - as data are representations of information (representation-oriented definition).
2. Metadata provides information on data - and about processes of producing and using data (contents-oriented definition).
3. Metadata are data that are needed for a proper production and use of the data they inform about (purpose-oriented definition).

In the same paper, metadata needed by users of statistical data are categorised as follows:

- Declarative metadata - concerning the usefulness of data, e.g. contents, accuracy and availability (what may be defined as a "quality declaration" of the data).
- Process-oriented metadata - concerning methodology, references, nomenclatures and, in general, more detailed information about how statistical data have been processed and then how they can be re-used.
- Global metadata and general knowledge - concerning information applicable to a wide range of statistical data.

These definitions, although quite general, show that ideally metadata should be very closely linked with the data that they should help us to find, manipulate and understand. Metadata specifications should relate to the structure and the use of statistical data. Specific metadata management tools (for the management of textual information, or for ensuring metadata quality) should therefore be designed to be integrated into statistical systems rather than as stand-alone entities. Technological developments are making it much easier to link different elements into an overall system, thus making the logical link between data and metadata more of a reality.

For Eurostat purposes, statistical metadata can be seen as information that makes a number, or a group of numbers, understandable as statistics to users, e.g. units, time period, coverage, definitions, classifications, methodology etc.. The mission of Eurostat is “to provide the European Union with a high-quality statistical information system”; therefore, this paper concentrates on metadata primarily as a tool to help users of statistical data.

Having said that about definitions, perhaps it would be better to concentrate more on needs relating to metadata, and to devise ways to organise, store, develop and disseminate metadata to meet those needs. The guidelines mentioned above are very comprehensive, perhaps too comprehensive to be implemented in one go. One possible approach is to pick aspects of these proposals, try them in practice, and modify them, if necessary, in the light of experience. This can be achieved by looking at user needs, gathering together existing metadata (harmonising the layout and contents where relevant), and see how far those needs are met. This will help to identify priorities for future developments.

2. Why do users need metadata?

Talking about metadata means talking about how our databases are organised and how we manage to fulfil the needs expressed by those who use (or who would like to use) our statistical data.

Basically, users need three things from us:

- assistance in the search for data, to find out which data are actually available and how they can be retrieved (data must be *accessible*);
- help to understand meaning and limitations in the use of the data: they need elements for a proper interpretation and a quality assessment of the data (data must be *documented*);
- help to assess the reliability and the quality of the data in detail: they need to know methodological aspects concerning the data, along the stages of the statistical life cycle (data must be *usable*).

Of course, the degree of support required is a function of each user's statistical knowledge, informatics expertise, use of the data and even general knowledge. The growth of statistical dissemination via Internet, for instance, leads to a higher demand for metadata, as the audience is not necessarily aware of the statistical context. At the same time, a higher degree of information is required to assess data quality and to help international comparability.

3. Eurostat's Needs

To satisfy users' needs, Eurostat must be able to collect, produce, maintain and disseminate harmonised, documented and high-quality statistical data. This is our general aim, in common with other statistical and international organisations. Due to the volume of metadata involved, it is also necessary to have efficient and user-friendly metadata management and dissemination systems. It therefore makes sense for us to discuss progress with other organisations, to share ideas and technology, and to try to achieve economies of scale. There is also the very important additional benefit that if several agencies develop similar systems, this can provide a significant boost to international statistical harmonisation.

The specific requirements of Eurostat include the need to be able to handle metadata relating to internationally agreed standards (definitions, classifications etc.), and that relating to national methodologies, used in Member States to produce the data they send to Eurostat. The first type can often be seen as a benchmark, against which the second type can be measured in order to assess quality and particularly comparability.

Following the launch of the single currency, the availability of timely and harmonised data for the Euro-zone has rapidly become a top priority. Information about those data, including definitions, methods and all administrative metadata, must be therefore organised systematically and consistently before being released. For this specific purpose, Eurostat has decided to subscribe to the Special Data Dissemination Standard (SDDS) developed by the IMF, in order to create a uniform and consistent layout of metadata for Euro-zone indicators. This means more metainformation and a better framework for the management of data and metadata throughout the whole European Statistical System.

4. International standards and models

There are several metadata standards in existence. International classification systems (national accounts, industrial statistics, labour statistics,…); general guidelines for statistical metadata on the Internet, published by UN/ECE after discussion within METIS Work Sessions; general and special data dissemination standards from the IMF; OECD metadata standard for Main Economic Indicators. At the same time, more projects are being developed and implemented all around the world with a view to metadata harmonisation. Among these: the Dublin Core project, a global initiative aiming at identifying a common set of information fields for humanities and social sciences; the UN Economic and Social Information System (UNESIS); FEDSTAT, the web site for statistics from 70 US Federal agencies; CANSIM from Statistics Canada, and many more.

These standards and projects are normally developed around metadata for particular types of statistics, and concentrate mostly on what could be considered as "high level" metadata rather than detailed, in-depth information. They also usually concentrate on metadata to present data relating to individual countries, rather than multi-national aggregates. Eurostat has responsibility for a wide range of statistics, and therefore needs a generic metadata model that takes into account the aggregation of data from national sources, and that should be compatible where possible with more specific models, and other international standards.

One approach being considered is the layer model, under which, metadata can be thought of as comprising a series of layers. A table of simple numbers is, of course, meaningless: numbers on their own are not enough to represent statistics. The highest layer of metadata is therefore the information that turns numbers into statistical data. This can include headings, units of measurement, time period, coverage, footnotes, source, as we can see in the following example:

Part-time Employment Rates (%) in Selected Countries (1998)

	Sector		
Country	**Agriculture**	**Manufacturing**	**Services**
A	15.2	53.4	13.5
B	22.6	26.4	46.2
C	38.8	8.9	79.9

Note: Figures for C are provisional, and may be subject to revision
Source: Eurostat

The addition of this type of information makes these data meaningful, and can also be used to facilitate searching for data. This level may provide enough information for some users, but others will want to go to the next level.

The second level of metadata contains more information about definitions, classifications etc. In this case, it would include the definition of part-time employment, and describe exactly what is included in the three items of the sector classification used here. For example, does agriculture include forestry and/or fishing? Information on the quality of the data, e.g. in the form of a quality declaration could also appear here. Normally these first two levels would concentrate on internationally agreed metadata.

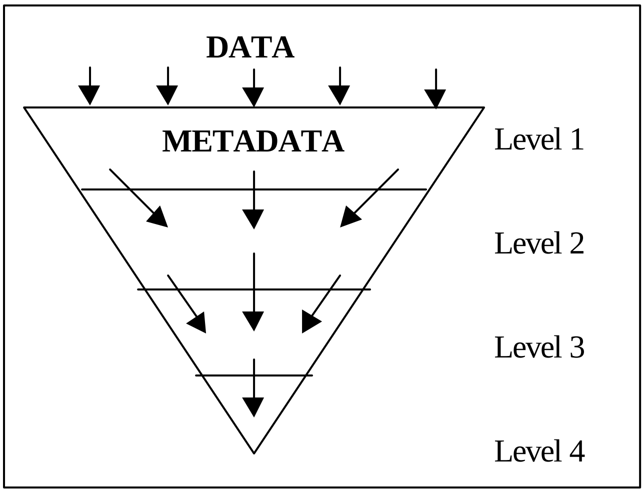
The third level of metadata is better described as methodology, and gives details on how the data is collected (survey or administrative source), and processed (aggregated, validated etc.). For an international agency such as Eurostat, this is usually in the form of methodological texts or papers, and/or information on national data collection and processing methodologies. More detailed metadata describing data quality may also be included in this level.

A fourth level of metadata also exists in some cases. This includes the legal basis for the data collection (e.g. national or international legal texts authorising or requiring the collection of data, or placing limits on the use and dissemination of the data), descriptions of basic registers and generalised information concerning survey or analytical methods.

This approach is broadly consistent with the categories of metadata for users of statistical data in the "Guidelines for the modelling of statistical data and metadata" as described in Section 1, above. Declarative metadata is present in level two, process-oriented metadata mainly in level three, and global metadata mainly in level four.

In terms of the collection and processing of statistical data, it can be envisaged that first level metadata is generally that which should be in a standard format, and closely linked to data as it is transferred from one process to another. Second level metadata is probably more suitable for transmission separately to data, but the volumes may be such that it is still useful to have standard formats (e.g. Claset for classifications). Third and fourth level metadata are less likely to be transmitted on a regular basis, and therefore less likely to require detailed standard formats. Free text is probably more suitable.

This model of levels of metadata also makes sense from the point of view of users of statistics. Generalists just wanting basic information often need go no further than the highest level, but specialists can progress through the levels to get the information they need. This should be made possible by a series of hypertext links, so that each click takes the user to a more detailed level. In terms of storing, organising and accessing metadata, it may be helpful to think of an inverted pyramid.



There is a wide range of metadata at level 1, but as the user progresses through the levels of metadata, the volume that needs to be stored should diminish, as definitions, methodology and legal basis are shared by increasing numbers of data-sets.

Within this general model, specific models and tools should be further developed for metadata within each of the levels, and there may also be a case for changing the number of levels in the light of practical experience. It is important that models are flexible enough to be able to cope with the speed of technological change, and new developments in the collection, processing and dissemination of statistical data.

5. Eurostat's current approach to the metadata problem

Eurostat work in the area of metadata can be split into short-term and long-term projects. The longer-term approach is to construct an integrated European metadata management system, directly linked to statistical data. Eurostat has funded several research projects in this area (e.g. IMIM, IDARESA) and also has several relevant medium to long term internal initiatives (e.g. EDEN, European Reference Environment).

At the same time, there has also been a growing need within Eurostat to develop systems for storing and harmonising metadata in the short-term, without waiting for the various longer-term solutions to become operational.

During the last few years, the reference layer of Eurostat has grown significantly, not only in size but also in the management of metadata. The Working Paper 2.1 (*Metadata Structures in Eurostat*) describes the architecture of the information systems in Eurostat, and particularly the use of metadata in the NewCronos reference database, with the aim of giving a concrete view of how the system actually works.

NewCronos is a huge set of statistical tables, structured hierarchically in 9 themes (General Statistics; Economy and Finance; Population and Social Conditions; Energy and Industry; Agriculture, Forestry and Fisheries; External Trade; Distributive Trade, Services and Transport; Environment; Research and Development) and then in domains, collections and, if needed, groups and subjects. This reference layer has to ensure the flow of data and metadata from the production to the dissemination layer, maintaining metadata in three languages. This all takes place within a framework where the harmonisation of codes and concepts is on-going, and the system itself is evolving towards a new structure.

If these projects and processes are properly co-ordinated this should not cause problems, and can even give a few benefits, such as the option to try new ideas quickly and cheaply on a small scale, whilst maintaining an overall vision. In the past co-ordination between projects relating to metadata management and development within Eurostat has not been as strong as it should have been. Pragmatic solutions have been adopted to meet specific needs, and the overall vision has not been so clear. User needs have rightly led to pressure for greater co-ordination efforts, and the challenge now is to bring these diverse strands together into a consistent policy framework. The two-track approach, of long and short-term solutions can work if managed carefully. It can provide tangible outputs, and support harmonisation work, whilst keeping in mind longer-term objectives.

5.1 "Business Methods"

One recent short-term development is the "Business Methods" project, which aims to develop a tool for the harmonisation and dissemination of metadata relating to business statistics methodology. "Business Methods" is actually a set of different systems that are in various stages of development, which are brought together using hypertext links within an Internet environment. It was considerably quicker to do this than to try to develop a single system to hold all of the metadata in a common format. This modular approach also has the important advantages that individual elements of "Business Methods" can evolve to cover other areas of statistics, and can be made available for other applications.

A key consideration in the development of "Business Methods" has been how to ensure coherency with future, wider metadata systems. This is not easy, as there are several longer-term projects, and their eventual outcomes are as yet unknown. Few informatics projects with a time-span of more than one year are likely to develop exactly according to the original plan, mainly due to the pace of technological development.

The solution has therefore been to try to bring metadata together using basic formats (e.g. html and pdf) and widely available software tools (e.g. Microsoft Access), thus keeping "Business Methods" as technically simple as possible. This makes it easier for non-informaticians to maintain, and also means that the metadata held in "Business Methods" should be relatively easy to transfer to future systems. In this way, "Business Methods" and its various components should not be seen as end products, but as steps towards more comprehensive future solutions, where data and metadata are fully linked.

5.2 The harmonisation of metadata for Euro-zone indicators

The EURO-SICS project for a common site of short-term statistical indicators of EU countries, agreed and launched in 1999, concerns the availability of more than 500 indicators agreed between Eurostat and Member States. Euro-SICS comprises a set of long time series, harmonised and national, available at a monthly or quarterly frequency, with detail at national level for all EU countries.

The domains concerned are: national accounts, money and finance, external trade, balance of payments, prices, industry and services, energy, retail sales, labour market and short-term business surveys. Some country-specific indicators will be proposed and directly updated by Member States, in addition to European aggregated data, already available via Eurostat.

One of the main challenges, in this regard, is to provide institutional users with a full set of methodological information for both harmonised and national domains and for each indicator. For this reason, Eurostat decided to subscribe to the SDDS (Special Data Dissemination Standard) developed by the IMF, in order to have a uniform layout and a general consistency with a reference international standard accepted by most of the individual countries. However, metadata repositories are not yet organised systematically and consistently: the layout, quality and level of breakdown of available information still vary considerably from one domain to the other.

At present, Eurostat is preparing methodological information for all European aggregates according to the SDDS, going beyond the minimum requirements for certain aspects such as the "methodology page", which is not compulsory in the original IMF layout. For country-specific series, methodological notes will be produced by member countries following the same scheme as the others. This exercise should be logically supported by a better partnership between Eurostat, Member States and other international organisations such as ILO and the OECD, that produced special standards and actually collect and maintain metadata for the same domains.

One important lesson from this experience is that attempting to harmonise first, before starting to disseminate, is very often a "pious" hope, whereas the pressure on data dissemination, coming from institutional and private users, is able to give a considerable boost to the process of producing and consolidating the corresponding metainformation.

6. Conclusions

Which priorities can we outline for future actions concerning metadata? Eurostat, at present, is particularly involved in a sort of twofold strategy: the standardisation of statistical information in several *production* domains and the *dissemination* of high-quality and timely data for the European Statistical System. For both lines of action, we must ensure that developments related to metadata are closely co-ordinated, so that they fit into the overall vision of the future production, reference and dissemination environments.

In the meantime, we must learn to live with non-perfect harmonisation of data and metadata, at least in the short run. The development of more efficient reference and dissemination tools will hopefully create their own pressure for greater harmonisation in the future. In this context, the growing demand coming from our users for timely statistical data is one of the most powerful tools for stimulating the production, exchange and dissemination of relevant and coherent metainformation.

In our integrated statistical system, various national and international bodies collect, maintain and disseminate data and metadata, with a high risk of duplications and inconsistencies. For this reason, we need to develop closer partnerships within the European Statistical System and with other international organisations, such as the OECD, ILO, UN and the IMF, to make sure that our metadata systems are compatible, wherever possible.