Computer and Communication Systems (LCI[1]) 2004[2]

J.Magalhães Cruz, FEUP

Electronic documents' formats

Electronic document

Definition

An electronic document is a document suitable for use by computerized means.

Its usage may consist of creation, alteration, storage and destruction.

Types

An electronic document, like a classic document (in a broad sense), can consist of text information, image information, sound information, animation information and other types of meta-information.

An electronic document may have a single type of information or be made up of a number of distinct types of information.

Typically, the manipulation of electronic documents can only be carried out by computer means and using appropriate software tools, often specialized in a given type of information.

Binary representation

An electronic document is represented internally (by the computer system where it resides, is manipulated or passes through) by a sequence of binary numbers (at least in current computers, which use the binary system!).

However, this binary representation is not unique, that is, the same document can have more than one binary representation, all perfectly appropriate.

Example: a document with only the decimal number "`1`" can be represented by "`00110001`"[3] or by "`11110001`"[4].

It will be up to the software tools used to manipulate the document to correctly interpret the representation used in each case. However, given the multiple possibilities of representation, an automatic interpretation is not

1 Licenciatura em Ciência da Informação (Bachelor of Arts in Information Science), `https://sigarra.up.pt/feup/pt/cur_geral.cur_view?pv_curso_id=454`

2 translated to English and slightly revised in 2021 and in 2024.

3 in ASCII (American Standard Code for Information Interchange)

4 in EBCDIC (Extended Binary Coded Decimal Interchange Code)

always possible so, in such cases, the human user will have to, in one way or another, assist in the interpretation of the information in the document.

Support medium

The electronic document or, more correctly, its binary representation must be supported by an appropriate physical medium.

Known physical media are magnetic disks (hard or flexible), optical disks, magnetic tapes, integrated circuits ("normal" memory, "flash" memory), etc.

Some physical media can hold documents for a long time without the aid of electronic media (e.g. hard disk); others can only hold documents as long as they are connected to an electronic media system (e.g. "normal" memory); this is because they need an active power source.

Since the media can hold thousands or millions of documents, the different documents must be grouped together or arranged in some way to prevent the information from getting mixed up and even corrupted (if an area occupied by one document is used by another document).

Files

A file is a logical grouping of information relating to an electronic document that is stored on a support medium.

A file can be stored on the medium in such a way that all the information (in binary representation!) is physically close, or scattered across the medium.

In any case, the file management software of the computer system being used will know exactly where on the support medium all the parts constituting the binary representation of the saved electronic documents are.

Exchange of documents

Binary and text representation

Many electronic documents contain information that unequivocally corresponds to text (e.g. a Philosophy paper); others may contain information of a different type, difficult to be translated into words (e.g. an abstract drawing). A third situation would correspond to a document containing information which is not necessarily textual, but which could be represented more or less accurately by a text (e.g. an image of fruit in a bowl).

Thus, in certain situations, a document may be stored in a file containing the binary representation of a combination of a set of basic characters, such as the alphabet of a given language. That document is said to be in "**text format**".

In other situations, it is difficult or even impossible to store the information of a document using a grouping of character representations; in this case, a representation of the document is made in "**binary format**". Now, it is very difficult to understand the representation made if we are not aware of it, because there are many possibilities to make such a representation.

As an example to illustrate the difference between the types of formats, try to see with a simple text editor the content of a `TXT` file and the content of a `JPEG` file. (You can also use a tool that allows you to "see" the binary representation of the information, e.g. `okteta`, under Linux.)

Formats

With either of these representations, text and binary, it is possible to make variations in the way information is represented. These variations will correspond to different "file formats". For example, a document with an image can be saved in `PNG` or `JPEG` format. A document with a romance story can be represented by a file in `TXT` or `XML` format.[1]

Each tool used for handling a given type of document (with a given type of information) can typically save the document in one of several file formats. However, it is customary to "prefer" to save it in a particular format, specific to the tool: the so-called "native" format. Typically, it is a binary format, even in cases where the information is essentially text (that is the case of a word processor's document).

On the other hand, a file in text format will necessarily have to specify the way its characters are represented in binary. In an example given above, the decimal number "`1`" was represented by "`00110001`" and by "`11110001`". These different representations are much easier to take into account (to interpret) than the different possible representations of binary formats.

Standardization

If you want to use the information in an electronic document in different ways (for example, you may want to insert a sentence - text - in a picture) or if you want to enable different people, possibly working with different computer systems, to access a document, you have to ensure that the information is correctly understood and manipulated.

This implies that documents have a normalized (standardized) representation, and that companies developing software tools understand, accept and follow the agreed representation standards.

And agreed by whom? Typically there are two situations: one is a *de facto* standard, where a given representation, perhaps created by a single company, is so widely used that other companies follow it in order to make their applications compatible. In this case, it is important that the standard is public so that it can be known and accurately used by any company or entity.

Another situation is that corresponding to a grouping of companies or institutions that decide to create standards by common agreement (examples of such groupings, international or national, are ISO - International Organization for Standardization, ANSI - American National Standards Institute, etc.).

The standards established or to be established for the representation of information can act at various levels: at the structural representation of a given type of document (plain text, formatting and embellishment information, structure by sections and chapters, etc.), and at the binary representation in the support medium (remember the case referred to of the "`1`").

The standards could also cover issues such as representing in "text format" information that is not necessarily textual, codifying the linguistic representation of different peoples, even allowing the observation of sentences in different languages in the same document, etc.

Above all, the exchange and sharing of electronic documents will only be fully possible if the standardization process is **open**, allowing any entity to access the standards, and if it is **complete**, specifying everything that could be ambiguous.

Problems

Compliance

Compliance with standards is essential for the exchange of information and for the user to be able to freely choose the software tools he/she wants to work with.

1 The latter format stores more information than just the sentences in the novel; however, it can be read - and understood - with a simple text editor, such as the `TXT` file.

Unfortunately, commercial companies that develop IT (Information Technology) products are usually only interested in standards-compliance if their products are in the "ascendant" phase of their market. That is because they are very eager to attract customers.

As soon as a computer company becomes dominant (leader) in the market, with a large number of customers, it is interested in assuring their products are unique, different from what the (weak) competition can offer, and so keep their current customers "captive". In this case, and from a strictly commercial point of view, the only standards they are interested in following are the "company standards"!

Text representation

The "text" format, despite being one of the easiest electronic document formats to standardize, must also obey certain requirements related to the planet's linguistic diversity. The English alphabet, for example, has 26 letters, but the Portuguese had only 23 just a few years ago; but, following the Portuguese example, a multitude of characters associated with vowels and some consonants (e.g. é, ê, ç...) should be added. The complexity of the situation increases when dealing with "different" alphabets such as Cyrillic, Hebrew, Japanese (with several variants), Korean or the Chinese ideograms.

This diversity of character encoding formats can be seen in the labels '`Latin-1`', '`ISO 8859-1`', '`CP 1252`', etc., which sometimes appear in specialized computer documentation or in the text of e-mail messages or WWW pages.

One standard representation that has been gaining visibility for some years now is the Unicode representation, which aims to unite the representation of the languages of the world into a uniform representation (but with respect of diversity). However, even using Unicode, there are character encoding variations that must be specified in order to share text documents correctly (`UTF-8`, `UTF-16`, ...)

Space reduction

There is often a need to compress a file, an electronic document, in order to minimize the space it takes up in storage or to speed up its transmission over distance.

In such situations file compression can be performed using specialized software tools that follow specific compression algorithms. This applies both to binary files and to text files. In this way a new file is obtained in a new format, in general, a binary file. Everything that has been said about the need to standardize file formats also applies here.

Interestingly, some of the binary formats used in certain applications are themselves already compressed, at least partially (e.g. `PDF`), and there is usually no advantage in using a compression tool in such cases (unless the algorithm used by the tool is much more efficient than the algorithm used to build the file in its original format).

Examples to try out: investigate `ODT`, `ZIP` and other formats that you have heard of and even have used.

… (to be continued, some day) ...[1]

References

– http://www.unicode.org/ - Unicode
– http://www.w3.org/ - W3C - World Wide Web Consortium
– http://www.iso.org/ - ISO - International Organization for Standardization
– http://www.ansi.org/ - ANSI - American National Standards Institute
– - ...

1 e.g. MIME Types