

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 2
English only

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Metadata (METIS)
(18 - 20 February 1998)

Item 6 of the provisional agenda

STANDARDS FOR STATISTICAL METADATA ON INTERNET

prepared by

UN/ECE Secretariat¹

¹ The material is based on contributions received from Canada, Sweden and OECD

1. Statistical data on Internet - current status.

At present, more and more Statistical Offices are looking at Internet as a means to disseminate statistical information. Currently, out of the 55 member countries of the ECE, 34 have the homepages of their Central Statistical Offices available on Internet. (Here the United Kingdom and United States are both counted as one entity. In fact, there are 7 different Web pages from statistical offices from UK and 17 from USA dealing with different areas of statistics. Regional statistical offices from parts of a country are not included.)

30 of the above mentioned Web sites make available some statistical data, mostly general figures, or online news releases and statistical publications. From these 30, 11 countries allow access to statistical databases and provide some database searching capabilities. These are: Austria, Canada, Germany, Israel, the Netherlands, Portugal, Spain, Sweden, Switzerland, United Kingdom and United States. This overview was made in December '97 (see also the Annex). Many Statistical Offices are in the process of developing their Web pages, therefore the situation is changing constantly.

Taking into account the development of Internet (according to some estimations the amount of information on Internet doubles in every 2 months), there is no doubt that more and more statistical offices will establish their home pages on Internet, making more and more statistical data available to the users. These statistical data have to be accompanied with adequate metadata. What kind and how much metadata would be adequate, and how should it be organized? When deciding that, the Statistical Office has to keep in mind that all the information put on Internet can, and will be, accessed from all over the world.

2. Internet specifics.

Although the requirements for statistical metadata on the Internet are not substantially different from the needs for statistical metadata in any other media, Internet has several specific features that should be taken into account. The most important of them are mentioned below:

(i) Internet is a common means to access statistical data for many different users with different needs for metadata.

It would be reasonable to distinguish between two broad groups of users: the general public (including mass-media) who are interested in aggregate data on a general level, and researchers, subject-matter statisticians etc., who are intensive users of detailed statistics. The needs of those groups for statistical metadata differ accordingly - ranging from a minimal set of metadata items (indicator, measurement units, time, geographical coverage) to the definitions of concepts, methodologies, inclusions and exclusions in variables, relation to major aggregates, classification systems used etc.

(ii) The large scale of usage makes more apparent the existing weaknesses in data, e.g. lack of coherence across different data sets.

The vast amount of statistical figures accessible through a common path and originating from different sources in different countries makes more visible the lack of statistical coordination and integration. The harmonization of concepts, streamlining of statistical definitions and classifications is essential in order to get

comparable statistics. The need for coherence is valid both inside a country, between the statistical office and other institutions, and on international level, between different countries and international organizations.

(iii) Properly organized metadata can facilitate the search for statistical data on the Internet.

Internet technology allows the organization of metadata in a way that facilitates the search for statistical data. However, it can be very time consuming and cumbersome to find the required data with the help of current Internet search mechanisms. The keyword search does not guarantee that the user will find actual statistical data on the requested topics. Also, the search engines do not access the most recent information.

In putting statistical metadata on the Internet, it must be kept in mind that metadata should permit searching for the data based both on keywords (that can be accomplished partially via Internet search engines) and on themes or subjects. It should be possible to search for all or selected fields of metadata. The user needs to be able to link variables to the themes or subjects, and to drill down from a subject to the variable level. Internet technology combined with proper organization of statistical metadata makes it a manageable task. As statistical metadata on the Internet is displayed continuously, it should also be updated regularly.

(iv) Homepages of statistical offices are one of the most efficient ways to access statistical data on the Internet.

The Internet user is confronted with the diversity of structures and organization of statistical data and metadata on the Internet. One of the easiest ways to find statistical data is through the WWW homepages of national statistical offices. At present, many offices are in the process of establishing their Internet homepages, and putting statistical data on the Internet. It would be useful for the statistical offices to have common guidelines for organizing those homepages. At the same time, it would assist the users in finding the requested information, if data from the national statistical offices was organized according to a common pattern, and the user would not have to waste time on getting to know the system.

(v) Metadata and structured organization of statistical data on Internet can be used as a tool to implement statistical data dissemination and pricing policies.

Internet makes it possible to organize statistical data (and metadata) hierarchically, so that different users can access different aggregate levels of data or metadata. For example, the most aggregate data can be accessible free-of-charge, and the user has to pay to get access to more detailed data and metadata. Or, the statistical office can offer free access to raw data, and to charge for the data that is worked upon (results from an analysis, aggregation etc.). Well organized and well presented statistical data on the Internet can also be a means of improving the public image of a statistical office, and of making the general public and government aware of their work.

The question whether to put statistical data and metadata on Internet is part of the data dissemination policy of a statistical office. Within this framework, it needs to be decided what data, when and how it is made accessible through Internet, and the person responsible for it.

3. Requirements for statistical metadata on Internet

The statistical metadata requirements for Internet are not essentially different from what metadata is needed for sensible use of statistical information, regardless of the medium of its presentation.

3.1. Users and categories of metadata

The minimal set of required metadata should be defined in relation to the needs of individual metadata users. Consequently, the following categories of metadata could be considered:

- metadata supporting clients in searching for the existence and availability of information;
- metadata used by clients in interpreting and properly using the data found;
- metadata used by clients in evaluating the quality of the data available to meet their needs.

It is also perhaps useful to distinguish two broad types of clients of statistical information and, hence, of the associated metadata, as follows:

- clients that want to know what ready statistics are available to quickly meet their need, in what form and at what price.
- clients who are seasoned researchers and intensive users of detailed statistics. These clients, including those inside statistical agencies charged with analysis and harmonization tasks, want to be able to explore the full underlying richness of the master retrieval and/or public use microdata files.

According to client surveys conducted by Statistics Netherlands (OECD, 1995) the characteristics of statistical products considered most important were: timeliness, comparability, available detail, continuity and the availability of definitions and other metadata. Analytical articles, reliability and layout were of less importance. Asking clients' opinion on statistical products that they actually used, showed a discrepancy between what clients want and what they get. Client assessment of reliability/overall quality and continuity was quite positive, but their opinions on timeliness, comparability and available definitions/metadata were less positive.

Surveys by OECD's Statistics Directorate (see OECD, 1995) confirmed that clients want to see improvements in timeliness and availability of metadata, directly linked to the data. The conclusions are valid for statistical data dissemination generally, and should be kept in mind also when discussing the metadata requirements on Internet.

These client surveys show that:

- there is an urgent need for improved metadata
- this need is not limited to a small subset of 'sophisticated' clients
- assessing quality/reliability is not the first thing on the mind of the 'standard' client.

The distinction between 'standard' users and 'sophisticated' users does not generally coincide with the use of aggregate/macro data versus the use of detailed/micro data. Many seasoned researchers only use macro-economic aggregates, but they still want to understand

fully what the limitations and strengths of the data are, including comparability in time and space. There are, on the other hand, many 'standard' users that look at detailed data, for example to see how the trade in a very specific commodity affects their business.

Therefore, it would be reasonable to base the minimum metadata requirements on the needs of 'standard' users. This means that we should list metadata which make it possible to find/access useful statistical data and to understand what these data mean (the availability and content dimensions in UN/ECE, 1995). Metadata on accuracy/reliability seem to be less interesting for a majority of the users.

3.2. Metadata supporting clients in searching for data

The key factors to be considered with respect to metadata for searching are:

- the need to organize the metadata to permit searching for the availability of data at the variable level
- the need to provide both key word and logical (based on themes or subjects) searching capabilities
- the need to link the variables to those themes or subjects to facilitate searching by subject in a way that permits drilling down to the variable level
- the need to link the variables to databases, products and services that contain those variables (or aggregations thereof)
- the need to be able to search all or only selected fields of metadata

Fortunately, the Internet technology, with its hypertext linking capability, facilitates the sort of searching described above. Repositories of metadata can be established in such a way that certain cells or fields of information are understood to be links to other cells or fields of metadata without hand crafting the links for each application.

This can work via a table of contents which is organized by themes/subjects that are linked to individual variables. References to selected relevant statistical outputs, at least at the level of themes/subjects, could also be included.

Additional metadata requirements may be needed at the availability/searching stage. Some of them may also be relevant for the contents dimension:

- periods covered by the data (for time series)
- release calendar
- references to corresponding methodological information on other media
- contact point

3.3. Metadata for interpreting the data (metadata on contents)

Several metadata-items can be presented jointly for micro and for macro data. They include:

- a description of the population
- an explicit description of exclusions (e.g. with respect to an international standard description for a population) may be more informative than a more neutral description of the population
- a description of the target object/ reporting unit
- definition of variables
- definition of measurement units
- a description of dimensions/cross classifying variables
- a description of corresponding classifications/codelists

- a description of the relationship between these classifications/codelists and international standard classifications, explicitly listing departures from international standards (where applicable)
- reference period
- timeliness (time period between reference period and publication)
- frequency/periodicity of statistics

3.3.1. Metadata for microdata

Metadata for microdata should be organized according to the proposed documentation template for an observation register (UN/ECE, 1995, figure 3.3), or at least in some other similar way containing the most important parts.

The metadata items specific for the interpretation of microdata can include:

- a. variable descriptions (questions asked, etc; questionnaire may be enclosed in order to give a more precise definition)
- b. physical organization and data set descriptions (record layouts)

3.3.2 Metadata for macrodata

The metadata needed for the interpretation of macrodata should be described according to the box structure (see UN/ECE, 1995, section 1.2.2).

The proposed structure consists of the following parts:

- a. Descriptions of the statistical population and selections within the population (scope of the data set).
- b. Descriptions of the cells in the structure (array), i.e. definitions of the variables which are estimated as well as of the estimations (sum, mean value, etc.) and measurement units used.
- c. Descriptions of the cross classifying variables and corresponding classifications (code lists).
- d. Descriptions of the time parameters (time periods, measurement time, etc).

Macrodata are derived from microdata by one or more aggregation functions. So for macrodata there is also a need to describe:

- aggregation/grossing up methods
- adjustments (e.g. seasonal adjustments)
- other manipulations with the data

3.4. Metadata to assess the quality of data

The number of metadata items for accuracy/reliability should be minimized. Some important items, which may belong in this category are the following:

- responsible agency (often called “Source” at the bottom of statistical tables). This item does not belong strictly only this metadata category, but the indication of a reputable agency is for many clients the best indicator of reliability of the statistics presented.
- measurement instrument/media (e.g. administrative sources, enterprise surveys, mail surveys of persons, face-to-face interviews of persons).
- comparability over time (for time series), including a description of breaks in series and missing data together with their explanations.
- comparability with alternative statistics (e.g. collected via a different measurement

instrument) for the same variable. An explanation of major differences and their causes would be useful.

- an overall description of overall accuracy, or the overall error, would of course be ideal for the users. However, many statisticians are reluctant to commit themselves to making/disseminating statements which are partly judgmental, even if they are in a better position to make such judgements than anybody else.
- lacking a summary description of the total error, a listing of error sources and their possible impact on the accuracy of the statistics can be useful to some users. Most users will, however, shy away from trying to make an overall assessment themselves, especially if the information about error sources is incomplete and heavily focused on a few easily quantifiable, but not necessarily important, error components. Instead these users feel justified to use the agency stamp (first item on this list) as a guarantee for reliability.
- various items in figure 3.3 of UN/ECE (1995), referring to the survey plan, the data collection and the statistical processing could also assist the very committed clients to make their own assessment of accuracy/reliability. Given the probably very small number of such diehards, this does not seem to be essential for our minimum requirements.

4. Some general considerations

In establishing standards, it is reasonable to take into account the following **general aspects**:

- Standards for metadata on the Internet should as far as possible be based upon existing standards for other kinds of presentation media.
- Technical standards should be discussed later on. One alternative could be to standardise according to SGML (Standard Generalised Markup Language).
- The standards should be valid for static as well as for dynamically produced tables on the Internet

Taking into account that whatever is put on the Internet is accessible to users from the whole world speaking different languages, some basic recommendations that seem to be self-evident but are not always followed:

a) the Internet language is English

therefore it is recommended make accessible also the English translation of the text as much as possible, otherwise the use of the data is limited.

b) indicate country/region name (preferably in English) in the WWW homepage title.

When working with Internet, the Web page title is displayed always on the top of the Internet window on computer screen. When using the bookmarks, the Web page title is displayed instead the full Internet address. It is not sufficient for the users to see there only "Statistical Office", "CSO" or text in national language.

c) indicate whether the Web pages contain statistical data (figures)

It might be useful to work out some classification of types of statistical data and metadata on the Internet, and to indicate this type as high in the hierarchical structure of the Web pages (table of contents) as possible. E.g.:

- metadata (in broad sense: information about available publications, statistical methodology etc.),
- report (text, possibly including figures, tables and charts, e.g. a news release),
- chart,
- table (containing statistical data),
- database (access to a statistical database, dynamic Web pages).

d) include a table of contents of the Web-site,

preferably indicating the topic/subject, data type (report, table etc) and time period covered.

e) indicate the date of the last update,

as high in the hierarchical structure of the Web pages as possible, preferably in the table of contents. If possible, indicate also the advance release times.

In order to make it easier for the users to work with statistical data and metadata on Internet, and to save users' time and energy, some more recommendations of a general nature:

- attach a local search mechanism, it helps users to target the search on your Web-pages, and also because the most recent info is not accessible through global Internet search engines;
- make navigation between Web-pages possible at any moment (e.g. with the help of frames, buttons);
- if there is a fee required, indicate this as high in hierarchical structure as possible (in table of contents);
- include possibility to save in table format; for downloadable files include the type and size; every downloadable piece should include the minimum set of metadata;
- keep the background simple, without multicolored patterns, and text in contrasting colours;
- do not use too many detailed images - the access time through Internet is too long.

REFERENCES

IMF (1996) "Guide to the Data Dissemination Standards", Washington, D. C., May 1996.

KELLER, Wouter J., Kalvelagen, Erwin M., Bethlehem, Jelke G. (1995) "Statistics on the Internet" (*paper prepared for the Conference on New Technologies in Statistics, Bonn, Germany, November 1995*)

OECD (1995), "Users' Views on Short-Term Indicators" (*paper presented at the OECD Ad Hoc Expert Group on Main Economic Indicators, 16-17 October 1995, Paris*)

PETIT, Gerald Pierre Beziz and Rob van Eck (1996), "List of Metadata Items for OECD's Main Economic Indicators" (*paper presented at the UN/ECE Work Session on Statistical Metadata, 22-25 October 1996, Berlin*)

UN/ECE (1995), "Guidelines for the Modelling of Statistical Data and Metadata". Methodological material. United Nations, New York and Geneva, 1995.

Statistical data and metadata from national statistical offices of the ECE member countries and selected international organizations on Internet

December 1997

COUNTRY	NSO Web page	Regional stat. Web pages	WWW address	Statistical data included	Language (E - if English text is available)	Restrictions to access (fee or subscription required)	Dynamic Web page / access to databases	Last data/last modified	Notes/comments	IMF DSBB available	Metadata submitted to OECD available on Internet
Austria	1		http://www.oestat.gv.at/	yes	G	yes	yes		English version coming soon	+	+
Bulgaria	1		http://www.acad.bg/BulRTD/nsi/i	no	E			94/Oct. 95	1 general table.; no change after Oct. 95		
Canada	1	8	http://www.statcan.ca/start.html	yes	E	yes	yes	Oct97/Dec97	site map included; navigation possible at any moment, good metadata; list of table/table of contents; database CANSIM - Canadian scio-economic inf. management system, trade data online (for a fee); concepts and definitions separately; release dates;	+	+
Croatia	1		http://www.dzs.hr/	yes	DSBB English, other Croatian					+	
Cyprus	1		http://www.pio.gov.cy/dsr/	CPI	E			97	CPI, some reports in text from (data from '95 and '96); catalogue of publications		
Czech Republic	1		http://infox.eunet.cz/csu/csu_e.html	no	E			95?			+
Denmark	1		http://www.dst.dk/internet/startuk.htm	yes	E			Oct97/Nov97		+	+
Estonia		1	http://www.ee/epbe/datasheet/index.html	yes	E			Oct97/Nov97	National Bank datasheets; mostly financial data		
Finland	1		http://www.stat.fi/sf/home.html	yes	E			end 96/Nov.97	economic trends in the form of charts; plans to create dynamic gateways to databases?	+	+
France	1		http://www.insee.fr/va/index.htm	no	E				some statistical data in online publications (e.g. abridged version of "French economy in 1997/98", "Conjoncture in France")	+	+
Germany	1	11	http://www.Statistik-bund.de/e_home.htm	yes	E/G	yes	yes	Oct97/ Dec97	database in German	+	+
Greece	1		http://thales.iacm.forth.gr/esye/	yes	E			94/?			+
Hungary	1		http://www.ksh.hu/eng/homeng.html	yes	E			Sept 97/ 97	national data by 22 categories	+	+
Iceland	1		http://eldur.stjr.is/hagstofa/	yes	E			Oct97/Nov97	"Statistics Iceland" - full publication online for free		+
Ireland	1		http://www.cso.ie/	yes	E			Oct97/Nov97	Statistical reports, graphs, key economic indicators	+	+
Israel	1		http://www.cbs.gov.il/engindex.htm	yes	E		(yes)	Oct97/Nov97	"Monthly bulletin of statistics online", IMF national accounts data	+	
Italy	1		http://petra.istat.it/	yes	Italian			Oct97/Nov97	publications online, zip-files can be downloaded, no direct data; interactive map!	+	+
Latvia	1		http://www.latnet.lv/ligumi/CSBL/	yes	E			Oct97/Dec97	press releases, basic socio-economic indicators		
Lithuania	1		http://www.std.lt/	yes	E			?/97	online publications	+	
Luxembourg	1		http://statec.gouvernement.lu/	yes	E			96/no date	Luxembourg economic portrait		+
Malta	1		http://www.magnet.mt/home/cos/	yes	E			June97/97	press releases, main economic indicators etc.		
Netherlands	1		http://www.cbs.nl/indexeng.htm	yes	E/N		yes	Nov97/Nov97	Dutch ec. indicators in pdf format, Statline (over 24000 visitors)	+	+

**Statistical data and metadata from national statistical offices of the ECE member countries
and selected international organizations on Internet**

December 1997

COUNTRY	NSO Web page	Regional stat. Web pages	WWW address	Statistical data included	Language (E - if English text is available)	Restrictions to access (fee or subscription required)	Dynamic Web page / access to databases	Last data/last modified	Notes/comments	IMF DSBB available	Metadata submitted to OECD available on Internet
Norway	1		http://www-open.ssb.no/www-open/english/	yes	E			Oct97/Nov97	statistics by subject, publications "Weekly bulletin", "Economic survey", "Yearbook" online	+	+
Poland	1		http://www.stsp.gov.pl/						Experimental web-site	+	+
Portugal	1		http://www.ine.pt/	yes	E/Portuguese	yes	yes	Oct97/97	access to INFOLINE for subscribers		+
Romania	1		http://www.kappa.ro/clients/cns/	few	E			96/April97	some very general figures and general information about Statistics Romania		
Russian Federation	1		http://www.fe.msk.ru/infomarket/ewelcome.html								
Slovenia	1		http://www.sigov.si/zrs/index_e.html	yes	E			Oct97/Nov97	Short-term economic indicators, "Yearbook" in pdf files	+	
Spain	1		http://www.ine.es/	yes	E/Spanish		some DB	Oct97/Nov97	TEMPUS DB, 350 000 time series, 1 selection possible at a time		+
Sweden	1		http://www.scb.se/indexeng.htm	yes	E	yes	yes	Oct97/Nov97	DB will be available in English in 1998	+	+
Switzerland	1		http://www.admin.ch/bfs/eindex.htm	yes	E	yes	yes	Nov97/Nov97	well structured homepage; key data corresponds to IMF DSBB; online access to STATINF (in French and German) through Telnet since 19 Nov. 1997)	+	+
Turkey	1		http://www.die.gov.tr/ENGLISH/index.html	yes	E			Nov97/Nov97	Results from several surveys, press releases, includes data dictionary (definitions); (plans to allow access to databases?)	+	+
United Kingdom	1+7	2	http://www.emap.com/ons97/	yes	E	yes	yes	Nov97/Dec97	Access to the official system of statistics through ONS; ONS databank online ; free-of-charge general data	+	+
United States	17		http://www.fedstats.gov/	yes	E		yes	97/Dec97	centralised access to official statistics through White House briefing room or FEDSTATS www-site	+	+
Uzbekistan											
Yugoslavia	1		http://www.szs.sv.gov.yu/homee.htm	few	E			96/97	Experimental server, some data available from 1996		
Web pages of National Statistical Offices available: 34 of 55 countries											
in 30 statistical data included											
in 11 some database searching capabilities											

**Statistical data and metadata from national statistical offices of the ECE member countries
and selected international organizations on Internet**

December 1997

COUNTRY	NSO Web page Regional stat. Web pages	WWW address	Statistical data included	Language (E - if English text is available)	Restrictions to access (fee or subscription required)	Dynamic Web page / access to databases	Last data/last modified	Notes/comments	IMF DSBB available	Metadata submitted to OECD available on Internet
International organizations										
CIS	1	http://www.unece.org/stats/cisstat/mainpage.htm	yes	E	yes	yes?	June97/97	DB "Statistics" online to subscribers?; main macroeconomic indicators for CIS countries		
Eurostat	1	http://europa.eu.int/en/comm/eurostat/eurostat.html	yes	E	yes	yes	Dec97/Dec97	The main EU statistical indicators online, Eurobases - access can be obtained via host (mostly commercial DB-s)		
FAO	1	http://apps.fao.org/	yes	E		yes	Nov97/Nov97	Faostat databases		
ILO	1	http://www.ilo.org/public/english/index.htm	no	E				Bibliographic databases on national laws, terminology etc.		
IMF	1	http://www.imf.org/	no	E				Dissemination Standards Bulletin Board; links to some national summary data sites		
IEA (International Energy Agency)	1	http://www.iea.org/	yes	E		yes	Sept97/Dec97	Key indicators by countries for 1994, connection to OECD datasets; monthly surveys in PDF files		
ISI (International Statistical Institute)	1	http://www.cbs.nl/isi/	no	E						
OECD	1	http://www.oecd.org/statlist.htm	yes	E	yes	yes	Dec97/Dec97	"OECD in figures" online, short term indicators etc.; OECD hot file - password protected key economic indicators, updated weekly		
UNESCO	1	http://www.unesco.org/general/eng/stats/stat.html	yes	E			94/?	Statistical yearbook 1996 online		
UNIDO	1	http://www.unido.or.at/start/services/statistics/navigator.html	yes	E			95 / Oct97	Selected statistical tables on comparative economic and industrial performance		
UN Statistical Division (New York)	1	http://www.un.org/Depts/unsd/mbsreg.htm	yes	E	yes	yes	/ Dec97	Publications online (several for subscribers only), social statistics estimates for 1997, annual report		
WHO	1	http://www.who.ch/whosis/whosis.htm	yes	E		yes	94/ Aug96	"Health for all" European database online (last available data for 94, some estimates for 95)		
World Bank	1	http://www.worldbank.org/html/Welcome.html	no	E				Some publications online in pdf format, including Annual report		
WTO (World Trade Organization)	1	http://www.wto.org/	no	E						
WTO (World Tourism Organization)	1	http://www.world-tourism.org/esta/statserv.htm	yes	E	yes	yes	95 / 97	Tourism statistics		