The National Archives

Digital Preservation Guidance Note:

1

Selecting File Formats for Long-Term Preservation

Document Control

Author:	Adrian Brown, Head of Digital Preservation Research
Document Reference:	DPGN-01
Issue:	2
Issue Date:	August 2008

Contents

1 Introduction

This document is one of a series of guidance notes produced by The National Archives, giving general advice on issues relating to the preservation and management of electronic records. It is intended for use by anyone involved in the creation of electronic records that may need to be preserved over the long term, as well as by those responsible for preservation.

This guidance note provides information for the creators and managers of electronic records about file format selection. Please note that The National Archives does not specify or require the use of any particular file formats for records which are to be transferred. Choice of file format should always be determined by the functional requirements of the record creation process. Record creators should be aware however, that long-term sustainability will become a requirement, both for ongoing business purposes and archival preservation. Sustainability costs are inevitably minimised when this factor is taken into account prior to data creation. Failure to do so often makes later attempts to bring electronic records into a managed and sustainable regime an expensive, complex and, generally, less successful process.

This guidance note sets out a range of criteria the aim of which is to help data creators and archivists make informed choices about file format issues.

2 Selection issues

File formats encode information into forms that can only be processed and rendered comprehensible by very specific combinations of hardware and software. The accessibility of that information is therefore highly vulnerable in today's rapidly evolving technological environment. This issue is not solely the concern of digital archivists, but of all those responsible for managing and sustaining access to electronic records over even relatively short timescales.

The selection of file formats for creating electronic records should therefore be determined not only by the immediate and obvious requirements of the situation, but also with long-term sustainability in mind. An electronic record is not fully fit-for-purpose unless it is sustainable throughout its required life cycle.

The practicality of managing large collections of electronic records, whether in a business or archival context, is greatly simplified by minimising the number of separate file formats involved. It is useful to identify a minimal set of formats which meet both the active business needs and the sustainability criteria below, and restrict data creation to these formats.

This guidance note is primarily concerned with the selection of file formats for data *creation*, rather than the conversion of existing data into 'archival' formats. However, the criteria described are equally applicable to the latter.

Selecting file formats for migration introduces some additional issues. Formats for migration must meet the requirements for both preservation of authenticity and ease of access. For example, the data elements of a word-processed document could be preserved as plain ASCII text, together with any illustrations as separate image files. However, this would result in a loss of structure (e.g. the formatting of the text), and of some context (e.g. the internal pointers to the illustrations).

There is also a subtly different conflict between the need for data formats that can be accessed and those that can be re-used. From a preservation and re-use perspective, data must be maintained in a form that can be processed. For the purposes of access, however, control of the formatting may well be the most important criteria, and in some cases it may be desirable for the data not to be able to be processed by end users. In some cases it may only be possible to reconcile these differences by using different formats for preservation and presentation purposes.

The following criteria should be considered by data creators when selecting file formats:

- Ubiquity
- Support
- Disclosure
- Documentation quality
- Stability
- Ease of identification and validation
- Intellectual Property Rights
- Metadata Support
- Complexity
- Interoperability
- Viability
- Re-usability

These criteria are elaborated in the following sections:

2.1 Ubiquity

The laws of supply and demand dictate that formats which are well established and in widespread use will tend to have broader and longer-lasting support from software suppliers than those that have a niche market. There is also likely to be more comprehensive community support amongst users. Popular formats are therefore preferable in many cases.

2.2 Support

The extent of current software support is a major factor for consideration. The availability of a wide range of supporting software tools removes dependence on any single supplier for access, and is therefore preferable. In some cases however, this may be counterbalanced by the ubiquity of a single software tool.

2.3 Disclosure

Those responsible for the management and long-term preservation of electronic records require access to detailed technical information about the file formats used. Formats that have technical specifications available in the public domain are recommended. This is invariably the case with open standards, such as JPEG. The developers of proprietary formats may also publish their specifications, either freely (for example, PDF), or commercially (as is the case with the Adobe Photoshop format specification, which is included as part of the Photoshop Software Development Kit). The advantages of some open formats may come at the cost of some loss in structure, context, and functionality (e.g. ASCII), or the preservation of formatting at the cost some reusability (e.g. PDF). Proprietary formats frequently support features of their creating software, which open formats do not. The tension between these needs is sometimes unavoidable, although the range and sophistication of open formats is increasing all the time. The use of open standard formats is however highly recommended wherever possible.

2.4 Documentation quality

The availability of format documentation is not, in itself, sufficient; documentation must also be comprehensive, accurate and comprehensible. Specifically, it should be of sufficient quality to allow interpretation of objects in the format, either by a human user or through the development of new access software.

2.5 Stability

The format specification should be stable and not subject to constant or major changes over time. New versions of the format should also be backwards compatible.

2.6 Ease of identification and validation

The ability to accurately identify the format of a data file and confirm that it is a valid example of that format, is vital to continued use. Well-designed formats facilitate identification through the use of 'magic numbers' and version information within the file structure. The availability of tools to validate the format is also a consideration.

2.7 Intellectual Property Rights

Formats may utilise technologies encumbered by patents or other intellectual property constraints, such as image compression algorithms. This may limit present or future use of objects in that format. In particular, 'submarine patents' (when previously undisclosed patent claims emerge), can be a concern. Formats that are unencumbered by patents are recommended.

2.8 Metadata Support

Some file formats make provision for the inclusion of metadata. This metadata may be generated automatically by the creating application, entered by the user, or a combination of both. This metadata can have enormous value both during the active use of the data and for long-term preservation, where it can provide information on both the provenance and technical characteristics of the data. For example, a TIFF file may include metadata fields to record details such as the make and model of scanner, the software and operating system used, the name of the creator, and a description of the image. Similarly, Microsoft Word documents can include a range of metadata to support document workflow and version control, within the document properties. The value of such metadata will depend upon:

- The degree of support provided by the software environment used to create the files,
- The extent to which externally stored metadata is used in its place. (For example if records are stored within an Electronic Records Management System).

In general, formats that offer metadata support are preferable to those that do not.

2.9 Complexity

Formats should be selected for use on the basis that they support the full range of features and functionality required for their designated purpose. It is equally important, however to avoid choosing over-specified formats. Generally speaking the more complex the format, the more costly it will be to both manage and preserve.

2.10 Interoperability

The ability to exchange electronic records with other users and IT systems is also an important consideration. Formats that are supported by a wide range of software or are platform-independent are most desirable. This also tends to

support long-term sustainability of data by facilitating migration from one technical environment to another.

2.11 Viability

Some formats provide error-detection facilities, to allow detection of file corruption that may have occurred during transmission. Many formats include a CRC (Cyclic Redundancy Check) value for this purpose, but more sophisticated techniques are also used. For example, the PNG format incorporates byte sequences to check for three specific types of error that could be introduced. Formats that provide facilities such as these are more robust, and thus preferable.

2.12 Re-usability

Certain types of data must retain the ability to be processed if they are to have any re-use value. For example, conversion of a spreadsheet into PDF format effectively removes much of its ability to be processed. The requirement to maintain a version of the record that can be processed must also be considered.

3 Evaluating formats: sources of information

A variety of practical information sources are available to support the evaluation of formats in accordance with these criteria. PRONOM, The National Archives' technical registry, is particularly designed as an impartial and authoritative source of advice on this subject, and is freely available online at www.nationalarchive.gov.uk/pronom/
The following sections indicate sources for evaluating formats:

3.1 Ubiquity

The relative popularity of a format tends to be a comparatively subjective measure, but is likely to be widely known within a particular user community.

3.2 Support

This requires consideration of the number of software tools which currently support the format, and of the ubiquity of those tools. In PRONOM, the level of software support for a format may be assessed using the 'Compatible software' search facility on the 'File format' tab. This will return a list of software known to support a given format. This can be supplemented by additional research, as PRONOM may not provide comprehensive coverage for all formats. In addition, this factor must be considered in conjunction with the ubiquity of the format (see 3.1).

3.3 Disclosure

In PRONOM, the degree of disclosure may be ascertained from the ‘Availability’ field on the ‘Documentation’ tab of a format record.

3.4 Documentation quality

PRONOM provides links to known documentation that is available for a format. In PRONOM, an initial assessment of the comprehensiveness of available documentation may be gained from the ‘Disclosure’ field on the ‘Summary’ tab of a format record. The authoritativeness may be ascertained from the ‘Type’ field on the ‘Documentation’ tab of a format record. A detailed judgement of documentation quality will require evaluation of the documentation itself.

3.5 Stability

The stability of a format may be judged by its age, and the frequency with which new versions are released. The number of versions of a format may be determined in PRONOM by searching on the format name: all known versions of the format will be listed.

PRONOM also records the dates on which versions of formats were released and withdrawn from current support – these may be used to judge the longevity of each format version.

3.6 Ease of identification and validation

In PRONOM, the availability of existing identification and validation tools for a format may be determined by using the ‘Compatible software’ search facility on the ‘File format’ tab. The search can then be filtered by software which can ‘identify’ or ‘validate’ a given format respectively.

3.7 Intellectual Property Rights

In PRONOM, known IPR restrictions for a format will be listed under the ‘Rights’ tab of any format record.

3.8 Metadata Support

Determining the degree of metadata support offered by a format may require a review of its technical documentation. PRONOM may be of assistance for locating such documentation (see 3.4).

3.9 Complexity

Complexity is a subjective measure, and can generally only be determined with reference to the relevant technical documentation. PRONOM may be of assistance for locating such documentation (see 3.4).

3.10 Interoperability

In PRONOM, the general level of interoperability for a given format may be judged by reviewing the number of software products which are available to create or render files in that format (see 3.2).

3.11 Viability

In PRONOM, the provision of error detection and correction mechanisms may be noted in the 'Description' field of the format record. Otherwise, this will need to be determined with reference to the relevant technical documentation. PRONOM may be of assistance for locating such documentation (see 3.4).

3.12 Re-usability

Re-usability is a complex measure and will vary depending on the requirements of a particular community of users. It can generally only be determined with reference to the relevant technical documentation. PRONOM may be of assistance for locating such documentation (see 3.4).

4 Conclusion

There are many issues to be considered when selecting file formats extending beyond the immediate and obvious requirements of the situation. It may not be possible to select formats that meet all criteria in every case; however, new formats and revisions of existing formats are constantly being developed. This guidance note should assist data creators to make informed decisions about file format selection from the ever-changing choices available.

The adoption of sustainable file formats for electronic records brings benefits to data creators, data managers and digital archivists. Selection decisions informed by the criteria described above will greatly enhance the sustainability of the records created.