File Formats in Digital Preservation

Sunita Barve,

National Centre for Radio Astrophysics, PO Box 3,
Pune University Campus, Pune 411 007, India
sunitab@ncra.tifr.res.in

Abstract. Digital libraries have been built all over the world. Libraries are engaged in creating and maintaining digital libraries. One of the main challenges in maintaining digital libraries is the digital preservation aspect. The aim of digital preservation is to ensure that digital records are filed and are made available throughout time. There are different digital preservation strategies such as migration, emulation, encapsulation etc. One of the important aspects or key part of any digital preservation activity is the format of the document in which the digital document is created and added in the digital library or repository. Each digital library consists of different documents with different file formats. Due to rapid obsolescence in hardware and software technology it is necessary for the libraries to look into the details of the file formats such as their characteristics, specifications of file formats, categories of file formats, standards used for creating file formats, usability of file formats in long term, etc. Thus file formats play an important role in any digital preservation. This article tries to discuss about general issues related to file formats, what challenges are there and how format representation information is necessary in any digital preservation aspect, which file formats are preferred formats for long term preservation as well as it will discuss about international initiatives dealing with file formats management registries such as PRONOM & Global Digital Format Registry (GDFR).

Keywords: File Formats, PRONOM, GDFR, Digital Preservation

1 Introduction

File format is one of the core issues in any digital preservation approach. Digital information is produced in a variety of standard and proprietary formats, including ASCII, common image formats, word processing, spreadsheets, database documents, formulae, charts, multimedia files and sound and video. As a result of such a heterogeneous nature of storable information, a high number of file formats are now spread, and many of them often need specific software to view or edit the file. These formats are

ARD Prasad & Devika P. Madalli (Eds.): ICSD-2007, pp. 239-248, 2007

continuously evolving and becoming more complex due to new features and functionality. Before preserving digital records it is necessary to know in which file format the digital record is created. For example, to view an older word processing file, one needs software that understands the encoding schemes of the original software and can display that properly on the screen. Thus several factors need to be considered while adding any document in digital repository.

2 What is File Format?

Digital information can be saved on any medium that is able to represent the binary digits ("bits") 0 and 1. The meaningful sequence of bits with no intervening spaces, punctuation or formatting is called as bit stream. A file is nothing more than a sequence of bits and the file format is nothing but interpreting the bit stream.

Retrieving a bit stream requires a hardware device, such as disk drive and special circuitry for reading the physical representation of the bits from the medium. After a bit stream is retrieved, it must still be interpreted. For interpreting a bit stream the implicit structure of the format needs to be known. Since a disk drive, or any other computer storage system, can only store bits, the computer must have a way to convert information in zeros and ones and vice versa. There are different kinds of formats for different kinds of information (Rothenberg, 1995). The first thing a file format specifies is whether the file is binary or ASCII, and second is how information is organized.

Brown (2006) defined file format as 'the internal structure and encoding of a digital object, which allows it to be processed, or to be rendered in human-accessible form. A digital object may be a file, or a bit stream embedded within a file'.

Different file formats specify how binary digits represent the intellectual content created by a digital object's author. An example of which is the Microsoft Word Format. Microsoft word format is a specification for the storage of textual data, along with formatting information. In order to understand the Microsoft Word Format, software is required to interpret and display the data for the user. Thus there are several file formats which are available today which are incredibly complex, making the binary code meaningless to a human observer if the required software is not available to interpret the format.

The aim of digital preservation is ensuring that records are filed and made accessible throughout time, but as a result of progress in software and hardware technology old formats soon become unreadable and unusable. Different research initiatives are focusing on this issue, trying to define

preservation-friendly standard formats, as well as strategies for records to be made available over time. There are many initiatives are taken place to read and convert old file formats.

3 Types of File Formats

There are different categories of file formats available today for different applications. The official categorization of file formats is the MIME type, provided by IANA. They define the following main categories of formats:

- application
- audio
- image
- message
- model
- multipart
- text
- video

Here are a few examples of different file formats used for different applications.

Text files	**Images**
Word Documents (DOC) Rich Text Format (RTF)	Bitmap (BMP) files Computer Graphics Metafile (CGM) Drawing Interchange Format (DXF) Encapsulated PostScript (EPS) files Joint Photographic Experts Group (JPEG) Graphics Interchange Format (GIF) Tagged Image File Format (TIFF) Portable Network Graphics (PNG)
Audio	**Video**
MPEG layer 3 (MP3) WAVE (WAV) Musical instrument digital interface (MIDI)	Movie (MOV) Windows Media Video (WMV) Audio Video Interleave (AVI) Flash (SWF) QuickTime Virtual Reality (QTVR) QuickTime Movie (MOV)
Spreadsheet	**Databases**
Excel (XLS)	Microsoft Database (MDB)
Presentation	**Markup Languages**
PowerPoint (PPT, PPS) *Portable Document Format (PDF)* files	(HTML, HTM, SGML, XHML, XML)

Compression	Other
Zone Information Protocol (ZIP)	Executable (EXE)

4 Challenges in Digital File Formats

1. Many file formats become obsolete due to several reasons such as
    - Developer of that file format goes out of business
    - Developer stops supporting that format
    - The market share of the developer declines
    - Supporting program of the software change significantly
    - Third party support is lacking etc.

    (A well known example is WordPerfect file format.)
2. Format depends on obsolete hardware or operating system.
3. Format is proprietary.
4. New versions of application software may not support earlier format versions.
5. Most application software developers produce file format documentation for the formats they design and develop. Not all of them make this documentation available and even if they do, it is not always accurate.
6. The number of file formats is incredibly high. The File Extension collection asserts they have indexed over 15,000 file name extensions (Guercio, 2004).
7. There are many public domain format sites such as My File Formats, Wotsit's Format, File Formats Encyclopedia etc. but they lack any vision or plan to sustain over long period on Internet.
8. Word processing and desktop publishing systems dominate the market for the creation of documents with complex structure and layout, and the software for such use typically models and stores document structure and layout in proprietary terms. Although the software may provide mechanisms for converting documents to common interchange formats, use of such mechanisms often results in the loss or inadequate rendering of content such as page structures and layout of headers, footers and section headings.

5 File Format Specifications

A file format specification indicates the proper sub division, encoding, sequence, arrangement, size, and internal relationships that uniquely identify the particular format and allow it to be properly interpreted and rendered.

For example, a format specification indicates the location of meaningful boundaries within the bit stream and whether a particular subunit should be interpreted as an ASCII character, a numerical value, a machine instruction, a

color selection, or something else. Without a format specification a file is just a meaningless sequence of 0s and 1s.

Whenever a file is saved, the internal representation of the text document is converted to its standard format. Inversely, when the file is read by another tool, the format is abstracted and converted into its internal representation. Hence it is necessary to know the syntactical and semantic specifications of any digital file.

A format specification provides the details necessary to construct a valid file of a particular type and to develop software applications that can decode and render such files (Digital Preservation Online Tutorial, 2004). The actual specifications of any file format may vary considerably in length, from 100 pages to over 1000, depending on the complexity of the format. In digital preservation activities it is necessary to know internal structure of any file. For example, TIFF file format grew from 37 tags in version 4 to 74 tags in version 6.0. New proprietary tags for TIFF version 6.0 are registered with Adobe, which does not make their specification available in public.

File format specifications are very important in digital preservation activities. If internal structure of a file is known it is easy to maintain the file in current file format.

6 File Format Registries

A format registry is a repository for format specification information or, in other words, descriptive, administrative, and technical metadata about digital formats, including the definition of the syntactic and semantic characteristics of the registered formats. This metadata defines the significant properties of digital formats with regard to the long-term preservation of digital objects.

Brown (2004) defined file format registries as authoritative and publicly available source of technical information, supporting identification, accession, preservation, and access of files. File format registries are expected to be persistent, trustworthy, and publicly discoverable.

The most well-known example of a format registry is the Internet Assigned Names Authority (IANA) MIME type registry. However, MIME registry does not prescribe any specific set of format attributes that must be disclosed. For example, the MIME type application/msword by itself doesn't help to determine whether a file is compatible with Microsoft Word 6.0 or 2000. Also MIME typing does not provide sufficient granularity to disambiguate important format distinctions, whether based on versioning or profiling. For example, the entire PDF family – PDF 1.0 through 1.4, PDF/X-1 through 3 (ISO 15930), and the proposed PDF/A standard – are all typed with a single identifier: application/pdf. (www.iana.org/assignments/media-types/).

A number of projects are investigating or developing systems which provide repositories of file format and representation information for use in digital preservation. The popular format registries are PRONOM, and the future GDFR which provide detailed information about internal specifications of file formats and tools required to render the format.

6.1 PRONOM Services

PRONOM was launched during February 2004 by the digital preservation department of the UK National Archives (http://www.nationalarchives.gov.uk/pronom/). It is a file format registry. It is an information system about data file formats and their supporting software products needed to open them. It is a web enabled database of information on file formats and their technical dependencies, including hardware, software and operating systems. The task of preserving digital objects requires a reliable, sustained repository of file format information.

PRONOM is being made available as an information resource for anyone who needs authoritative information about software products, their support lifecycles and technical requirements, and about the file formats which they support.

The PRONOM system developed by PRO captures information about tools used for generating, manipulating, and rendering objects on a per-format basis (Abrams, 2003).

It contains 550 file format descriptions as of February 1st, 2004, but does not allow direct access to any specifications they may have stored (Clausen, 2004). The most important propriety of the file PRONOM format registry is the persistence to act as a resource which a digital repository can actually point to (Guercio, 2004).

6.2 Global Digital Format Registry

Detailed knowledge of the internal properties of digital formats is necessary to interpret properly the full information content of digital objects. All digital repositories need to be able to identify, validate, characterize and process those objects on a format-specific basis. Digital format representation information is of potential use to all institutions and individuals engaged in digital preservation.

Harvard University and the Massachusetts Institute of Technology are leading an initiative to establish a centralized registry of file format information. The initiative is currently at an early stage but has already seen international interest and contribution from a range of institutions and organisations facing digital preservation problems. (http://hul.harvard.edu/gdfr/)

Global Digital Format Repository (GDFR) will function as a sustainable public service for the collection, maintenance, and dissemination of authoritative information about the significant syntactic and semantic properties of digital formats and of systems that support and manipulate those formats.

GDFR was conceived of as a single centralized repository of format representation information. The scope of GDFR is to "maintain persistent, unambiguous bindings between identifiers for digital formats and representation for those formats". The GDFR will function as the persistent memory of the digital preservation community to ensure that the format knowledge often taken for granted today will remain accessible to the community in future.

The main objective of registry is to support a range of preservation functions such as:

1. Automatic identification of file formats – given a digital object; what format is it?
2. Verification of digital objects compliance to a relevant file format specification – given an object that is supposed to be of format F; is it?
3. Delivery - given an object of format F; how can it be rendered?
4. Transformation – given an object of format F, to what formats can it be converted to?
5. Risk assessment – given an object of format type F; is it at risk of obsolescence?
6. Characterization – Given a format F; what are the representation specifications of F?

7 Classifications of File Formats

7.1 Proprietary formats

Proprietary formats are licensed and their full documentation is not always available. The user cannot modify the format freely. Their license agreement gets changed. There may be restrictions for using and modifying any proprietary file formats. In proprietary formats their format code sequence is not available to the end user.

7.2 Open formats

Open formats are always fully documented, they are not licensed, and the user can freely modify the format structure. One can use open formats for unlimited period. There are no license fees for open formats and there are no patent owners for open formats as well as their full documentation is available permanently. One can also make modifications in these formats.

8 Open Formats in Digital Preservation

Digital preservation has to guarantee the integrity, understandability, originality, authenticity, and accessibility of digital records and data over long term. To enable this, preservation file formats have to fulfill a number of requirements. Their syntactical and semantical specifications must be in public, they must be free of patent and license fees, and ideally they are standardised by a recognised standardization body. Wide use and acceptance improve long term perspective of file formats. Preservation formats must be free of any cryptographical and compression techniques, their specification should be self-contained, and they should be storage media-independent. It becomes clear from the above that, generally speaking, *open formats* are to be preferred over proprietary ones, for digital preservation since they allow for unlimited use without license fees or patent issues, and the fully available documentation eases their future handling.

It is easy to migrate open formats to a newer version as their specifications are available openly hence it is easier to maintain these formats. Thus for preserving digital documents on a long term it is recommended that each digital document should be translated into *standard form* that is independent of any computer system.

9 Recommendation Of Using File Formats in Digital Libraries

It is necessary to consider the following principles while creating any digital document in any format to make the format available for the long term such as (Christensen, 2004):

- The format should be simple to describe, understand and implement
- The format should not depend on specific hardware
- The format should not depend on specific operating systems
- The format should not depend on proprietary software
- The format should be robust against single points of failure

In the digital library literature that are many contributions that suggest which are the file formats more appropriate for preservation issue. The preferred formats should be those that remain usable for a significant amount of time. Four types of basic file formats are considered within the digital library communities: text, image, sound and video and for each categories specific standard format are suggested as described in the table below. These formats can be referred to as preferred formats as they will remain usable over a significant amount of time (Guercio, 2004).

Type of file	Format suggested
Text	Unicode (ASCII), XML and PDF/A, ODF (Open Document Format)
Image	Raster: standard TIFF for master copies(no-compression, high resolution), JPEG for safety copies or distribution, PNG (Portable Network Graphics) Vector: CGM, EPS, DXF, SVG
Sound	Compressionless WAV (PCM-coding)
Video	MPEG, OMF (Open Media Framework)

XML (eXtensible Markup Language) is now accepted as the universal format for data and document exchange, and has actually become the *lingua franca* of the Information Age. XML shows the great promise of data longevity (or future proofing), in a situation in which hardware, software, and network protocols continue to change. XML and PDF are often presented as the two rival formats, with the idea that if you wish to preserve a record long-term you should choose one of the two. PDF and XML are complementary to the point that, in terms of preservation, it is actually better to use them both, rather than choose one of the two. And actually, choosing both standards is also a way of 'sharing the risk': if within a hundred years one of the two formats will no longer be readable, the other might still be so.

PDF cannot be used as an archival format hence long term solutions are needed to keep digital PDF records accessible for a long time length. The PDF/A format has been expressly introduced for the purpose. In order to satisfy preservation criteria the PDF/A attempts to achieve the objectives of device independence, self-containment, and self-documentation. Self-containment is defined as the degree to which a PDF/A file may contain all the necessary resources for performing interpretation and rendering in a reliable way and as expected to, while self-documentation is defined as the degree to which a PDF/A file would document itself in terms of descriptive, administrative, structural, and technical metadata.

10 Conclusion

It is extremely important to standardise the document format by publishing its internal specifications and making them available to public. File formats internal specification information plays an important role in digital preservation activity. It is therefore necessary to use open formats while adding any documents in the digital repository to make these documents available over long term.

References

[1] Abrams, S. L., Seaman, D.(2003). Towards a global digital format registry, Information Technology and Preservation and Conservation workshop, 2003, Berlin.

[2] Brown, A. (2006). Digital Preservation Technical Paper 2. The PRONOM Unique Identifier Scheme. *DPTP-02*, Issue2, p. 1-9. http://www.nationalarchives.gov.uk/aboutapps/pronom/pdf/pronom_unique_identifier_scheme.pdf

[3] Christensen, S. S. (2004) Archival data format requirements. *Stats Bibliotek Report of the Royal Library, Denmark.* http://netarkivet.dk/publikationer/Archival_format_requirements-2004.pdf

[4] Clausen, L. R. (2004). Handling file formats. *Stats Bibliotek*. http://netarchive.dk/publikationer/FileFormats-2004.pdf

[5] Digital preservation management : implementing short-term strategies for long term problems : online tutorial (2004) http://www.library.cornell.edu/preservation/tutorial/presentation/presentation-02.html

[6] Guericio, M., Cappiello, C. (2004). File formats typology and registries for digital preservation, DELOS Report, 54 p. http://www.erpanet.org/events/2004/vienna/Vienna_Report.pdf

[7] Internet Assigned Numbers Authority, “MIME Media Types”, URL:http://www.iana.org/assignements/mediatypes/

[8] The Representation and Rendering Project, University of Leeds: ”Survey and assessment of sources of information on file formats and software documentation, Final Report”, 2003, URL:http://www.jisc.ac.uk/uploaded_documents/FileFormatsreport.pdf

[9] Rothenberg, J. (1995). Ensuring the Longevity of Digital Documents. *Scientific American*, January 1995, p. 42-47.