

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Metadata
(Washington, D.C., United States, 28-30 November 2000)

Topic (ii): Metadata modelling and terminology issues

**TOWARDS RECOMMENDATIONS ON FORMATS RELEVANT TO THE DOWNLOADING
OF STATISTICAL DATA FROM THE INTERNET**

Submitted by Statistics Netherlands and U.S. Bureau of Labor Statistics ¹

Contributed paper

I. INTRODUCTION

1. This report presents the problems involved with the choice of a format for documents that make statistical data available on the Internet. The aim is to start a discussion among participants at the METIS 2000 Work Session, in order to enable the formulation of recommendations approved by the community represented at METIS.

2. The paper was prepared by Daniel Gillman (U.S. Bureau of Labor Statistics) and Jean-Pierre Kent (Statistics Netherlands). This work was backed up with useful comments by Håkon Berby (Statistics Norway), Marco Pellegrino (Eurostat) and Roger Dubois (Eurostat).

II. MOTIVATION

3. The paper was requested by the Conference of European Statisticians, as a complement to a previous METIS paper entitled "Guidelines for statistical metadata on Internet". Paragraph V.3 "Metadata assisting postprocessing" was considered incomplete without such recommendations.

4. The authors judged that the problem was not only a question of individual competence, and should be defined in the context of existing policies. It was decided that the current paper would present the problem and state the questions to be asked in order to formulate the requested recommendations. The final paper will make use of the answers to these questions. Participants are kindly requested to answer the questionnaire in their respective offices in order to provide the input for the final paper.

III. HISTORY

5. Before the advent of the computer and electronic storage media, statistical organisations disseminated data on printed paper in the form of tables and reports. Users of statistical data needed to acquire these tables and reports either from the statistical organisation itself or through some intermediary such as a library or data archive. Access to respondent data (microdata) was difficult, if not impossible.

6. Things changed considerably after computer use became widespread. First, magnetic tapes were used to store and transfer data. Tapes were heavy, did not hold much data (by today's standards), and had

¹ Prepared by Jean-Pierre Kent, Statistics Netherlands, and Dan Gillman, U.S. Bureau of Labor Statistics.

low reliability and life span. The introduction of the PC and the subsequent miniaturisation of computer components gave rise to the diskette and later the CD-ROM.

7. Computer networks came into use at the same time as the early development of the PC. Networks allow many computers to be interconnected. The transfer of data through networks greatly simplifies that activity, because magnetic tapes, diskettes, CD-ROMs, and similar media all require people to effect a transfer. Networks make purely electronic transfer of data possible.

8. The use of networks increased substantially throughout the 1980s and into the 1990s with the development of Local Area Networks (LANs) and Wide Area Networks (WANs). LANs were implemented so computers used within an organisational unit could communicate, share data and use common software. Large organisations with many sub-units developed WANs so that computers connected to different LANs could communicate with each other.

IV. WORLD WIDE WEB

9. The ultimate WAN is the Internet, which interconnects computers, WANs, and LANs throughout the world. This development has greatly increased the capacity for users to access data from remote sources. Statistical agencies have benefited as a result, and now users routinely download data (tables, reports, documents, etc.) from their sites.

10. The two most widely used methods for transferring data over the Internet are File Transfer Protocol (FTP) and HyperText Transfer Protocol (HTTP). FTP is primarily meant as a method for transferring files to or from a remote computer. The user needs to invoke some other system to read the file after downloading it. HTTP is the method for transferring data across the World Wide Web. Specially formatted files, i.e. files formatted and tagged in HyperText Markup Language (HTML), are transferred to the user's machine where the HTML is interpreted by browser software, e.g. Netscape or Internet Explorer, and the contents of the file are displayed on the user's screen.

11. Web browsers are becoming more sophisticated as time goes on. Browsers interpret files based on a mime-type as presented by the server. Usually, this type is determined by the server on the basis of file extensions, but that is not always the case. Different mime-types are assigned for files of the same extension depending on whether download or display is allowed. Each type induces different behaviour from the browser. Display behaviour by the browser is implemented through plug-ins. MS-Word (*.doc) files are mapped to the mime-type application/MSWord by the server. At the client side, in the browser, the mime-type is recognised and the plug-in launches MS-Word to read the file.

12. The advantages of this are enormous to the user with the right kinds of plug-ins. Users without the correct plug-ins must resort to downloading files or using FTP to transfer files to their machines. Then they must launch the correct applications themselves in order to read the files.

V. THE FUTURE

13. At present, technology is developing at a rapid pace, and the near future will offer new possibilities. WAP (Wireless Application Protocol) allows Web access through mobile telephones. E-book (electronic book) makes it possible to download large documents and read them offline on a machine that emulates the dimensions and functionality of a book. It is not clear whether these access forms will mature and take hold of the market, or whether they will be quickly superseded by other approaches. It is clear, however, that the PC as we know it at present will not be the only means of access to documents on Internet.

14. The ultimate Internet access could well turn out to be a system in which the user need not do more than specify his or her subjects of interests. Pieces of software currently known as "smart agents"

will keep browsing the network for new information, extracting whatever is relevant, and presenting it in the user's favourite form.

VI. FROM FILES TO DOCUMENTS

15. The discussion so far is focused on files. Originally, the Web contained static files (pages) only. The emergence of active protocols such as the Common Gateway Interface (CGI), and Active Server Pages (ASP), made it possible to generate HTML pages dynamically. It was not long before dynamic Web pages were created from the output of databases. Now, with the development of Java, JavaScript and ActiveX, even the Web browser interfaces can be made as flexible as Windows.

16. In spite of the advantages of these techniques, there is a potential problem. In the current state of technology, search engines are only capable of finding text in files that are present at the time of the indexing activity. So whatever text is created on the fly in browse time is not accessible to these search mechanisms. A site providing dynamic content is compelled to provide its own specific search engine, which defeats the purpose of generic search engines, i.e., find content without knowing where to look for it.

17. An important paradigm shift since the advent of the Web is to consider all files as documents. In the past, the accent was on form: files were accessed through file descriptions. The structure and format of a file was all important. The content was considered to be implicitly understood by humans or irrelevant for software. Nowadays we rather refer to documents, stressing more the structure of the contents. A document can consist of one or more files. HTML is particularly suited for this document approach, because it can tie different files together through hypertext links. The location of the files forming one document is irrelevant in this concept. They can be distributed over different servers all around the world.

18. Most recent is the development of eXtensible Markup Language (XML) for use on the Web. XML was developed to replace HTML, although browsers do not interpret it very well yet. An immediate advantage of XML is that content and layout can be defined independently. Where HTML tags primarily define document structure in terms of layout sections, XML tags are used to define the structure of the content. This makes it possible for software to access the document in a meaningful way. The layout is defined in a separate XSL script (XML Style Language). This makes it possible to vary the layout without affecting the document structure.

VII. OTHER TECHNIQUES

19. The Web and its document handling capabilities are only one way to enable users to download data from statistical offices. As mentioned earlier (10) FTP is simple and easy to use, and it does not require the Web. Other possibilities include direct access to specialised databases or the use of electronic data interchange languages and formats, such as GESMES (GEneric Statistical MESsage).

VIII. FORMATS ON INTERNET

20. A number of formats are appropriate for documents on the Internet. They can roughly be divided into two categories: the document oriented formats, and the application oriented formats.

VIII.1. Document oriented formats

21. The two main formats for documents on the Internet are HTML and XML. Both are well suited for browsing. Additionally, XML can be accessed for specific purposes by software that understands the meaning of the user-defined tags. Their main advantage lies in the fact that they can be accessed by browsers: whoever has access to Internet usually also has a browser installed on the computer. For documents intended to be downloaded, however, hyperlink tags should be avoided. There is no guarantee

that a document consisting of a number of interlinked files will be completely downloaded and correctly accessed locally.

VIII.2 Application oriented formats

22. Alternatively, documents can be stored on the Internet in a format specific to an application. Such a document, of course, can only be accessed on a computer on which the required application has been installed. This can be an obstacle to the accessibility of the document. A general advantage of application oriented formats is that a document is usually fully contained in a single file, ensuring the integrity of the downloaded version. Hereafter, a few application formats will be considered.

VIII.3 Read-only formats

23. The author of a document would usually prefer that his text not undergo modifications by third parties. Content and form integrity cannot be guaranteed by applications with editing functionality. This is what makes read-only documents interesting. Such documents are produced by one piece of software (the generator), and accessed by another (the reader). PDF (Portable Document Format) is the best known read-only format. It is accessible through Acrobat Reader, which is free of charge. A read-only format protects a document against inadvertent or naive changes. Malicious tampering, however, can only be excluded through certification algorithms based on encryption methods. PDF does not support such certification. Neither does any other widely used format on Internet.

VIII.4 Word processor formats

24. Documents mainly consisting of text can be published in a format used by a word processor. The main word processors are MS-Word (Microsoft) and WordPerfect (Corel). They can read and write files in each other's formats without causing disturbing biases. They can both read and write in Rich Text Format (RTF), a format defined for application independent interchange of documents.

VIII.5 Spreadsheet formats

25. Documents containing cross tabulations or accounting records can be published in a spreadsheet format. This is also suitable for flat databases. The main spreadsheet applications are Excel (Microsoft), Quattro (Corel) and Lotus (IBM).

VIII.6 Database formats

26. It will not really be practical to make complex database material available for downloading. The preferred method is to install a searchable database on the server, and allow the user to select the material to be extracted. This can subsequently be presented in HTML or XML pages created on the fly. It is, however, also possible to publish limited material in a database format. The formats most in use for small databases are Access (Microsoft) and Paradox (Corel).

VIII.7 Unknowns

27. Knowledge of available formats is not sufficient for selecting a preferred format for statistical data. The purpose of the document, the needs of the users, and various preferences of the office play a role in this choice.

28. Let us start with the user parameters. We are interested both in the user's wishes and needs and in the technology at his or her disposal. Often, the best format for the user is determined by his needs. For example, if the user simply wants to print a copy of a report, then PDF format is probably best. If the user wants to analyse data, then a spreadsheet or database format is best. Obviously, if the user wants to

analyse data, then software for accomplishing this must be available to him, too. Therefore, the user's needs must match the formats and software available to him.

29. On the statistical office side, the aim is to supply the information in a form accessible to all relevant users. This can mean that the same document will be made available in different formats. Indeed, some sites give you a choice of formats for the same document. This is actually the case of the METIS site. There can, however, be reasons to limit the number of versions of a document.

30. There can be a financial constraint: for instance, the software to produce PDF files is subject to a license fee based on the number of people allowed to use it; or the cost of producing multiple formats is prohibitive. These can be reasons for a statistical office not to make documents available in certain formats.

31. There can also be a technical reason for not publishing documents in a variety of formats. Availability of storage space, aesthetics of a web site, web site security concerns, and complexity of maintaining multiple formats can contribute to such a decision.

32. Legal considerations can also be of influence in the choice of a format. A statistical office owns the content, and wants to make sure the source is mentioned if part of the content is integrated into another document for further publication. This can be the case of a journalist making use of statistical charts in an article about the economic situation. This can lead to a preference for a format that firmly ties the chart and the reference together, such as PDF.

33. These considerations, along with other preferences, can lead to the issuing of a policy regulating the choice of formats for documents on Internet. It is not clear to us to what extent the leading statistical offices have such policies. It would be pointless to formulate recommendations without knowing more on this subject. Therefore, the rest of this paper will be devoted to asking a few questions. The METIS participants are kindly requested to find out what the answers are in their respective offices, and to send them back to the authors.

34. The preferred communication channel for the answers is e-mail. The authors have the following addresses: jknt@cbs.nl and Gillman_D@bls.gov.

Question 1: Policy Does your office have a policy on document formats? We would like to know whether the choice of formats is left to the people in charge of producing the documents or uploading them, or if this choice is subject to predefined rules. If such a policy does not exist (yet), is there a plan to formulate one?

Question 2: Criteria What are the criteria to be taken into account when choosing a document format? We are interested in the role played by the content, the user, the office, or any other relevant aspect.

Question 3: Co-ordination Does your office co-ordinate its document format decisions with any other groups or organisations?