UNITED NATIONS STATISTICAL COMMISSION
and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS
METHODOLOGICAL MATERIAL

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

UNITED NATIONS
Geneva, 2001

CONTENTS

PREFACE

The methodological material "Recommendations on formats relevant to the downloading of statistical data from the Internet" was prepared at the request of countries participating in the activities of the Work Sessions on Statistical Metadata organised by UNECE Statistical Division within the programme of work of the Conference of European Statisticians. This material has a link to the "Guidelines for statistical metadata on the Internet" (published in 2000 in the "Conference of European Statisticians Standards and Studies' series under no. 52). Its aim is to give more detailed recommendations for statistical data and metadata formats to be used on Internet following the principles outlined in the "Guidelines".

The document was prepared by UNECE consultant Jean-Pierre Kent from Statistics Netherlands and Daniel Gillman from the U.S. Bureau of Labor Statistics in cooperation with experts of the UNECE member countries and international organisations. It was reviewed at the Work Session on Statistical Metadata in November 2000. National Statistical Offices of the UNECE member countries and Eurostat, the European Free Trade Association (EFTA), Food and Agriculture Organisation (FAO), International Labour Office (ILO), International Monetary Fund (IMF), Organization for Economic Cooperation and Development (OECD), United Nations Educational, Scientific and Cultural Organisation (UNESCO), United Nations Industrial Organization (UNIDO) and United Nations Statistics Division (UNSD) participated in this meeting. The material reflects the outcome of the discussion on the document.

At its 2001 plenary session, the Conference of European Statisticians agreed to publish this methodological material and to distribute it to interested statistical offices and other users.

SUMMARY

The aim of the publication is to assist national and international statistical offices in choosing the formats for making statistical data and metadata available on Internet. The recommendations are of a general nature, as no generally usable formats have yet been developed that would be specifically targeted for statistical data. It can be expected that in future such formats will appear. A promising initiative in this respect is the international Task Force lead by IMF to develop a common standard for Statistical Data and Metadata Exchange (SDMX) based on the XML language.

The use of networks, especially Internet, has substantially changed the way in which users access statistical data. It has considerably increased user expectations concerning data availability, comparability, timeliness and ease of use. The provided formats and their characteristics (readability, provision of metadata, possibility to further use the data, time needed for downloading, etc.) are an essential part of the availability. Customer research in statistics shows that, for users, the quality of the delivery can be as significant as the quality of the data. Very important are the factors at the interface between the user and statistical office (reliability of supply, service quality and delivering on time, trust, personal relationships and responsiveness). All these aspects should be kept in mind when deciding about the format of data dissemination.

There are a number of formats available for statistical agencies to provide data and documentation to users. For agencies that wish to make these products available on the Internet, the number of useful choices is limited by

- the resources available to the agency for providing data and documents in multiple formats;
- the requirements of the user, i.e., which software the user has available and what actions the user wishes to take with the product;
- the inherent limitations of each format, especially as they relate to appropriate metadata and read/cut/write access.

For the reasons provided in this paper, the authors recommend different formats depending on the agency/user situation. XML and PDF can be considered formats with the greatest potential for widespread use.

An ideal situation would be to make data available in many different formats to meet the needs of all users. In real life, however, the number of different formats is limited because of financial constraints or technical reasons, like availability of storage space, aesthetics of a web site, web site security concerns, and complexity of maintaining multiple formats. Legal considerations can also influence 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 lead to a preference for a format that excludes editing, such as PDF. If it is desirable for the users to be able to further process, analyse or integrate data, then a text processing format (e.g. MS-Word) would be appropriate for a natural language document, or a spreadsheet format for tabular data.

It can be recommended that the statistical office would decide on a policy regulating the choice of formats for documents as part of its general Internet dissemination strategy.

1. INTRODUCTION

This report presents the problems involved with the choice of a format for documents that make statistical data available on the Internet, and formulates recommendations based on the properties of the available formats.

The paper is structured as follows. It starts with a historical perspective on publication media, from printed material to the Internet, with particular attention to the World Wide Web. Then it presents a shift of focus from file to document, in the context of fast evolving technology. After this, a typology of formats is introduced, to allow a more abstract discussion of formats. We finally draw upon these considerations to formulate some recommendations.

2. PUBLICATION MEDIA

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.

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 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.

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

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 that 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.

3. INTERNET

The ultimate WAN is the Internet, which interconnects computers, WANs, and LANs throughout the world. It combines the functions of a network, allowing the sharing of data and programs, with those of a dissemination medium. 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, press releases, etc.) from their sites.

The three most widely used methods for transferring data over the Internet are HyperText Transfer Protocol (HTTP), e-mail, and File Transfer Protocol (FTP).

- HTTP is mainly used as a "pull" method, and can be seen as the electronic continuation of the traditional paper-based publication media: the author or publisher makes the information available, and customers take the initiative of accessing it and acquiring it. HTTP is the preferred protocol for the interactive part of the Internet, the World Wide Web (WWW). HTTP is supported by web browsers, programs that interpret HTML (HyperText Markup Language) or XML (eXtendable Markup Language) and allow users to follow links and select information.

- E-mail is the main "push" method: the author has a specific reader in mind, and the recipient's address is part of the protocol. Mail messages are written and read in dedicated programs that take care of both encoding and dispatching. A message can be supplemented by any number of files in any format (attachments). This makes it the method of choice for sending documents or references to them (update notification, the document itself being pulled) when the users are known in advance.

- FTP combines the push and the pull approaches. It is primarily meant as a method for transferring files to or from a remote computer, and is particularly suited for batch work. FTP can also be used interactively, and is the preferred method for large file transfers.

A few other protocols exist, but they are not widely used yet and will not be discussed in the present paper. We must, however, pay sustained attention to the technological evolution, in order to anticipate widespread acceptance of other protocols and the emergence of new ones.

These protocols do not all support the same types of activities. Here are the differences, from the point of view of a national statistical office:

- HTTP allows individual users to access the site either anonymously or through an authentication protocol, and interactively find their way to the data. They need not know in advance what they are going to look for: if the HTTP site is well structured, navigation information is clearly and completely comprised in the links that are part of the displayed pages. Therefore, HTTP is the digital successor of traditional paper-based dissemination methods, and an HTTP site on the World Wide Web is the method of choice for making information available to the world in general.

- E-mail allows the office to send specific information to one or more selected users. This can be useful if the users are personally known and have been allowed to express their data needs. E-mail is, therefore, very well suited as the vehicle of a subscription system.

- The use of FTP entails that the user has access to the FTP site at file level, which can be the source of security problems. User rights can be restricted through the use of user names and passwords. Anonymous "guests" can be granted read-only access. One is usually reticent to grant FTP write access to a large public. A further limitation of FTP is its less widespread availability in relation to HTTP.

FTP is mainly suited for communication between statistical offices. The typical situation in which FTP is the preferred protocol is the transfer of national data from a national office to an international office. The international office can open its FTP site to national offices, allowing them to upload data as soon as they become available. Alternatively, the national office will grant access to the international office, which will periodically search the site in order to download whatever new data have been added.

4. FROM FILES TO DOCUMENTS

The different available protocols illustrate an important paradigm shift that has taken place since the advent of the Web. In the past, the accent was on form: files were accessed through file descriptions. The formal structure and format of a file was all important. The content was considered implicitly understood by humans and irrelevant for software. Nowadays we tend to refer to documents, with more stress on the structure of the contents. Unlike e-mail and FTP, which are not concerned with the content of files, HTTP reflects this shift. HTTP is not file-oriented but

document-oriented. A document can consist of one or more files. HTTP can use the URLs embedded in hypertext links of the HTML or XML formats in order to tie different files together. The location of the files forming one logical document is irrelevant in this concept. They can be distributed over different servers all around the world.

The Web has clearly been influenced by this shift. Originally, the Web contained static files (pages) only. The emergence of active mechanisms such as the Common Gateway Interface (CGI), Active Server Pages (ASP), Java servlets and Java Server Pages (JSP), etc. made it possible to generate HTML pages dynamically, based on many scripting languages. It was not long before dynamic Web pages were created from the output of databases. Now, with the development of Java, JavaScript and ActiveX, on the client side, even the Web browser interfaces can be made as flexible as a true application.

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 or generated 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, unless it is generated for them. It is, however, not yet standard behaviour for search engines to access the contents of a database, or to activate browse-time mechanisms. Therefore, a site providing dynamic content is compelled to provide its own specific search engine, which defeats the purpose of generic search engines, i.e., to find content without knowing where to look for it.

The most recent development is eXtensible Markup Language (XML) for use on the Web. XML was developed to replace HTML, although browsers do not yet interpret it as well as they might. 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 allows software to access the document in a meaningful way, by using content information. 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. Therefore, XML can be seen as the most recent step in the shift of focus from form to content. The persistent use of older, non XML-enabled browsers is not a limitation to the use of XML. An intermediate server (or maybe the web server itself) can apply a style and deliver simple HTML files for the browser. In this case, content and presentation remain separate – even if the presentation is not processed by the client.

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 this is not always the case. Different mime-types are assigned for files of the same extension depending on the intended use (download or display), expiration date, etc. Each type induces different behaviour from the browser. Display behaviour by the browser can be implemented through plug-ins. MS-Word (*.doc) files are mapped to the mime-type application/MS-Word by the server. On the client side, in the browser, the mime-type is recognised and the plug-in launches MS-Word or the free word-viewer to read the file.

This approach combines downloading and browsing in one operation. It greatly simplifies access operations for the user with the right kind of plug-ins. Users without the correct plug-ins must resort to downloading files to their machines. Then they must launch the correct applications themselves – if available – in order to read the files.

5. THE FUTURE

At present, technology is developing at a rapid pace, and the near future will offer new possibilities that might eventually cause us to revise some or all of the present recommendations. This dynamism is yet

another reason for the use of standards and for keeping content and presentation separate as much as possible. We expect to see the development of new access methods along the following three paradigms:

- *The human-centred approach*. People carry around documents in order to read them in their spare time or to use their content as a support for their work. The only difference with past habits will be that the medium has become electronic. The present technology, called 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. In the future, we will have lighter, thinner and cheaper devices that look more like sheets of paper.

- *The machine-centred approach*. Speed and capacity of hardware is expanding, and software draws upon this growing power to accomplish more numerous and complex tasks in less time. This will spawn a trend where the machines will take charge of searching, selecting and presenting information. 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.

- *The communication-centred approach*. We will want to be able to access any information at any time from anywhere. This requirement will be supported by devices that access the net through a wireless communication medium. The current state of this facility is called WAP (Wireless Application Protocol), and allows Web access through mobile telephones.

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, however, evident that the PC as we know it at present will not be the only means of access to documents on Internet. It is important to note that the three approaches sketched above are not mutually exclusive. It is easy to imagine a device that combines wireless communication with offline document use. We could use such a device to send our smart agents to fetch some information, bring it back via the wireless telephone line and store it for offline use. Therefore, these technologies are highly complementary, and should be expected in the near future to merge into a general system of wireless co-operative networking, in which semantic-level communication hides the implementation-level databases, engines, formats and protocols.

In this context, the sharp distinction between collection and publication tools will tend to fade. This trend is being enabled by the development of XML-compliant data description languages that allow the implementation of an infrastructure for transferring data from registers to statistical offices. Under this paradigm, the initiative of communication can be taken both by the office and the data provider. The currently relevant technologies are IQML (Intelligent Questionnaire Markup Language, a project currently under way under the Fifth Framework – see [4]) and XBRL (eXtensible Business Reporting Language – see [1]).

The ultimate Internet access could well turn out to be a system in which users need do no more than specify their subjects of interest, and leave both the selection and the presentation to the machine. When this comes into effect, there could be little correlation left between the format of the data on the net and their format on the offline device.

6. FORMAT TYPOLOGY

In order to discuss the advantages and drawbacks of the different formats available, it is useful to establish a classification of formats. A number of criteria are available for such a classification. The kind of content can provide a useful approach, because a format defines a technical representation of the content. The

intended usage also gives helpful indications. For the purpose of this document, we will use two approaches. One will be to look at the relationship between data and metadata. The other will take into account the human vs. machine readability of the document.

Metadata - The meaning and usefulness of data is primarily determined by the metadata through which the data can be accessed. It is therefore logical to take the relationship between data and metadata as a basis for classifying documents and their formats. This approach is particularly appropriate in this paper, considering that it is intended as a complement to [3]. In the present paper we will consider the following three cases:

- *Separate metadata*: the file contains data only. The metadata are to be found elsewhere. Typical examples are fixed-format and comma-separated ASCII files, which comply with a structure defined elsewhere.

- *Loose metadata*: both data and metadata are present in the same file or in different files tied together through HTML hyperlinks. However, the relationship between data and metadata is not formally defined. This is the case of most formats designed for interactive processing, such as word processing formats. Metadata are laid out in a way to guide humans to establishing a correct interpretation. The rules to interpret the metadata are not defined by the format, but rather by human culture.

- *Tight metadata*: both data and metadata are present in the file, in a way that makes their relation unambiguous. Databases are a good example of this type of format.

These categories are not always exclusive. For example, XML embeds all data in a hierarchical metadata structure, and is thus an example of tight metadata. However, the metadata comprised in an XML document can more or less heavily draw upon external definitions given in an XML-Schema, a DTD, or an RDF. Therefore, XML also illustrates the category of separate metadata. GESMES – GEneric Statistical MESsage, Eurostat's format for transferring data from national statistical offices – also illustrates this hybrid approach: the data are tightly embedded in metadata, but most of the metadata are made of references to definitions specified elsewhere.

Readability - We access documents for two main types of goals: either to become acquainted with the information they contain, or to submit this information to further processing such as analysis or integration with other data. In the first case, we need to be able to read and understand the document. In the second case, we want our software to be able to access the data correctly. From this point of view, we have three types of document formats:

- *Human-readable formats*: this category primarily comprises formats designed mainly as successors to paper-based communication. Word processing formats (MS Word, WordPerfect, etc.), read-only formats (PDF) and graphical formats (GIF, JPEG, etc.) are the main examples of this category. The concept of human readability must be understood in a broad way. It implies not only the possibility of viewing, but also the possibility of understanding the meaning of what we see without any assistance from software beyond the normal display functionality. In this sense, a comma-separated ASCII file is not human-readable: although we can read the separate values, extra processing is needed for us to understand the content.

- *Machine-readable formats*: those formats are not readable for humans. Viewing the content of such documents is supported by selection and presentation functionality. Databases offer a typical example of this category.

- *Universally readable formats*: here we find formats designed for interactive

processing of the data. Most new formats, e.g. XML, are of this type: human-readable and machine-processable.

Interdependence - These two typological views on document formats are interdependent and do not lead to a 3 by 3 matrix of document types. Separate metadata are supposed to be loaded in a machine before they can be used to interpret data. Separate metadata and human readability do not go hand in hand.

On the contrary, loose metadata are typical for human readability. Most natural language based documents have loose metadata. Loose metadata do not provide the machine with the formal unambiguous information that it needs in order to make proper use of metadata.

Finally, tight metadata can easily lead to complex structures that are hard or impossible to process with the eye. Database files are an example of tight metadata. In this case, extra processing is required for humans to access and understand the data.

We need to pay special attention to universally readable formats: being accessible to both humans and software, they illustrate the difference between the metadata needs of people and machines. Spreadsheets, for example, have two layers of logic: in the presentation layer, there are loose metadata, which are used only by people, in order to understand the meaning of the numbers. In the computational layer, however, values have statuses such as constant, formula, cell reference, etc. These tight metadata determine the structure of the model and allow the machine to display the correct values. Therefore, spreadsheets use both loose metadata for human interpretability of the meaning and tight metadata for machine computability of the values.

A similar remark can be made of tabular data with row and column titles, presented in ASCII form. This is an example of human-readable data with loosely coupled metadata. However, such a document is easy to describe in a formal way, in terms of column positions and widths. A metadata document with such information is probably available, because the data could not have been formatted properly otherwise and it could be used to enable software to access the table in a meaningful way. This is again an example of two disjointed sets of metadata: separate metadata for the machine (column positions and widths), and loose metadata for the human user (titles, labels and footnotes).

7. RECOMMENDATIONS BY DOCUMENT CATEGORIES

These considerations allow us to divide document formats into the following six categories:

A. *The bare data file:* This is a file containing only values. The metadata are elsewhere, either embedded in program code, or described separately in a human-readable format. The values can either appear in predefined positions (fixed-format), or they can be separated by a predefined code, usually a comma (comma-separated). This is the oldest form of digital information. It was thought of before machines had much memory capacity, when storage was on punched cards and working space was measured in kilobytes. This format is very dangerous, because it offers no operational guarantee that data are used with the appropriate metadata. The risk of incorrect use has become particularly important with the size and ease of access of the Internet. With today's memory capacity in mind, it makes sense to unconditionally advise against loading such files on an Internet site.

B. *The graphical file:* This is a family of formats capable of storing pictures (GIF, JPEG, BMP, etc.). They are perfectly suited for purely graphic information, such as photographs or drawings, meant for human eyes. They are not suited for the

presentation of data, because they are not able to represent data elements internally. Therefore, graphical formats are not suited for managing metadata either. So we recommend never using graphics as the basic format of a document. Graphical information should always be embedded in a document containing at least metadata, and preferably data as well, in a non-graphical form.

C. *The natural-language file:* This is the category of human-readable documents with loose metadata. Their main purpose is to present textual information on paper (or on screen). In its simplest form, the ASCII text file, the machine sees the content as an array of characters structured only in terms of lines or of paragraphs. More sophisticated formats allow software to define the layout and to control the structure in terms of pages, paragraphs and chapters. This is the case of text processor formats, such as MS-Word and WordPerfect. It is also the case of read-only formats such as PDF and Envoy. HTML, the HyperText Markup Language of the Internet, also falls into this category.

Although all these formats adequately handle natural-language information, they are not equally suited for documents to be made available for downloading. Here is a synopsis of their qualities and drawbacks:

- ASCII (American Standard Code for Information Interchange) is the most universal, because it is so simple. It requires no specific software. It can be sent as raw data to the screen or to the printer. However, this format is severely limited in terms of functionality. It has no layout control. It has only very elementary means of structuring text, such as separating paragraphs with an empty line, showing titles in capital letters, or starting a new chapter with a page separator. The only means to arrange text in tabular form is through the use of spaces or tabs. This causes problems when the reader uses another font, font size, margins, or page size than that used by the author. Further, it cannot embed graphical objects.

  ASCII has a very limited set of 94 characters. Extended codes comprising up to 127 extra characters, and called 'code pages', have been thought of for support of a variety of languages, but use of these extensions requires separate metadata for the code page identification. All this makes ASCII quite unsuited for anything but very simple unstructured text in English. This format inflicts an old-fashioned, unprofessional aspect on data, and thus should be avoided by professional publishers.

- HTML can be interpreted by browsers, and any one accessing the Internet can be expected to be able to use one. Therefore, HTML can be considered a reasonably universal format. HTML, however, cannot embed objects. It can only refer to them. The different parts of an HTML document are usually stored in separate files. It is the task of the browser to follow the links and present the different elements together as a whole. This is what browsers normally do when browsing, i.e. accessing a document on line. If an HTML document needs to be downloaded for off-line access, this feature becomes problematic. It is difficult to download all the parts of a document in such a way as to guarantee that off-line reading will produce the same result as on-line browsing. It also lays a potential burden on the users: if they move or delete the document, it is their responsibility to treat all the files together. Therefore, we recommend limiting the use of HTML to documents that will be browsed on line. If, however, HTML is put to use for downloadable documents, the author should be careful to limit the structure

to something HTML is able to store in a single file.

- Word processor formats used not to be accessible without the software they belonged to. MS-Word files needed to be accessed through MS-Word, and WordPerfect files needed to be accessed through WordPerfect. In their present versions, word processors are able to read files produced by a wide range of other word processing programs. The resulting presentation of the document, however, is not always satisfactory. The same can be said of the Rich Text Format (RTF), designed for platform-independent document interchange. Although the leading word processors support this format, one is not always sure of seeing the intended picture when viewing an RTF file in a different program from the one it was made with. The main drawback of word processing formats, however, is that it is very easy for users to edit the documents and pass them on. The user is not always aware of modifications taking place: the program tries to adapt the document to the standard page format used locally, and when asked to save the changes, the user might inadvertently click "yes" by habit. This can cause tables to become misformatted, or pictures to appear on the wrong page. Subsequent users will be unaware of this and consider the document to be the sole result of the statistical office's work. Therefore, word processing formats should be avoided, unless end-user editing is part of the intended purpose of the document.

- Read-only formats are both flexible and accessible. They can accommodate all the presentation functionality of word processors, and the programs to access them are available to download free of charge. The most popular read-only format is the Portable Document Format (PDF). The access program for PDF documents is provided by Adobe and is called Acrobat Reader. To make PDF documents universally accessible, it suffices to make Acrobat Reader accessible for downloading. A read-only format, and particularly PDF, is the best choice if the integrity of the document is an issue – which is usually the case of published data.

  - A read-only format protects a document against inadvertent or naive changes. Malicious tampering, however, can only be excluded through certification based on encryption algorithms. PDF does not support such certification. Neither does any other widely used format on the Internet.

D. *The data managing formats:* This category comprises many formats designed for software access and processing of data. These formats have tight metadata, and require presentation functionality to produce views or reports. A database management system offers the most typical example of this type. Data processing software in general either uses standard database formats, or defines its own proprietary format. In the context of statistics, SPSS offers an example of such a format. Transfer protocols, such as GESMES, also belong to this category.

   Access to such formats depends on the presence of the software they belong to. It is not reasonable to count on the presence of such software on the user's machine. Therefore, these formats should only be used on the basis of mutual agreements between the publishing office and the intended users.

   Some offices make use of proprietary formats supported by specific software. Use of such formats for downloadable documents is acceptable, provided the software is made available for downloading on the same site.

E. *Mixed formats:* This category comprises formats that support universally readable

documents, with loose metadata for human readability, and formal metadata for automatic processing, as explained above. The main exponent of this category is the spreadsheet format. These formats, supporting both the presentation of the results and further processing by end users, seem to be widely accepted by users for presentation purposes and as a sort of "intermediate" format towards further processing. The remarks made under point C also apply here:

- If end-user editing is not part of the intended purpose of the document, spreadsheet formats, like word processing formats, should be avoided.

- Tabular ASCII files, like natural language ASCII files, lack presentation functionality and look clumsy. They should always be avoided.

F. *XML:* This format deserves a special mention. It is the newest of all, and has features that make it very attractive for communication of data. XML is more than just a format. It provides the tools for defining new, XML-conformant formats for documents of a specific category. The structure of a type of documents can be specified in an XML-Schema, a Document Type Definition (DTD), or a Resource Description Framework (RDF). The author, or the end user, can specify the layout in a separate document, which conforms to the rules of the XML Style sheet Language (XSL). Browsers use this definition, if available, in order to present the content in a way that is compatible with the intentions of the author. However, browsers are able to give a meaningful presentation of an XML document even if it has no access to the structure definition or to an XSL specification.

XML's main advantage is that the content of a document is accessible not only to browsers, but to any XML-enabled software that is able to use the proper definition. This puts XML in a position to take over the functions of most other formats. It can support tight metadata like database formats. It can mix tight metadata and loose metadata like spreadsheets. It can manage natural language with loose metadata like word processors and HTML. It can also support tight metadata referring to separate metadata as some transfer formats do. GESMES illustrated this last point by switching to XML for its syntax. However, XML supports hypertext linking, which enables it to build complex documents from multiple files, like HTML. Therefore, the warning issued about HTML also applies to XML. When publishing downloadable data in XML, one should avoid using links to structure a document. One should aim to produce a single file containing all data. This being said, XML is the format of choice whenever data are highly structured and use of specific software is not critical.

8. OTHER CRITERIA

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.

We are interested both in the user's wishes and needs and in the technology at his or her disposal. We can roughly distinguish two types of user needs: reading and processing. In the first case, the users want to become acquainted with the content. They want to display or print the document. Maybe they wish to cut and paste some of it for use in their own documents. This kind of use is best served by non-editable natural language formats. The example of choice is PDF – although not all versions of PDF support cutting and pasting. If, for any reason, some editing functionality is desirable, then a text processing format (e.g. MS-Word) would be appropriate for a natural language document, or a spreadsheet format for tabular data. In both cases, the transfer would take place under the HTTP protocol, under control of the user. Alternatively, e-mail can be

used from the office, if the documents are tailored to the needs of specific users under a subscription contract.

Processing could mean a variety of activities, the most important of which are analysis and integration. This type of use applies to machine-readable documents, because the processing will be done through software. Statistical institutes very often provide data with proprietary software. Those formats can be useful for browsing, retrieving and exporting data (especially in the case of unexpected changes in the underlying data format) but they need to be assisted by good "software-related metadata", helping the user to understand how the software works and/or how to obtain support.

On the statistical office side, the aim is to supply the information in a form accessible to all relevant users. As different users may need data for a series of different reasons, 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.

The choice of formats will have to depend on bilateral agreements between the producer and the user. Users would be expected to be mainly research agencies (universities) or international statistical offices. For this purpose, single-file XML should always be the format of choice if a schema or a DTD is available for the type of document at hand. Otherwise, one could resort to a data transfer protocol (e.g. GESMES), or a database format (e.g. MS-Access).

There can, however, be reasons to limit the number of versions of a document.

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

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.

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 in 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 and excludes editing, such as PDF.

These considerations, along with other preferences, can lead to the issuing of a policy regulating the choice of formats for documents on the Internet. Therefore, the present recommendations can apply at two different levels. They can be used on a case by case basis by the people in charge of producing the documents, or by those who upload the documents and maintain the web site. Alternately, they can be used by policy makers who translate them into rules that are mandatory for the production and uploading of documents on the Internet.

A word about content - The choice of a format should not be considered as a goal, but as a means of supporting content. Making the relevant content available, easy to find and understandable should always be the central issue. For regular customers, the content should also be consistent with whatever has been provided in the past. These considerations lead us to lay particular stress on the following points:

- Recurring data on the same subject should be integrated in time series.

- Revised data should be made available under reference to the data they replace.

- Data should always be accompanied by extensive metadata, as recommended in [3].

Conclusion - There are many formats available for statistical agencies to provide data and documentation to users. For agencies that wish to make these products available on the Internet, the number of useful choices is limited. The limitations are characterised by the following:

• The resources available to the agency for providing data and documents in multiple formats.

• The requirements of the user, i.e., which software the user has available and what actions the user wishes to take with the product.

• The inherent limitations of each format, especially as they relate to appropriate metadata and read/cut/write access.

For the reasons provided in this paper, the authors recommend different formats depending on the agency/user situation. Formats with the most potential for widespread use are XML and PDF.

List of abbreviations

ASCII: American Standard Code for Information Interchange
ASP: Active Server Pages
CD-ROM: Compact Disc Read Only Memory
CGI: Common Gateway Interface
DTD: Document Type Definition
FTP: File Transfer Protocol
GESMES: GEneric Statistical MESsage
HTML: HyperText Markup Language
HTTP: HyperText Transfer Protocol
IQML: Intelligent Questionnaire Markup Language
LAN: Local Area Network
PC: Personal Computer
PDF: Portable Document Format
RDF: Resource Description Framework
RTF: Rich Text Format
URL: Universal Resource Locator
WAN: Wide Area Network
WAP: Wireless Application Protocol
WWW: World Wide Web
XML: eXtendable Markup Language
XSL: XML Style Language

References

[1] Aplin, E., Slater, R. – *Statistical EDR : The Australian Experience*. Paper submitted at the Electronic Data Reporting Workshop, Canada, 25–27 September 2000.

[2] *Best Practices in Designing Websites for Dissemination of Statistics.* United Nations, 2001 (Conference of European Statisticians Statistical Standards and Studies No. 54).

[3] *Guidelines for Statistical Metadata on the Internet*. Geneva, United Nations, 2000 (Conference of European Statisticians Statistical Standards and Studies No 52).

[4] Kunzler, U. – *Electronic Collection of Raw Data (eCoRD) – a European Perspective.* Paper submitted at the Electronic Data Reporting Workshop, Canada, 25-27 September 2000.