WP 14
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 9-11 February 2004)

Topic (iv): Using metadata for searching and finding statistical data in websites and portals

STATISTICAL METADATA ON THE INTERNET REVISITED

Invited Paper

Submitted by Statistics Norway[1]

I. INTRODUCTION

1. Metadata has been an issue in many statistical meetings, conferences and research projects supporting the production and dissemination of official statistics. The development of Internet as the main channel for disseminating such statistics has put emphasis on metadata and also led to new developments in this area. The Conference of European Statisticians discussed Internet and metadata at its 1997 plenary session. Following this, Statistics Norway prepared some draft guidelines for Statistical Metadata on Internet, which, after discussions, were adopted in 1998 (Statistics Norway 1998).

2. The guidelines for statistical metadata on the Internet from 1998 consist of recommendations on metadata to explain the meaning of statistical data, and other information to help the user to find and use relevant statistics. The 1998 paper also includes a number of recommendations on good practices for statistical web services.

3. The guidelines were followed up in a survey in 1999 (Byfuglien 1999). The status with regard to compliance with the guidelines was mapped by self assessment by 12 European NSIs. Most of these NSIs claimed to comply with the majority of the points mentioned in the guidelines, though all of them had a potential for improvements. In general, they found the guidelines useful.

4. The guidelines have been revisited by investigating the web services of selected National Statistical Institutions in 2003. This has partly been done in connection with work on documentation in the Eurostat LEG on quality, and also linked to the work in the research project Metanet (2003).

5. Questions considered are:

- What is the status regarding the implementation of the recommendations?
- Are the recommendations still valid, or are some of them outdated?
- Are there new recommendations that should be included?

[1] Prepared by Hans Viggo Sæbø, hvs@ssb.no

6. The paper summarises observations done during this investigation and with regard to the guidelines mentioned. Some new features and challenges for statistical metadata on the Internet are also considered.

II. FRAMEWORK

A. What is metadata?

7. The term "metadata" is often used synonymously with "documentation". However, on the Internet *structure* is a key word when it comes to presenting statistics (e.g. organisation of a huge amount of information and hypertext techniques), and documentation for a wide variety of users without a high degree of structure will tend to be useless. Hence, in this paper we will use the term statistical metadata to mean structured information about statistics.

B. Quality and metadata

8. Many National Statistical Institutes (NSIs) have started a systematic quality work. Statistics Norway's work on this is described by Sæbø, Byfuglien and Johannessen (2003). Quality can most simply be defined as *fitness for use* in terms of user needs. The dimensions of product quality for statistics are often described according to Eurostat's criteria (Eurostat 1998):

- Relevance and completeness
- Accuracy
- Timeliness and punctuality
- Comparability and coherence
- Accessibility and clarity

9. Cost constraints are important, and costs always have to be considered in connection with quality indicators. Statistics must also be objective, and personal integrity must be protected. Good product quality is necessary to satisfy user needs, but improving processes (in the production of statistics) is the key for better product quality at an acceptable cost.

10. Documentation and metadata can be looked upon as a part of quality as well as a precondition for improving quality. Metadata are necessary for the clarity of statistics to be found on Internet, and metadata assisting search and further processing also contribute to relevance and accessibility, by reducing the costs of retrieving statistics. Information about production processes is often needed in order for the users to understand the statistics as well. Such information is of course crucial for improving production processes.

C. Types of metadata

11. Metadata for users of statistics are only one of several types of metadata. One way of classifying statistical metadata is:

- Metadata for the users of statistical information (for finding and navigation, explaining and post-processing)
- Process documentation for internal users (for control and improvements)
- Metadata for external data providers (to provide correct data)
- Quality information

12. This paper deals with the first group of metadata. However, there is some overlap between the groups. As mentioned, users of statistics often need to know something about the production processes and certainly something about the product quality as described according to the Eurostat quality dimensions. Some of the 1998 guidelines refer to process and quality information. The broader picture of metadata types should therefore be kept in mind.

D. Internet dissemination then and now

13. Technology and web practices have developed rapidly. Technological constraints of (unsophisticated) users and NSIs are not issues to the same extent. The usage of statistics on the Internet has exploded. The increase in the number of external "hits" on Statistics Norway's web site from 8 million in 1998 to about 40 million in 2003 is typical. For most NSIs, Internet is now their main channel for disseminating statistics. This means that statistics are presented only on the web or at the same time or before they are published on paper, and in a format (e.g. html) that can benefit from hypertext (even if there are still a lot of publications in pdf-format on the web).

14. Two trends can be observed for the dissemination of statistics on the web compared to former paper dissemination:

- Large and complicated tables have been replaced by small and simple tables and graphics, to cover the needs of the media and the general public in particular.
- Statistics are often available in databases where more advanced users can select and construct their own tables.

15. This represents both new challenges and possibilities for metadata. More self-service requires more and better explanatory metadata; metadata that are not linked to pre-defined tables to the same extent as before. Larger volumes of statistics on the web put more demand on "discovery" metadata for looking for statistics and navigation.

16. Pullinger (2003) has described work on discovery metadata. Lamb (2003) has considered some challenges in research on metadata for official statistics.

E. Guidelines from 1998

17. The 1998 guidelines paper classifies the different types of metadata on the Internet by metadata assisting:

- Search and navigation
- Interpretation
- Post-processing

18. The guidelines are briefly summarised as follows:

Metadata assisting search and navigation:

- Metadata providing general information about the statistical web site (sitemap, frequently asked questions, descriptions of institution and subject areas, product overview, contacts and links)
- A hierarchical subject matter classification
- List of key words
- Local search engine

Metadata assisting interpretation:

- Title/content description of tables (normally statistical population, geographical coverage, observation unit, classifications)
- Measurement unit
- Labels for rows/columns with proper definitions
- Time period
- Regional unit
- Comparability over time
- Footnotes highlighting specific precautions
- Sources of data (agency)
- Explanation of standard symbols in tables
- Any information on copyright
- Contact points for additional information

19. In addition, some recommended metadata have been proposed, such as comparability with alternative sources and links, descriptions of production methods, information on errors and accuracy and description of background and purpose of the statistics, definition of concepts and variables and information about standards used.

20. Metadata assisting post-processing basically include metadata assisting interpretation in addition to data to allow downloading and further processing using suitable tools (e.g. spreadsheets, databases, packages for statistical analyses). Information on possible restrictions and limitations for further use, such as suppression of data and rounding, is also mentioned in the 1998 guidelines.

III. EVALUATION OF STATUS

A. Evaluation method

21. All types of the metadata comprised by the 1998 guidelines are metadata or documentation for users of statistics, and the status has been evaluated by investigation of the web services of 20 statistical institutes in Europe, North America and Australia. The same technique has been used in connection with work on documentation in the Eurostat LEG on quality, and also supporting the work in the research project Metanet (2003). However, this study was limited to Europe and concentrated on metadata needed for the interpretation of statistics. It focused on systematic documentation according to templates with fixed headings. The present search has been extended to cover all the metadata covered by the 1998 guidelines. In particular, this means that the search has been looking for metadata for discovery and navigation in addition to metadata for explanation and usage.

22. Statistics can be presented in several ways on a web site, and documentation often varies according to the type of statistics and presentation. It can be convenient to distinguish between:

- Statistical news/press releases/daily statistics
- Other statistics in html-format, where tables can be copied and/or downloaded
- Statistics presented in publications or on sheets only available on the web in pdf-format
- Statistics in databases that can be accessed by Internet, and where the users normally can form and select their own tables:
    - Databases with all types of statistics
    - Databases with time series (mainly short term statistics)
    - Databases (only) with specific datasets, i.e. data from specific surveys, censuses etc.

This grouping of statistics and presentation modes has been used as a reference in the search for documentation.

B. Overview

23. Many of the metadata considered in the 1998 paper are simple metadata linked to statistical tables, often to be found in headings, labels and footnotes. These metadata are necessary, and normally presented in or linked to tables on the Internet as well as on paper. However, the developments mentioned towards simpler predefined tables, graphics and databases in addition to web technology with links and hypertext possibilities have provided opportunities for new and more advanced metadata solutions. Hence, some of the 1998 guidelines are not very challenging any more (even if they should be followed when applicable), and others and more ambitious solutions have come into focus. This in particular concerns the recommended metadata about concepts and definitions, description of the production of statistics and quality information about accuracy, comparability and coherence. The huge increase of statistics available on the Internet has put more emphasis on metadata for finding statistics and navigation. The need for (common) standards and structures of metadata has also come into focus.

24. Most recommendations included in the minimum set of guidelines have been fulfilled in the sense that such metadata exist on the Internet. Guidelines for metadata linked to statistical tables are followed both for tables in pdf- and html-formats. However, there are great differences between the countries when it comes to the organisation and structure of metadata for statistics presented primarily for electronic dissemination (in

html-format or databases). In many cases it is difficult to find the metadata, making both navigation and understanding of the statistics difficult. A more specific evaluation follows.

C. Metadata assisting search and navigation

25. General information about the statistical web site with descriptions of institution, products, contacts, and links is normally easily accessible from the institution's home page. Most of them have a sitemap, but only a few have f.a.q. (frequently asked questions).

26. One can, however, discuss the convenience of f.a.q. in the case of a statistical institute. Forms for asking questions are normally available on web, and what is important is that such questions are registered, analysed and taken into account in the structuring of the web service. Managers of statistical web sites should have their emphasis on ensuring that popular statistics are easily accessible rather than on developing f.a.q.. Experiences from Norway are that statistics about prices (e.g. CPI) and population are in frequent demand, in addition to statistics on names. These types of statistics are easy to find from most NSI home pages.

27. Most NSIs have a hierarchical subject matter classification and a local search engine. However, less than half of those investigated have a list of key words for navigation purposes. There is a potential for improvement here. Ongoing work on thesauri for statistics should be mentioned in this context. Work on such thesauri for example takes place in Eurostat (2003b) and in the data archive networks and projects, see reference to relevant LIMBER (Language Independent Metadata Browsing of European Resources) web site.

D. Metadata assisting interpretation

28. While most of the metadata presented in or closely linked to statistical tables (such as footnotes) on paper are normally available also when the tables are presented on the web, great differences between the countries are revealed when it comes to more comprehensive metadata, and the way they are linked to the statistics presented (in html-format or in databases). Only a few countries offer systematic documentation linked to the statistical news, but some countries have this kind of documentation linked to statistics available in databases.

29. One exception from the statement that table type metadata are presented as well on web as on paper is the explanation of standard symbols in tables. This type of metadata seems to have been forgotten or well hidden in html format on the web. Since web tables are often simpler than paper tables covering book pages, one could argue that the need for explaining symbols has been reduced. However, there are examples of tables where these symbols vary within the same web service. This may be due to use of different software. This is still a detail that should be checked by the web managers.

30. Systematic documentation about the production of statistics, definitions of concepts and quality is typically named "about the statistics" or "facts about statistics". This is largely documentation according to a fixed scheme or template. About one third of the countries have such information linked directly to the statistics, most of them primarily for statistics in databases. In addition, some countries have such information available in a more general form (covering several statistics), linked to the home page or a special page for metadata. Some countries also provide definitions of concepts by using hyperlinks. Web technology provides good possibilities for providing systematic documentation according to a fixed scheme or template, and in addition to the need to develop this kind metadata there is a need to standardise the content. Box 1 is an example showing the items linked to all statistical news from Statistics Norway. This set up is typical for the NSIs with this kind of metadata. The IMF General Data Dissemination System (GDDS) and Special Data Dissemination System (SDDS) are other examples of systematic documentation schemes, see reference to relevant IMF web site.

31. To support the further work to promote good documentation practices in the European NSIs, the Eurostat Quality LEG is carrying out annual surveys on documentation status in the European NSIs. The last survey was carried out in June 2003. Here the status regarding systematic documentation schemes (e.g. according to a template with fixed headings) was asked about. The results are presented in Eurostat (2003a).

About half of these NSIs claimed that they had or planned to have such schemes in place on the Internet in the near future. The answers correlate well with the status observed by scanning the web services.

32. The 1998 recommendation also mentions information about standards (classifications) used and information on possible suppression of data (for post processing purposes). It is natural to explain both these issues in a standardised documentation scheme (see box 1), but many countries also present the most common standard classifications on their web sites, in some cases there are databases containing classifications. Such classifications are international, and links to international sources are also given.

Box 1. About the statistics - Example from Statistics Norway

1. **Administrative information**
   - 1.1. Name
   - 1.2. Frequency
   - 1.3. Regional level
   - 1.4. Type of statistics (register, census, sample survey)
   - 1.5. Subject group
   - 1.6. Responsible division
   - 1.7. Authority
   - 1.8. Response burden
   - 1.9. EU regulation (if relevant)
   - 1.10. International reporting
   - 1.11. Funding
2. **Background and purpose**
   - 2.1. Purpose and history
   - 2.2. Users and applications
3. **Statistics production**
   - 3.1. Population
   - 3.2. Data sources
   - 3.3. Sampling
   - 3.4. Collection of data
   - 3.5. Control and editing
   - 3.6. Calculations
   - 3.7. Confidentiality
4. **Concepts, variables and classifications**
   - 4.1. Definition of the main concepts and variables
   - 4.2. Standard classifications
5. **Sources of error and uncertainty**
   - 5.1. Measurement errors
   - 5.2. Non response
   - 5.3. Sampling errors
   - 5.4. Other errors
6. **Comparability and coherence**
   - 6.1. Spatial comparability and comparability over time
   - 6.2. Coherence with other statistics
7. **Accessibility**
   - 7.1. Internet address
   - 7.2. Publications
   - 7.3. Storing and use of basic material
   - 7.4. Other documentation

E. Metadata assisting post-processing

33. In addition to the metadata mentioned these metadata comprise information to allow downloading and processing using suitable tools (e.g. spreadsheets, databases, packages for statistical analyses). Most statistical web sites have the possibility to convert statistics to spreadsheet formats (typically the generalised csv-format), and many to other formats as well (in particular data from databases). Several countries also offer possibilities for some on line post processing, for example manipulation of tables, graphics and maps, and also free downloading of software. One example is the statistical software PC-Axis that is used and offered by the Scandinavian countries and Spain.

34. In general, most types of software are able to read and convert to the most common data formats today, and the need to develop more or new metadata to allow for downloading is not the same as 5 years ago.

35. However, the downloading possibilities do not apply to statistics only presented in pdf-format, and web sites with large amounts of this (paper) format will certainly have an improvement potential. Efficient downloading from html-formats and even spreadsheets may be difficult due to different use of symbols in tables (also use of nil versus blank and digit grouping symbols).

IV. CONCLUSIONS AND CHALLENGES

36. Even if the 1998 minimum guidelines regarding statistical metadata on the Internet at large are followed, the volume and complexity of the web requires easy access and a good metadata structure. This is an important challenge for the NSI web services today.

37. Discovery metadata or metadata for search and navigation have become more important than before. In this area most web services offer a number of possibilities. There is a potential for improvements, in particular for those NSIs that still have not followed all the 1998 recommendations (for example to have a list of key words), but in general regarding easy access to the recommended possibilities (for example to the sitemap and to the hierarchical subject matter classification).

38. When it comes to metadata assisting interpretation, the more ambitious recommendation regarding information on background and purpose of the statistics, production methods, definition of concepts and standards, errors and accuracy, comparability and alternative sources, still represent a challenge for most institutions. This in particular concerns structuring and making this information easily available. There is a need to standardise this kind of structured documentation, which should be available by links directly from the statistics presented in tables, graphs and maps, or from databases where statistics can be selected.

39. The quality of metadata itself is another crucial issue. The 1998 guidelines and this paper have concentrated on the *presence and accessibility* of metadata. However, it is clear that even if structured documentation about the production of statistics exists, the descriptions can vary in length, correctness and how easy it is to understand them. This is certainly the case in Statistics Norway, where we are just now about to check the quality of "about the statistics". To improve this, we will provide those responsible for the statistics with some examples of "best practices" for different types of statistics (e.g. survey based statistics and register statistics).

40. Regular updating is necessary for a metadata system. It is not enough to update metadata once, they have to be checked and if necessary modified for every new release of the statistics.

41. Technical solutions ensuring that metadata can easily be updated together with the data, and only once and one place (i.e. database) is a precondition for an efficient metadata system. This is a great challenge for most of the NSIs, working with linkage of several metadata systems (for example different systems for data collection, production and dissemination of statistics, in addition to databases for standard classifications, see Statistics Norway 2004). The work on this is beyond the scope of this paper, but it is crucial for the quality and updating of metadata.

REFERENCES

Byfuglien (1999):"Compliance with UNECE guidelines for statistical metadata on the Internet. An overview and an example". Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, 22 - 24 September 1999.

Eurostat (1998): "Definition of quality in statistics". Doc. Eurostat/A4/Quality/98/General Definition.

Eurostat (2003a): "2003 LEG implementation status report". Draft document G0/LEG-IMPL/32.

Eurostat (2003b): "THESEUS - A multilingual thesaurus for accessing Eurostat's reference databases". Document for the Metadata Production and Exchange Workshop. Luxembourg, 3-4 April 2003.

IMF: Information on dissemination standards and metadata: http://dsbb.imf.org/Applications/web/dsbbhome/.

Lamb, J. (2003):"Metadata and Official Statistics: Future directions". Paper to ISI Conference, Berlin, August 13 - 20 2003.

LIMBER project website: http://www.limber.rl.ac.uk/.

Metanet (2003): "METANET PROJECT: WG 4 Deliverable - Adoption issues".

Pullinger, D. (2003): "Integrated searching and communication of statistical information using discovery metadata". Paper to ISI Conference, Berlin, August 13 - 20 2003.

Statistics Norway (1998): "Guidelines for statistical metadata on the Internet". Statistical Journal of the United Nations ECE 15 (1998) 169-176.

Statistics Norway (2004): "Variables documentation system in Statistics Norway". Contributed paper to the Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, February 9-11 2004.

Sæbø, H.V., Byfuglien, J. and Johannessen, R. (2003): "Quality Issues at Statistics Norway". Journal of Official Statistics, Vol. 19, No. 3, 2003, pp. 287-303.