erpatraining

File Formats for Preservation
Austrian National Library, Wien
10-11 May 2004

Briefing Paper

File formats are a crucial layer, indeed a hinge between the bits in storage and their meaningful interpretation. The proper access to and display of content depends entirely on the ability to decipher the respective bitstream, and therefore on precise knowledge about how the information contained within is represented. Consequently, file formats are one of the core issues of any digital preservation approach, and file format obsolescence is a major challenge for anybody wanting to preserve digital files.

ERPANET proposes to examine this subject from different viewpoints, to present and analyse best practice, and to give participants working guidelines and practical insights.

Requirements for file formats

Repeatedly, requirements for file formats suitable for preservation have been formulated by the preservation community. Some of the most important requirements will be discussed below.

Preservation file formats should allow for maximum persistence and should minimise risks of obsolescence. There is wide consent that this will best be attained through the use of standard formats. Widely used standards are very likely to be supported longer than average, and a critical mass of users is more likely to advocate conversion or preservation efforts. While both official standards and de-facto standards offer these benefits, the preservation community has long underlined the importance of open standards. These are standards whose definition is freely available and not owned by any particular company. Preservation institutions prefer open standards because their open specification facilitates reusing them or converting them.

Preservation file formats should carry with them as many features of the original format (i.e. the format in which they are created) as possible, at least the most important ones. It is evident that this requirement needs to be specified further. In fact, discussions about authenticity have shown that opinions differ about what features are crucial to a document's authenticity. It is clear, therefore, that detailed requirements for preservation formats need to be informed by respective authenticity requirements. Furthermore, the question of compression steps in here. For some types of documents such as images and audiovisual files, compression algorithms are commonly used to minimise storage space or download times. Consequently, preservationists are confronted with compressed files. Since only certain compression algorithms are lossless (i.e. preserve all of the document's features), this raises concerns about authenticity and information loss. Generally, recommendations tend to rule out lossy compression algorithms for preservation purposes. With the tremendous growth of audiovisual material to be preserved, this is likely to cause further discussion.

To facilitate future access, preservation formats should be easily useable or easily convertible in then current formats. Again, this is supported by standard formats.

Finally, since no format will last forever, a good preservation file format must also offer a promising exit strategy. It should make it as easy as possible to lift the essential content off the file. Thus it will allow future migration to a then common, but as yet unknown format. Ideally, it will also facilitate any other preservation method that might be chosen some decades in the future.

The role of preservation institutions

All preservation institutions usually choose one or a range of formats to preserve their holdings. Ideally, the documents and records they hold will already be delivered in those formats. This greatly facilitates preservation decisions and annihilates authenticity concerns. However, this scenario is far from being realistic, since most preservation institutions have little influence on the formats the producers of records, documents, and data use. Therefore they must restrict themselves to accept a somewhat broader range of formats and convert them to one or more preservation formats that meet the requirements outlined above.
Attempts to impose particular file formats on the producers of digital documents have also yielded research into a so-called "Universal Preservation Format". As of today, it seems however that this approach is not promising.

A novel approach: file format registries as a support to preservation

Faced with the sheer plethora of file formats currently or formerly in use, the majority of them proprietary and non-standard, the preservation community has repeatedly expressed the need for a file format registry. Different answers to this need have already been undertaken, yet further work is deemed necessary. On a limited (national or other) scale, the UK National Archives have recently presented their PRONOM Application. On a more global scale, the MIME Media Types registry provides some basic registry functions, but this is not deemed satisfactory to answer long-term preservation concerns. An ad-hoc working group, mainly from the digital library community, is currently undertaking a new approach to a global digital format registry.

In a similar vein, automatic format detection and classification is often seen as a promising assistance to digital preservation, possibly in connection with format registries. The seminar will explore this possibility as well, offering first-hand experience and best practice recommendations.

OAIS and file formats

A number of preservation institutions have recently adopted the Consultative Committee for Space Data Systems' "Reference model for an Open Archival Information System" (OAIS). This is a high-level reference model that serves to structure and standardise preservation processes. The OAIS does not go into any details; in particular, it does not address file format questions. However, it must be borne in mind that the OAIS introduces a new kind of information, the Archival Information Package (AIP), and while it does not specify this, it is clear that the file format of an AIP is an important detail of an OAIS-compliant preservation approach. Discussion about how to implement the AIP must therefore be informed by the file format discussion and contribute to it.

Possible candidates for preservation file formats

A range of file formats are already being used for preservation purposes. Digital preservation research has recommended different file formats. The ERPANET Case Studies reveals some best practices for preservation file formats. The following short overview lists the most important of these for three kinds of documents that the seminar will examine.
Text documents: plain ASCII; PDF; XML; TIFF
Image documents: TIFF; JPEG2000
Audiovisual documents: WAV; BWF; MPEG.

All these formats have significant benefits, but also shortcomings. The seminar will provide guidelines to assess these and choose the best file format for each preservation purpose.