

Work Session on Statistical Metadata
(Geneva, Switzerland, 18-20 February 1998)

Item 6 of the provisional agenda

**COMMENTS ON “GUIDELINES CONCERNING STATISTICAL METADATA ON
INTERNET”**

Submitted by Eurostat

These notes contain some general considerations on the subject, together with a series of comments to the following Papers:

- (1) Standards for Statistical Metadata on Internet, prepared by the UN/ECE Secretariat.
- (2) Guidelines for Statistical Metadata on the Internet, prepared by Statistics Norway.

The above papers were discussed within the EUROSTAT Directorate for “Statistical Information Systems, Research and Data Analysis”.

The following comments and suggestions have been prepared under the responsibility of Unit A3, in charge of “Reference Data Bases”.

I. INTRODUCTION: STANDARDS AND GUIDELINES

1. Document (1) has a more general structure and can form the basis for a further draft of the proposal, but a few corrections and some additional specifications seem to be needed. Document (2) provides some useful corrections to the first proposal, from a more limited and pragmatic point of view.

2. A set of guidelines for the dissemination of statistical meta-information is desirable, for helping data suppliers and, indirectly, end users. The degree in which those guidelines will be actually used depends, however, on our ability to meet the requirements of both users and suppliers of statistical information. No world-wide standard on metadata can be bureaucratically enforced from the top, in a commercial domain whose main feature is actually its freedom of initiative. The position expressed by document (2) for the use of the term *guidelines* is therefore acceptable.

3. A final proposal, anyway, should start from a clearer definition of the following aspects: a) Internet specifics; b) user groups; c) metadata classes.

II. INTERNET SPECIFICS

4. First of all, the interaction between existing information systems and the dissemination channel should be carefully considered. Being the world-wide-web a global information network, each guideline must be carefully addressed to the international community, where the knowledge of national environments can vary significantly within the audience.

5. In spite of the fact that the data available for dissemination are the same, the requirements for managing statistical meta-information on the Internet may differ significantly from what we have already experienced with other media, for various reasons:

- A wider audience may actually lead to a higher need for metadata. The audience for Internet is potentially larger and more geographically widespread. Consequently, average users probably need a more solid assistance in terms of meta-information, if they are not aware of the statistical context.
- The larger scale of the audience and the higher usage make all kind of inconsistencies far more visible. It follows that a higher degree of information is required to assess data quality and, generally speaking, to help international comparability.

- Metadata are displayed continuously on the Internet, so users expect them to be regularly updated: this means that freshness cannot be considered as a second-category option.
- Internet makes the interaction between data and metadata easier, up to the point where that distinction actually begins to fade. Even the less experienced Internet client is used to navigate from page to page, looking through all kind of information: a statistical end user can easily go from press releases (data merged with metadata) to methodological notes (metadata) to data bases (data) with attached footnotes (metadata) and then back to other data and more text files with comments and so on, possibly blending different sources to have a good product at the end. The key points, for assisting this kind of navigation, are the general structure of the whole “information base” we put on the Internet and the rules for the accessibility. Advanced search tools are needed, of course: the information bases should be organised accordingly, combining the standard hierarchical organisation by general themes and domains with a specific statistical thesaurus.
- The organisation of the information base in hypertext format is a specific feature induced by Internet, although it can be also regarded as a quality challenge. The set-up of appropriate management and control tools (for the organisation of web pages and the management of hyperlinks, for instance) can lead, at the end, to an improvement of the quality for the whole information base.

6. What follows from the above listed points is that, probably, statistical data supplied on the Internet need to be more metadata-assisted (not only in volume, but also in quality) than in traditional dissemination channels, normally targeted to more delimited user groups.

III. USER GROUPS

7. A higher effort must be made, in order to understand who are the users we should address to, what they need in terms of meta-information, why (for which purpose) and, finally, how such requirements can be met by the suppliers of statistical information, on the Internet market. Evidently, it would not be reasonable to focus on the needs of a “standard” general user, whose existence can be also questioned.

8. Users of statistical information can be, broadly speaking, either statisticians (who can re-process the data) or end users. Within the “end users” group, we find subject-matter researchers, political decision-makers, public officials, executives, teachers, students, librarians, journalists or residual general public. Each group can be obviously split up into smaller subgroups, and each user can possibly belong to more than one group. But also statisticians can be divided into many different groups, according to the organisation or company they belong and according to the knowledge they have of a particular topic. A tentative study of user groups could start from a broad distribution of the main users by competence, knowledge about data sources, information needs and ability (inclination) to use informatic tools.

9. The minimum set of required metadata should be then defined with reference to different user groups:

- a) general audience, including mass-media: those users are generally not very skilled and not too familiar with statistics (Eurostat’s experience seems to suggest that approximately one third of the users has no opinion at all about freshness, comparability and coverage of statistical data);
- b) skilled users with a good level of specific competence, but with no time or inclination to search and retrieve a wide load of information: those users look for a customised service, but they are able to assess the general likelihood of data;

- c) expert users, skilled in searching, retrieving, assessing quality, interpreting and eventually producing statistical information on their own.

IV. METADATA CLASSES AND USER-ORIENTATION

10. The definition of “minimum metadata” as an isolated exercise is quite impossible, as it is extremely difficult to discuss metadata independently from the data they should help to find, manipulate and understand. Press releases, one of the most common dissemination tools and also a leading source of statistical information for end users, are a clear example of how data and metadata are often merged in a combined and self-explanatory information product.

11. Data cannot be drastically separated from metadata because metadata specifications normally depend on the structure and the use of statistical data. The existence of specific metadata management tools (for the management of textual information Vs numerical information, or for ensuring metadata quality) does not seem to justify a sharp logical separation. Above all, if statistical metadata are “data which are needed for proper production and usage of statistical data”¹, we cannot leave the user out of our considerations. It follows that the usual purpose-oriented classification should be integrated with a clear user-oriented approach.

12. Document (1) proposes the following purpose-oriented classification of metadata: a) metadata supporting clients in searching for the information; b) metadata used by clients in interpreting and using the data; c) metadata used by clients in evaluating the quality of the data (page 4).

13. A first remark is that point c) and point b) are not logically separated: as correctly pointed out by document (2), the quality assessment and the interpretation of data must be considered as part of the same logical process (page 2).

14. Another remark can be done about “metadata supporting clients in searching for data” (Document 1, page 5), where the concept of “variable level” is not specified and can be even misleading, especially when used in association with automatic searching tools. A thesaurus-like approach is somewhat larger and it is not limited to the concept of variable.

15. In a user-oriented approach, the main purposes for the use of metadata can be better outlined as follows:

- a) users need to be assisted in the search for the data, to find out which data are actually available and how they can be retrieved;
- b) users need to understand the meaning and the limitations in the use of the data: they must be provided with all main elements for a proper interpretation and a quality assessment of the data (data must be well documented);
- c) expert users must be able to assess the reliability and the quality of the data in detail: they need to know each methodological aspect concerning the data, along all the stages of the statistical life cycle.

16. Referring to those three pillars, we can try to draw up a list of descriptive attributes that should allow a basic organisation of statistical meta-information:

¹ United Nations Statistical Commission and Economic Commission for Europe, *Guidelines for the Modelling of Statistical Data and Metadata*, New York and Geneva, 1995, page 2.

- **Search:** all the information items needed in the search for the data, such as: classification plan based on themes, domains or subjects and so on; access to a keyword search; access to the latest press releases; sources; coverage; basic definitions (including measurements units); periods; update calendar; contact points (e-mail) and references for further information. Last but not least, textual metadata needed by the user in order to be able to handle the software tool properly (software-related metadata).
- **Interpretation:** metadata for the interpretation and elements for a quality declaration of data (contents, methodological notes, model assumptions, survey information, accuracy, error sources, comparability over time and space,...). The information should be displayed in the easiest way for the user, for instance by using attached footnotes. References for further information should be normally included.
- **Process-oriented metadata:** access to a comprehensive source for methodology, concepts and definitions, nomenclatures and, in general, more detailed information about how statistical data have been processed and how they can be re-used (survey information, reporting methods, data manipulation, quality checks,...).

17. Quite obviously, *process-oriented* metadata partly overlaps with the metadata for general interpretation, because methodology and quality issues form a common ground for both categories, at different levels of investigation.

18. Each class of metadata seems to have some interest, for a world-wide audience. In the international community, particularly, comparability seems to be a vital issue, and real comparability relies on the awareness of how data have been collected and processed.

V. FINAL RECOMMENDATIONS

19. Considering metadata classes and user groups at the same time, in a sort of a bi-dimensional table, we should be able to propose a series of analytical guidelines for the dissemination of each class of metadata (search, interpretation, process-oriented) at each user level (general level, skilled level and advanced level).

- **General level Metadata.** The so-called general public (including journalists to a large extent) is normally interested in aggregate data at a general level, with a “minimum set of metadata items”, associated to a sort of assisted access. They need all metadata concerning the *data search*, and a general information on the contents.
- **Medium level Metadata.** Researchers, decision-makers and other subject-matter users need, in addition to the previous group, a more solid assistance in terms of metadata for interpretation and data analysis, as well as a more advanced software support.
- **Advanced level Metadata.** Expert statisticians and full-time researchers should receive full metadata assistance, with reference to statistical methodology, concepts and definitions, nomenclatures, data quality and process-oriented metadata.

VI. SUMMARY

- A **set of guidelines** for the dissemination of statistical meta-information should meet, in the first instance, user's requirements.
- **Technology is never neutral**: if the guidelines refer to the world wide web, they have to go into Internet specifics. Probably, statistical data supplied on the Internet will have to be more metadata-assisted (not only in volume, but also in quality) than in traditional dissemination channels, normally targeted to more delimited user groups.
- **Internet can change the client's attitude towards data retrieval**. Being used to browse and navigate from page to page, the Internet user is no longer willing to accept any unjustified (or perceived as such) lack of information. Consequently, the average user of statistical information should be left free to shift from press releases to general data bases and then to methodological notes, to any other related domain and maybe back to more explanatory notes.
- **Data must be well documented. Users must be provided with the elements for a proper interpretation and a quality assessment of the data**. Internet, in general, is a remarkable quality challenge for statistical dissemination: this will be particularly evident with the adoption of new generalised formats.
- **Explanatory elements are important for all users**: they may be embodied into plain explanatory texts, flags or footnotes, by using each technical facility offered by Internet.
- **The minimum set of required metadata should be defined with reference to both user groups and general purposes**. Considering metadata classes and user groups at the same time, in a sort of a bi-dimensional table, we should be able to propose a series of analytical guidelines for the dissemination of each class of metadata (search, interpretation, process-oriented) at each user level (general level, skilled level and advanced level):

General level Metadata. The so-called general public (including journalists to a large extent) is normally interested in aggregate data at a general level, with a "minimum set of metadata items", associated to a sort of assisted access. They need all metadata concerning the *data search*, and a general information on the contents.

Medium level Metadata. Researchers, decision-makers and other subject-matter users need, in addition to the previous group, a more solid assistance in terms of metadata for interpretation and data analysis, as well as a more advanced software support.

Advanced level Metadata. Expert statisticians and full-time researchers should receive a full metadata assistance, with reference to statistical methodology, concepts and definitions, nomenclatures, data quality and process-oriented metadata.