How to Evaluate the Ability of a File Format to Ensure Long-Term Preservation for Digital Information?

Nicolas Lormant[1], Claude Huc[2], Danièle Boucon[1], Christine Miquel[1]

[1] SILOGIC, ASC-SP, 6 rue Roger Camboulives, BP1133,
31036, Toulouse Cedex 1, France
{nicolas.lormant, daniele.boucon,
christine.miquel}@silogic.fr
[2] CNES, DTS/MID/VM/SVD, 18 avenue Edouard Belin,
31 401 Toulouse Cedex 4, France
claude.huc@cnes.fr

Abstract. Today, it is crucial to consider how file formats affect the use of data stored in long term archives. File formats contain part of the representation information. They describe the data organization, their syntax and sometimes their semantic. Thus, they are playing an essential role in the understanding and in the future access to the data. We present a set of criteria to evaluate the abilities of a file format to be adapted to the preservation needs of a given type of information. The goal of these criteria is to ensure that the chosen archiving format make the data preservation easier, is free of rights, independent of future technological evolutions, and that its implementation cost is as low as possible. We also present the results of two studies based on this set of criteria. The first one is related to PNG (for graphical information), while the second one is related to PDF (for documents). These two studies allowed us to refine and validate the set of criteria.

1 Introduction

Many petabytes of digital data are generated every day throughout the world. This data is not restricted to purely scientific subjects. It covers the entire spectrum of human activity, including science, engineering, justice, culture, government and commerce, and can exist in any digital format, including sound, image, video, documents and files. Most if not all of this data needs to be preserved.

The preservation of data involves more than just storing it as a bit stream on a long-term medium. Representation information also needs to be preserved. This information will be needed in order to interpret the data when it is eventually retrieved from the storage medium.

This is achieved by storing a description of the syntax and semantics of the bit sequence, i.e. by the use of a storage format that forms part of the representation information.

Most data can be stored in a variety of formats. For example, an image could be stored in GIF, JPEG, PNG or TIFF format, and a sound could be stored in AIFF, OGG VORBIS, MP3 or WAV format.

The format in which the data is recorded therefore plays a critical role in preserving the data. However, the format must also make it easy to access and exchange the data. A methodology is thus required for the evaluation of the capabilities of any given format.

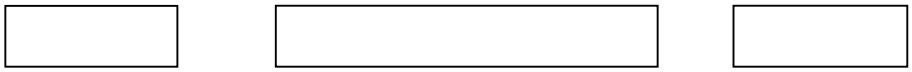

Figure 1. In order to recover the information, both the bit stream and its representation information are necessary.

2 Necessary Condition

The preservation of digital data requires not only the preservation of the actual bit stream, i.e. the physical integrity of the data, but also a knowledge of the information needed in order to interpret these bits. A full and accurate description of the syntax and semantics of the data, i.e. the representation information of this data, must also be available.

The data format used defines most of the representation information.

The necessary condition can therefore be stated as follows:

FORMAT-1	***The format of the data must be fully and explicitly specified and the format specification must be known to the body responsible for preserving the data.***

Failure to comply with this condition may result in a number of problems. For example:

- The irrecoverable loss of the data resulting from a loss of knowledge of its format.
- The need for costly migrations as a result of the use of formats incorporating proprietary information.
- The need to re-enter data recorded in proprietary commercial formats owned by companies that no longer exist.

This condition aims to eliminate all unpublished formats from the list of candidate formats for digital data preservation as the representation information is not under the control of the body responsible for archiving the data.

If this condition is complied with, two possible situations may arise:

- The software needed to read the data and transform it into intelligible form still exists.

– The availability of a complete specification of the format makes it possible to write new software to recover the data at any time.

It should be noted however, that the specification of the data format is only part of the representation information as defined in the OAIS model. The format information should always be included in the representation information, but this should also include all the information needed to understand the semantics of the data.

3 Criteria for evaluating a format

We showed in the preceding section that an exhaustive knowledge of the format is essential for the preservation of digital data. However, it is not of itself sufficient. A number of other rules and recommendations are also required, the extent of which depends on the area of application.

3.1 Principal Rules

3.1.1 Suitability of the Format for Representing all the Information

The format should be chosen to suit the type of information to be preserved. For example, if the aim is to preserve the appearance of a digital document, an image format such as PNG or JPEG would be a good choice. However, if the significant content of the same document needs to be preserved, a format such as the draft ISO PDF/A standard would be a better choice.

It should also be noted that some formats (e.g. JPG) result in a partial loss of information as they make use of lossy compression. Depending on the context of the data preservation, this loss of information may or may not be acceptable.

In general, the more a format takes account of the semantics of the data to be preserved, the easier it will be to interpret the data file. This semantic may be expressed by the capability of the format to structure the data and to introduce high level abstraction.

FORMAT-2	***The format of the data must be suitable for representing the semantics and complexity of the information to be preserved.***

3.1.2 Standard Format

This rule does not formally prohibit the use of published proprietary formats, but this type of format is not advised if others are available.

Firstly, the future existence of the proprietor of the format cannot be guaranteed. Secondly, the commercial policies of the proprietor (e.g. the GIF format) and/or their technological decisions (e.g. the frequent changes to the specification of the PDF format) may change in the light of market conditions and may prevent the archiving body from maintaining continuing access to the data.

In addition to intellectual property rights pertaining to the format itself, consideration should also be given to rights over the underlying technologies, including compression algorithms, character sets, floating point formats, etc. Care should be taken to ensure that no proprietary component has been included in the format in this way.

Finally, some formats have the ability to include private data areas (e.g. the 'private chunks' in the PNG format). If these areas are used, a detailed specification of the content should be kept and the data in these areas should be stored using a standard format.

FORMAT-3	***The use of standard formats is recommended. The use of proprietary elements within a standard format should be avoided.***

Exception In the absence of any standard, a format specified by an open collegiate group should be chosen (e.g. W3C). This may be considered to comply with this rule.

3.1.3 Modification of the Data

Some data may be preserved with a guarantee that the information contained will not be altered in any way. Publications are an example of this type of data preservation. Situations also exist where the data must be preserved for several decades or longer, but where it must remain possible to modify, correct or add to the information contained.

Some digitized text documents (such as those describing scientific data) may fall into this category. If an error is detected in the description of the data, even many years after the data was created, it must still be possible to correct the document describing this data. This rule does not apply in all situations, or to all categories of data, but it may be a considerable constraint if this requirement for modification exists.

FORMAT-4	***If a need to be able to modify the data has been identified, the choice of the data format must take account of this constraint.***

3.2 Additional Recommendations

The aim here is to establish a list of criteria for evaluating formats rather than to state strict rules in the manner described above. In an imperfect world, it may not be possible to comply with all these recommendations simultaneously. Depending on the context of the data preservation project and the constraints affecting it (technological, thematic, budgetary, etc.), a greater or lesser weighting may be accorded to each criterion.

3.2.1 Tools and Creation Facilities

FORMAT-5	***The choice of format must take account of the availability and cost of the tools and other facilities needed to create the data.***

An analysis of the capabilities and complexity of existing tools is essential in order to evaluate both the cost of the process and the workload involved.

3.2.2 Acceptance of the Data by an Archiving Service

FORMAT-6	***It must be possible to verify automatically that a data file complies with the format specification, and with the rules and restrictions specified for data preservation.***

The archiving service responsible for preserving the data must be in a position to check that the data complies with the specified formats. However, these checks may also require human intervention. The tools used to check the data may already exist or may need to be developed.

FORMAT-7	***The ability to extract all or part of the metadata from the data is a definite advantage.***

The automatic extraction of all or part of the metadata usually results in a saving of resources and makes the work of describing the data easier.

3.2.3 Preservation of the Data

The criteria for the compactness and complexity of the formats are usually linked. In general, the more compact a format, the more complex it is. The compactness has an effect on storage costs and input/output performance. The complexity affects the difficulty (and hence the cost) of developing the necessary tools and experience shows that complexity is also a source of errors. In any given project, a compromise must therefore be found between compactness and complexity.

FORMAT-8	***The use of unnecessarily voluminous formats should be avoided.***
FORMAT-9	***A simple format is preferable to a complex format.***

3.2.4 Use of the Data

The widespread use of a format by one or more user communities is usually a good indication of the number of tools available for creating, reading and manipulating the data. It also usually offers a better guarantee of the longevity of the format and multi-platform support. Care must still be taken, however, as some widely used formats are not suitable for archiving purposes. These include formats in which a software layer masks the internal organization of the bit stream.

If the format chosen for archiving is not a format in current use, the complexity of implementing converters should be taken into account.

The ability to extract subsets of the data from the archived file on demand is a definite advantage. In general, consideration should be given to the creation of value added services either from the archiving format itself, or from formats into which the archived data can easily be converted.

FORMAT-10	***Widely recognized and used formats should be preferred.***
FORMAT-11	***The choice of format must take account of the availability and cost of the tools needed to convert between formats and to display the data.***
FORMAT-12	***The choice of format must take account of the availability and potential of developments in value added services.***

4 Case Studies

4.1 The PNG Format

The Portable Network Graphics (PNG) format is a raster image format (the images are represented by a compressed grid of pixels) which is simple, portable, free of all rights and licenses, flexible and robust. All versions of the format are both upwards and downwards compatible. Version 1.2 (the current version) is evaluated here from the point of view of the preservation of non-vector graphics data.

4.1.1 Situation with Reference to the Rules and Recommendations for Data Preservation

Rule	Evaluation	Situation
FORMAT-1	The specifications of the PNG format are published and free of all rights. (http://www.libpng.org/pub/png/spec/iso/)	Good
FORMAT-2	Raster format Lossless compression 'TrueColor' images up to 48 bits/pixel. γ correction. Progressive display. Separation of data (raw image) and display information (filtering, gamma correction, transparency, etc.). Ability to add text information.	Good

FORMAT-3	ISO/IEC 15948: 2004 standard (Published March 3, 2004). W3C recommendation (November 10, 2003). Internal components: ZLIB: Published, free of rights, RFC-1950 (IETF) DEFLATE: Published, free of rights, RFC-1951 (IETF) LATIN-1: ISO 8859-1 UTF-8: ISO/IEC-10646-1	Good
FORMAT-4	Possibility of modifying all or part of the data using an image editor (e.g. Photoshop).	Good
FORMAT-5	Almost all the software currently on the market supports PNG for display, reading, manipulation, etc.	Good
FORMAT-6	The developers of PNG also distribute a set of free utilities for checking the integrity of files, their compliance with the specifications, the content of any 'chunks', the content of metadata, etc. The file structure consisting of chunks makes it easy to develop utilities to check for compliance with specific rules for a given project.	Good
FORMAT-7	The developers of PNG also distribute the pngmeta utility for the extraction of chuncks dedicated to metadata. The file structure consisting of chunks makes it easy to develop utilities to extract the metadata stored in additional chunks.	Good
FORMAT-8	A PNG file is around 30 % smaller than the equivalent GIF file.	Good
FORMAT-9	The PNG format is highly structured and therefore relatively simple.	Good
FORMAT-10	Since 1996, the PNG format has become the preferred alternative to the GIF format. It is supported by all the image editing software packages currently on the market.	Good
FORMAT-11	The conversion of images from the PNG format into other current image formats (JPG, TIFF, PS, etc.) is possible using any of the image editing software packages currently on the market. The creators of PNG also supply a library of utilities including a converter into the 'Portable Pixmap' range of formats.	Good
FORMAT-12	The chunk based structure makes it easy to develop: - Tools to search the metadata. - Tools to extract all or part of the image. - etc.	Good

4.1.2 Restrictions on Use

The structure of the PNG format makes it possible for developers to implement their own methods and tools (filters, corrections, etc.) via the creation of 'private' chunks.

However, as these chunks are not included in the specification, their handling by the graphics tools currently on the market cannot be guaranteed.

Either the private chunk is simply ignored, thereby resulting in a loss of information, or it is misinterpreted. For example, if a private chunk does not comply with the naming conventions and has the same name as a public chunk while containing different data, this may result in severe degradation of the image or even prevent the remainder of the data from being recovered.

The author of a private chunk must therefore take explicit steps to avoid this type of problem by complying with all the naming conventions and registering the necessary additional identification information at the start of the chunk.

The use of critical private chunks is not recommended as it makes the resulting PNG file non-portable.

4.1.3 Conclusion

The PNG format is clearly ideal for the archiving of raster images as it meets all the criteria for data preservation.

It is a free and fully published format. The chunk based design makes it possible to represent all the information needed for the preservation of a raster image, including the associated metadata. It is important to note that the core PNG specifications ensure that every decoder is both upwards and downwards compatible. This means that a PNG file written using the latest version of the format should be readable by all current decoders, and by all earlier decoders.

PNG is an ISO/IEC standard and is recommended by the W3C. This format is also widely supported by currently available applications for reading, creating and converting images (PaintShop, PhotoShop, The Gimp, Netscape, Internet Explorer, etc.). As the format is Open Source, it is also possible to develop your own applications.

The format is well structured and relatively simple, with a large number of basic utilities supplied directly in the libpng library. These include pngcheck for checking chunks in a file, pngmeta for direct access to text chunks, and pnmtopng and pngtopnm for converting Portable Pixmap files to PNG and vice-versa.

4.2 The PDF Format

The PDF format is intended for use in the creation, display and exchange of electronic documents completely independently of the software applications or operating systems used to create the document and the medium used to display it (monitor, printer, etc.).

A PDF document consists of a collection of objects describing the appearance of one or more pages. These objects may be accompanied by interactive components and/or high level application data.

A PDF document may contain any combination of text and images.

A PDF file contains the objects defining the PDF document together with the associated structural information in a single sequence of bytes.

Version 1.4 (the current version is 1.6) is evaluated here from the point of view of the preservation of documentary data.

4.2.1 Situation with Reference to the Rules and Recommendations for Data Preservation

Rule	Evaluation	Situation
FORMAT-1	The specifications of the PDF format are published but remain the property of the Adobe corporation. http://partners.adobe.com/public/developer/pdf/index_reference.html	Correct
FORMAT-2	The richness of the PDF format enables it to represent all the information capable of being contained in a document.	Good
FORMAT-3	The PDF format is not a standard. The specification changes rapidly (~18 months between revision). In addition, a large number of proprietary components may be included within a PDF document. Standard formats do, however, exist in the form of PDF/X (ISO 15929 and ISO 15930), developed to meet the needs of the pre-press industry, and PDF/A (ISO 19005-1 in course of publication) designed for the long term preservation of data.	Poor
FORMAT-4	At the present time, PDF documents can only be modified by using commercial software packages that are often expensive. The modification of a 'non-linearized' document may damage the integrity of the data to be preserved.	Poor
FORMAT-5	A wide range of free and commercial software is available for the creation of PDF documents.	Good
FORMAT-6	There currently exists no tool of this type for the PDF format (or for PDF/A). However, one could be developed in the same way as was done for the PDF/X format (Certified PDF).	Poor
FORMAT-7	The extraction of metadata is possible providing that only 'tagged PDF' files are used. It may be the case that XML would be more useful in this regard.	Insufficient
FORMAT-8	A PDF document is smaller than the equivalent Word document.	Good
FORMAT-9	PDF is certainly not a simple format, but it is no more complex that the few other candidates for the preservation of documentary data.	Insufficient

FORMAT-10	PDF is currently a de facto standard. It is widely used by a large number of communities (pre-press, pharmaceuticals, etc.).	Good
FORMAT-11	A large number of software suites are available for the conversion of PDF files into other common formats (text, Microsoft Word, Excel, etc.). However, the results are not always perfect and these software suites are relatively expensive.	Insufficient
FORMAT-12	The various software suites on the market all offer different value added services. It is also likely that the growing popularity of the format and the future standardization of the PDF/A format will result in the appearance of new value added applications.	Good

4.2.2 Restrictions on Use

The many functionalities available within the PDF format may make it difficult if not impossible to guarantee the preservation of documents. This is particularly the case with embedded multimedia objects in PDF documents. The inclusion of multimedia content such as sound or video sequences should be prohibited if the documents are to be preserved. In general, the insertion of files into a PDF document is undesirable as it is difficult or impossible to assess the suitability of the file format or contents for preservation.

It is also dangerous to allow the referencing of external objects such as files or hypertext links. Hypertext links may be permissible if they are inactive (i.e. if it is not possible to connect to a URL or email address directly from within the PDF document). The use of pointers to external files should be forbidden.

In order to avoid future problems with rights, the following points should also be observed:

- Limit the use of character fonts to standard fonts only.
- Limit the use of compression algorithms to standard compression algorithms only (e.g. CCITT 4).
- Prohibit encryption of the contents. Encryption algorithms are usual subject to licensing. Also, by definition, encrypting a document constitutes a major obstacle to its preservation.
- Ensure that the character fonts used are embedded in the document.
- Ensure that the color spaces used are independent of the terminals used to create or display the document.
- Avoid the use of hidden content.
- Avoid the use of transparency.
- If forms are present in the document, these must not execute any action, regardless of their type.

Finally, in order to facilitate the implementation of value added services, it is strongly recommended that the use of the PDF format is limited to the use of the linearized and tagged versions (use of metadata described using XMP).

4.2.3 Conclusion

The qualities of the PDF format are undeniable and its widespread use by large multinational organizations demonstrates its position as one of the major formats in use at the present time. PDF is currently the only document format on the market that can represent composite data independently of the terminals and operating systems used. The PDF Acrobat reader is distributed without charge by Adobe and the specifications of the various versions of the software are published.

Having said that, the situation relating to the use of PDF as a solution for the long term preservation is far from idyllic as, in the end, PDF is just too powerful. It can be seen from the usage restrictions that a large number of the functionalities of the format are incompatible with the needs of archiving.

It is also clear that the dependence on Adobe is very high and there is no guarantee that future versions of the Acrobat reader will be able to recover documents prepared at the present time. There is also no guarantee that Adobe will continue with its current commercial policy (published specifications and free Acrobat reader).

Work on the PDF/A standard has already revealed the extent of the restrictions that will have to be applied to the current format. There are currently no tools available to check these restrictions and to certify PDF/A documents in the same way as those available for PDF/X (e.g. Certified PDF).

It is also clear that the requirement to make the content self-documenting by labeling and structuring the documents ('tagged PDF') is resulting in PDF documents becoming structured in the same way as XML files. PDF/A will probably become the best archiving format to be based on PDF, but other solutions not based on PDF may evolve such as XML.

5 Conclusion

We have shown that the choice of a data storage format for preserving and accessing digital data is an important one. The format itself contains some of the representation information, and this results in the access to the data being highly dependent on the rapidly changing technological environment.

In order to minimize this dependence as far as is possible, we have proposed a set of rules intended to guarantee control of the specifications of the storage format throughout the life of the stored data, thus ensuring that the data can be accessed at any time during this period.

We have also proposed a set of recommendations to take account of the problems facing archiving services, including storage volume, software maintenance, ability to check the information, and value added services.

It is unlikely that all these criteria can be met simultaneously, but they do enable a choice to be made of the best available compromise between the requirement for preservation and the need to access the information in any given context.