UNITED NATIONS

E

Economic and Social Council

Distr.
GENERAL

CES/1998/32
23 April 1998

ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Forty-sixth plenary session
(Paris, 18-20 May 1998)

GUIDELINES FOR STATISTICAL METADATA ON THE INTERNET

Paper contributed by Statistics Norway[1]

I. INTRODUCTION

1. The Conference of European Statisticians discussed Internet and metadata at its 1997 plenary session. Following a request of the Bureau of the Conference, Statistics Norway prepared, on the basis of earlier discussions, some draft Guidelines for Statistical Metadata on Internet. The comments made on that draft by the Work Session on Metadata (Geneva, 18-20 February 1998) have been taken into account in preparing the current version. As recommended by the Bureau, the Guidelines are being submitted to the 1998 plenary session of the Conference.

2. The paper proposes some guidelines for minimum standards for statistical metadata on the Internet. Internet represents a lot of challenges and provides a lot of possibilities, but it is believed that minimum standards should be realistic and narrowly focused, rather than broad encompassing, at this stage. As statistical institutes are at different levels of developing a web service, the document should be regarded more as guidelines than as strict standards.

[1] Prepared with the assistance of a working group composed of Canada, USA, EFTA, Eurostat, OECD and the UN/ECE secretariat.

II. OBJECTIVES

In addition to the proposed guidelines, the paper mentions some more ambitious possibilities. However, technology with its challenges and possibilities changes rapidly. This may necessitate future revisions also in the proposed guidelines, whether they are regarded as standards or just as recommendations.

3. The Guidelines are intended for National Statistical Institutes and other national and international bodies disseminating statistics.

4. The main objectives of the Guidelines are the following:
   (I) to support the **world-wide dissemination of statistical data** on the Internet;
   (ii) to promote **a consistent interpretation of statistics from different sources** by considering data quality and international comparability as strategic issues;
   (iii) to promote a **proper usage and processing of public statistics.**

5. Besides metadata which explains the content of statistical data, the Guidelines also consider some other information relevant to searching and usage of statistical information on the Internet. They do not include more advanced features, such as metadata for search at the micro data level, restricted user access to databases, metadata issues linked to pricing, statistical data confidentiality and the general transmission of statistical data via Internet (e.g. as part of data collection).

6. The Guidelines are intended to be independent of any specific technologies. Some general technical and design issues are discussed briefly in Section VI “Other recommendations”.

III. METADATA ISSUES SPECIFIC TO THE INTERNET

7. Although the basic requirements for metadata on the Internet are much the same as those for any other medium, there are some particular features of the Internet which should be taken into account when considering the development and presentation of metadata. The most important are mentioned below.

8. **The broad variety of users**
It can be assumed that users of statistical information on the net show wider variation in their ability to search, to understand and to use statistical information than “traditional” users of statistics. The audience on Internet is larger and more geographically widespread. What is obvious in a national context may be less so in an international one. Consequently, a wider audience may lead to greater requirements of metadata.

9. **The large amount of available information increases the need for efficient navigation and search**
The issues of search and navigation on the Internet are of greater importance than in other dissemination media. Some recommendations on how to organise

homepages and search facilities in a user-friendly way could therefore be useful.

10. **Easy linkage of information (e.g. hypertext)**
At the same time, the Internet provides new features for searching, linking and selecting information. Appropriate control tools for the organisation of web pages and the management of hyperlinks can thus improve the quality of dissemination.

11. **Providing statistics world-wide makes inconsistencies apparent**
The availability of statistical information from different sources makes for more visible methodological differences and inconsistencies. Consequently, relevant metainformation is needed to assess the comparability of statistical data.

12. **Updating is an essential quality**
Internet is a particularly efficient tool for rapid dissemination of data, and users expect data to be current. Frequent and efficient routines for updating both data and metadata are therefore important.

13. **The interactivity of Internet provides new opportunities**
For statistical offices, the possibilities to analyse user feedback and to monitor user behaviour on the Web site permit further improvement of the Web-site itself, as well as statistical data dissemination through other media. For users, the interaction with the information base (for example, through self-service in databases) and with data providers (by E-mail) provides better possibilities for obtaining the required statistical data.

IV. USERS OF THE INTERNET

14. A greater effort should be made in order to understand who are Internet users, since their requirements for metainformation may vary substantially. According to the targets of statistical agencies, the users of statistical information on Internet can be either internal (that is, statisticians responsible for the production of statistical information) or external users of statistical information.

15. Amongst the external users, subject-matter researchers, political decision-makers, public officials, executives, teachers, students, librarians, journalists and others can be identified. It is therefore essential that the statistical agency takes account of the variety of users.

16. Another way to classify external users is by their level of skill in statistics and their interpretation. Bearing this in mind, the following groups can be distinguished:

- **users with limited skills in statistical analysis** (general public);
- **skilled users** with limited inclination to read the metadata, e.g. preferring ready-made compilations;

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

17. Irrespective of relevance to a specific group, the user of statistical information on Internet in general needs metadata for the following functions:

- to see what data are available, to assist in the search for information;
- to interpret the information; and, if needed,
- to assist post-processing of the information (downloading and further application).

Relating to specific user groups, however, the scope of metadata needed for the above-mentioned functions may differ.

V. GUIDELINES FOR DIFFERENT TYPES OF METADATA

18. Based on the general purposes of metadata, distinction can be made between three main types, namely between those assisting search and navigation, those assisting interpretation, and those assisting post-processing of data.

19. However, it is very often not possible to make a strict distinction between metadata elements relevant for the different types of users (see para 16). Depending on the requirements of a specific user, it should be therefore possible for him/her to decide into how much detail to go accessing the metadata.

V.1. Metadata assisting search and navigation

20. It is important to be aware of how users set about searching, and how ability to search varies in relation to statistical information. Many users will search for statistical information linked to a specific problem or based on general interest. Users' ability to specify their need and their understanding of statistical terminology may be rather limited. Simple as well as varied search facilities are therefore important.

21. It is also advisable to provide hyperlinks between different levels of information (e.g., statistical tables, press releases and general information pages). The Web-site architecture should always allow users to understand where they are, find what they are looking for, and find out where they can go next.

22. **Metadata providing general information about the statistical Web site** These metadata use to be a part of existing statistical Web services, Their role is to help users to understand the general framework of the Web site. These include:

- Sitemap/table of contents of the Web site;
- Frequently asked questions, useful as a first guide for using the site;

- Site news announcing new features (a new document or services, or an important update);
- Descriptions of statistical subject areas;
- Description of the statistical institution (legal framework, organisational structure etc.);
- Description of the statistical system (role of different partners, etc.);
- Reference publications, product overview, general publications;
- Contact persons or e-mail addresses for more information;
- Release calendar;
- Links to other WWW statistical sites;
- Feedback facilities (via e-mail) allowing to order products and services available on the Web.

Such facilities can also make it possible:

- to invite users to express their views about a particular site;
- to obtain information from users;
- to create visitors= books and conferencing systems where users can send in their own contributions.

23. **Metadata assisting search.**
These metadata can be implemented with the help of the following:

- Press releases or other textual/tabular information giving an overview of a specific topic or subject area, with links (hypertext) to more detailed information will provide a good starting point for many non-skilled users;

- A list of key words linking everyday language to statistical tables/graphs available either directly on the Internet and/or in other electronic or printed/published form is a useful feature;

- Local search engine based on free text search - this is what many users expect. One should however be aware of the implementation problems, as it is important to establish links via synonyms from everyday language to technical terms often used in tables. Integration with a thesaurus is useful;

- A hierarchical subject matter classification is preferred by many users to be able to “drill” down to a specific subject area and table. Too many hierarchical levels are not efficient (maximum 3-4).

V.2. Metadata assisting interpretation

24. The requirement to supply information for interpretation will to some extent depend on the subject area, the type of information provided and the target user groups for any specific service. It is obvious that the needs may vary substantially between general users, media, researchers, etc..

25. However, following the general objectives, and taking into account the needs users of all skill levels, as much information as necessary to make a

correct interpretation and to avoid misuse should be made available. It may be important to supply information on the level of comparability with data from different sources.

26. A further challenge is to link data/metadata in such a way that users can choose the option they really need to use and interpret statistical information correctly.

27. It is therefore recommended that for any form of statistical data presented on the Internet (e.g., statistical indicator as a separate figure, graph or table) the following metainformation should be available:

Minimum set of metadata required for the correct interpretation of statistics:

- Title/content description (depending on subject area/content);
  Often including the following elements:
  - Statistical population;
  - Geographical coverage;
  - Observation unit;
  - Classifications and standards applied;
- Labels for rows/columns in tables and elements of graph;
- Definitions of labels;
- Measurement unit;
- Time reference/period;
- Regional units;
- Comparability over time (break in series, missing data);
- Footnotes highlighting specific precautions;
- Source of the data (agency compiling the data);
- Explanation of standard symbols in tables;
- Any information on copyright, restrictions of usage;
- Contact points for additional information;

Recommended metadata for better assessing the quality and comparability of statistics:

- Comparability with alternative sources;
- Links to press releases/summary of findings;
- Description of methods used in collection, revision, calculation; and estimation of the statistics;
- Information on error sources and accuracy of the statistics;
- Description of background and purpose of the statistics; concepts, variables and standards used.

V.3. Metadata assisting post-processing

28. This type of metadata should be included when statistical data are downloaded for further application.
There are two important requirements:

(i) Metadata listed under the 'basic standard' in para 27, should either be attached to the downloadable data, or should be easy to access and download in a separate operation. In addition, one may foresee some

other metadata, like possibilities/restrictions for re-use and combinations (e.g., linked to confidentiality, suppression of data, etc.).

(ii) Data/metadata should allow further processing using suitable tools, like spreadsheets, databases, packages for statistical analysis, packages for tabulation/graphic presentation/thematic maps. Larger tables should be easily downloadable. When putting data on Internet, statistical agencies should be aware of the possibilities and limitations of the different formats with regard to downloading of statistical data.

29. For example:

- Tables or text in picture formats (gif, jpg) are not practical for further analysis and processing.
- General and open formats (comma separated ASCII etc.) should always be available.
- Proprietary formats (spreadsheets, statistical packages) may be useful to simplify further processing and include basic metadata. When a proprietary format is used, reference to the software, or a possibility to download it should be also included.

VI. OTHER RECOMMENDATIONS

30. In general, it is recommended to adapt or develop some guidelines for each specific site concerning the design of pages, formats for downloading, etc.. The general Internet “Good practice guidelines” are valid but the special nature of statistical data should be taken into account. Studying comparable sites to find a fitting practice is useful.

31. **Management:** The statistical agency should consider management and cost issues linked to the establishment and updating of a Web site, including the allocation of resources.

32. **User monitoring:** Internet dissemination provides good possibilities for the analysis of users’ demands. Statistics can be developed on the usage of different parts of the service, and possibly also according to user types. It is possible to log the requests on the site, keep track of which navigation parts are followed, which keywords are used for search, etc. These statistics and feedback should be used for improving the Web pages, as well as for statistical data dissemination through other channels.

33. **Language:** Since a Web site addresses an international community, it is recommended that an English version should be available for a major part of the Web service, at least the Home page and some general information.

34. **Update information:** Updating is a critical part of a Web service. Automated, dynamic links to general databases are often preferable to ad-hoc manual updating of static pages. Date of last update should be indicated linked to the homepage and to specific information items/pages.

35. **What's new?** It is recommended to announce a new feature (a new document or service or an important update) on a regular basis. The page can be in a single language but must indicate the language versions in which any new document is available. The labels **"New"** and **"Update"** can be recommended only for documents or services of particular interest. The hyperlinks to a document should be included after the document is ready (hence no **"under construction"** message).

36. **"Title", "Classification"** and **"Keywords**". These are metadata designed to facilitate access to documents. Each web page should have a well defined title in order to allow the user to find the right page, after a certain number of steps, without using the “Back” button too many times. “Title” and “Keywords” are particularly important since they can be read by search engines, thus they serve to index a document. Since they are vital if a search engine is to find a document, their use can be highly recommended.

37. **Country/region name:** It is recommended that the country/region name be included in the title of the WWW homepage. In the World Wide Web environment, simply “Statistical Office”, “CSO” or text in national language is not sufficient.

38. **Graphics/special icons:** The users attention must not be distracted from the content of the information. Statistical data and text, especially tables, need simple background with no multicoloured patterns. Care should also be taken to ensure that images do not slow down the loading of a page onto screen.

39. **Area that can be displayed on the screen:** The format of the screen is different from the normal printed page, therefore just uploading documents in the same format as they are printed is not always sufficient. This applies particularly to large tables, where the user may loose text on rows and headings.

40. **Technology:** When choosing technology one should be aware of possible limitations on the user side for accessing pages using the most advanced solutions (Frames, Java Applets, ActiveX, etc.). Technologies adopted for Internet should not exceed the processing capacity of a relatively wide and sometimes unsophisticated user base.
