

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

Item 6 of the provisional agenda

**COMMENTS ON
“STANDARDS FOR STATISTICAL METADATA ON INTERNET**

Submitted by EFTA ¹

¹ Prepared by Jan Byfuglien.

Guidelines for metadata as part of statistical information: some comments on the documents “Standards for statistical metadata on internet” and “Guidelines for statistical metadata on the internet”

1. The objective of this note is to raise some issues concerning the document on “Standards for statistical metadata on Internet” (Working Paper No. 2) prepared by UN/ECE Secretariat (A). The document “Guidelines for Statistical Metadata on the Internet” (Working Paper No. 23) by Hans Viggo Sæbø and Svein Longva, Statistics Norway (B), which may be regarded as an alternative proposal, has also been taken into consideration.

2. In general, document (A) provides many useful observations and as such should form a basis for developing a final proposal. However, there is a lack of clarity in parts of the proposal, and it may in some cases be too far-reaching.

3. Document (B) takes a more pragmatic approach, resulting in a more limited range of guidelines. But this document may also need some revision to serve as guidelines.

4. Some of the issues, which may need further consideration and elaboration, are:

- What are the objectives of this “standard” and how should the “standard” be handled and followed up?
- Is the user analysis valid and, are the requirements as deducted from this, given right priority?
- Are the proposals clear and consistent enough concerning structure, terminology and concepts?

I. What should be the overall objectives of a “minimum standard”?

5. It may be useful to try to clarify the scope and objective of a possible standard in this field as a basis for identifying priorities and developing a more concrete proposal for guidelines or a standard.

6. Report (A) does not give any clear idea of what objectives the proposed “standard” should serve. Report (B) indicates that at present one should more talk about guidelines than “strict standards” and that one should have a procedure for revisions, partly because technology changes.

7. It is obvious that at present one is not preparing for a standard in a strict sense, and that it would be no obligation for anyone to follow the proposal. In order to be clear about this from the start any proposal should therefore be called “guidelines” or “recommendations”, and it should be part of the proposal to put in place a procedure for evaluation and follow up.

8. A part of the issue is also to what extent it is possible and/or useful to discuss “statistical metadata” in isolation from “the real thing”, and whether it at all is useful to talk about “statistical metadata” or rather metadata as part of statistical information? It is anyway important to focus on statistical information as an integrated data/metadata information product, and it may appear very difficult to agree on metadata terminology and standardisation etc. without some common concepts of the statistical information itself. The proposal should thus not limit itself too much to “pure” metadata, but to some degree focus on (basic) statistical information put on the Internet. Take as an example a Press release, which are useful and put on the Internet by many NSIs (even if not included in any of the proposals): is it data and/or metadata?

9. A tentative description of the objectives of the guidelines for metadata and statistical information on the Internet could be:

- To ensure wider dissemination and usage of public statistical information (that is by having a low introductory level as well as more advanced features).
- To ensure that statistics from different suppliers can be understood in a consistent way (that is by harmonising concepts and by applying much of the same functionality).
- To avoid misuse and misinterpretations of public statistics (that is by closely integrating data and metadata, and making metadata easily available).

II. Who are the users on the Internet, and are they different from other users?

10. Developing guidelines for metadata as part of statistical information on the Internet should further in principle start from an understanding of the user requirements - and user possibilities - in this particular context. Analysis of user needs/requirements is, in my opinion, one of the weak points of both proposals. However, this may not be a problem if one is able to identify and target the lowest and least sophisticated level of needs/requirements at this stage, and having a strategy for developing the guidelines for more sophisticated usage as experiences are collected.

11. Both reports start from the common assumption that the needs for metadata on Internet are much of the same as for any other medium. This may be true if the user composition is more or less the same as for other media, and if the technology does not have any major impact on the way data/metadata are searched and handled.

12. However, as also pointed out in the reports, there are some specific features of the Internet to be taken into consideration, and which in fact may lead to a somewhat different conclusion.

13. One important consideration is that the potential users on the net perhaps show wider variation in their ability to search, to understand and to use statistical information than “normal” users accessing printed publications. Their base level for understanding statistics may be lower than usual, partly because the net is reaching a wider international audience. The conceptual problems may thus also be larger as the data/metadata are targeting an international community. What is obvious in a national context may not be so in an international one.

14. The assumption that the requirements are more or less the same as for serving the general public (as understood so far) may therefore need further consideration when developing the guidelines.

15. The most basic and simplest access level on the Internet should thus be very elementary, and it may be necessary to give more basic meta information than in publications for the national market, and thus also make more extensive use of the technological possibilities.

16. The guidelines could therefore also be structured along these lines, starting from the elementary and basic needs and indicating more (possibly optional) advanced facilities.

17. The two categories of the user groups as given in report (A) (p. 2 and p. 4) may thus also be too broad, and a more in-depth discussion of user requirements taking into account user needs, user abilities and accessibility to Internet, may be useful for developing any metadata guidelines.

18. Some other assumptions presented, for instance that “metadata on accuracy/reliability to be less interesting for a majority of the users” may thus also have to be modified.

III. What types of metadata?

19. The different metadata types identified in report (A) are:

- Metadata supporting clients in searching for available data
- Metadata for interpreting the data
- Metadata to assess the quality of data

Report (B) uses the following categories:

- Metadata for interpretation
- Metadata for navigation and search
- Metadata for retrieval

Metadata for search

20. The type a) in (A) and type b) in (B) should more or less correspond, but the proposals partly differ, and a joint proposal will need some consolidation and clarification.

21. The possibility for key word search and a (“logic”/“hierarchical”) subject matter classification seems to be the main common element of both proposals.

22. It may, however, be necessary to elaborate somewhat what is meant by “key word” search. What are basic searchable elements?

23. It may also be pointed out that a hierarchical subject matter classification is an obvious requirement, but due to international accessibility (see point 2 above) one should seek harmonisation of the subject matter classifications used, and a proposal in this respect would be appropriate.

24. Report (A) stresses the need to “permit searching for the availability of data at the variable level”. This, however, raises a fundamental problem, as the “variable” concept is not specified.

The “variable” concept is further neither clear from several metadata initiatives undertaken and/or under way, and it may thus not be productive to give a proposal in this direction at the moment.

25. One may rather specify that search more should be possible on the content of the available statistical information (which seems more in line with the subject index of statistical yearbooks). This will mean more the type of bibliographical information or a thesaurus, which not necessarily links to “variables”. This may even included references to available printed products.

26. Under this point one may also discuss and include more general information such as release calendar, publication lists and ordering, organisational overview, legal framework and link for instance to main indicators and press releases (which neither of the reports mentions).

Metadata for interpreting the data

27. The two reports differ also when specifying metadata for interpretation.

28. Report (B) takes a traditional statistical table as a point of departure, which is a pragmatic approach, as metadata should be linked to and integrated with an actual table or graph, containing the information. The report does not include the more detailed background information (concepts, definitions, estimation etc.) as part of the “absolute minimum requirements” on the basis that this is not included in general publications such as a statistical yearbook (at least in some countries). This reasoning may be questioned on the basis of the user needs, as well as on the basis of the possibilities of the technologies.

29. Report (A) is somewhat more theoretical, and possibly too ambitious on this point. It is further lacking some important metadata elements such as footnotes, data source and regional level. The report has also introduced a separate category for “metadata to assess the quality of data”, which rather should be seen as part of metadata for interpreting data. And it is somewhat strange that it is argued that “metadata items for accuracy/reliability should be minimized” as such information often is crucial to make correct interpretation and usage of the information, which should be in the interest of all NSIs, also when using the Internet as a dissemination medium.

30. The conclusion on this point is that the list of metadata items as presented in Report (B) can be used as a basis, but one should consider to add some items in order to cover more information important for the interpretation of statistical tables, such as concepts, definitions etc. (which also more or less would cover point 3.4 of report (A)).

Metadata linked to downloaded data

31. Report (B) proposes a separate list of metadata items under “Metadata for retrieval”. This is reasonable, as it is important to maintain the link between data/metadata in downloaded tables. The list should provide a good basis for further elaboration/clarification.

Other issues

32. There is some need of consolidation between the two reports under this general point. Report (A) may give some unnecessary details at this stage. Some recommendations on design issues etc. could be dropped or perhaps put in an annex to the guidelines.