



**Digital Preservation
Guidance Note:**

1

Selecting File Formats for Long-Term Preservation

Document Control

Author: Adrian Brown, Digital Archives Analyst

Document Reference: DPGN-01

Issue: 1

Issue Date: 19 June 2003

Contents

1	INTRODUCTION	4
2	SELECTION ISSUES	4
2.1	Open Standards	6
2.2	Ubiquity.....	6
2.3	Stability	6
2.4	Metadata Support	6
2.5	Feature Set.....	7
2.6	Interoperability	7
2.7	Viability	7
2.8	Authenticity	7
2.9	Processability	7
2.10	Presentation	7
3	CONCLUSION.....	8

1 Introduction

This document is the first in a series of guidance notes produced by the Digital Preservation Department of The National Archives, giving general advice and guidance on issues related to the preservation and management of electronic records. It is intended to be used by anyone involved in the creation of electronic records which may need to be preserved over the long term, and by those responsible for preservation.

This guidance note provides advice on general issues which should be considered by the creators and managers of electronic records when selecting file formats for use. The National Archives does not specify or endorse 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-creating process. However, record creators should be aware that long-term sustainability may be a requirement, both for ongoing business processes and archival preservation. Sustainability costs are inevitably minimised when these factors are taken into account prior to data creation – attempts to bring electronic records into a managed and sustainable regime after the fact tend to be expensive, complex and, generally, less successful.

This guidance note therefore describes a range of criteria which will help data creators and archivists to make informed choices about file format issues.

2 Selection issues

File formats encode information into a form which 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 by longer-term considerations. An electronic record is not fully fit-for-purpose unless it is sustainable throughout its required life cycle.

The practicality of managing any large collection of electronic records, whether in a business or archival context, is greatly simplified by minimising the number of separate formats involved. It is therefore highly desirable to identify the minimum 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 in the latter case.

Selecting file formats for migration introduces some additional issues. Such formats 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 *processable* and *formatted* data formats. From a preservation and re-use perspective, data must be maintained in a processable form. For the purposes of access, however, control of the formatting may well be the most important criteria, and in some cases it may actually be desirable for the data not to be processable by end users. In some cases it may only be possible to reconcile these differences by using different formats for preservation and presentation purposes. Additional criteria to be considered when selecting file formats for migration are therefore also provided.

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

-  Open standards
-  Ubiquity
-  Stability
-  Metadata Support
-  Feature Set
-  Interoperability
-  Viability

The following additional criteria should be considered for migration:

-  Authenticity
-  Processability
-  Presentation

These criteria are elaborated in the following sections:

2.1 Open Standards

Those responsible for the management and long-term preservation of electronic records require access to detailed technical information about the file formats in which those records are preserved. Formats for which the technical specification has been made available in the public domain are therefore 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 come at the cost of some loss in structure, context, and functionality (e.g. ASCII), or the preservation of format at the cost of processability (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. However, the use of open standard formats is highly recommended wherever possible.

2.2 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 which only have a niche market. Popular formats, which are supported by as wide a range of software as possible, are therefore to be preferred where possible.

2.3 Stability

It is desirable that 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.

2.4 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 thereof. This metadata can have enormous value both during the subsequent 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 2000 documents can include a range of metadata to support document workflow and version control, within the document properties. The value of such metadata provision will depend both upon the degree of support provided by the software environment used to create the files, and the extent to which externally-stored metadata is used in its place (for example, if records are stored within an Electronic Records Management System). However, in general, formats which offer metadata support are preferable to those which do not.

2.5 Feature Set

It is a given that formats should be selected which support the full range of features and functionality required for their designated purpose or business process. However, it is equally important to avoid choosing over-specified formats. In general, the more complex the format, the more costly it will be to manage and preserve.

2.6 Interoperability

The ability to exchange electronic records with other users and IT systems is frequently an important consideration. Formats which are supported by a wide range of software or are platform-independent are therefore highly desirable in many situations. This feature also tends to support the long-term sustainability of the data by facilitating the migration of the data from one technical environment to another.

2.7 Viability

Some formats provide error-detection facilities, to allow detection of file corruption which 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 which could be introduced. Formats which provide such facilities are more robust, and therefore preferable.

2.8 Authenticity

The authenticity requirements for electronic records are complex, and beyond the scope of this guidance note. However, the creators of records for which authenticity must be demonstrated, either now or at some point in the future, must consider this when selecting file formats. Broadly, the format must preserve the content (data and structure) of the record, and any inherent contextual, provenance, referencing and fixity information.

2.9 Processability

Certain types of data must retain their processability to have any reuse value, even though the requirements of authenticity demand that the archived version must not be altered through reprocessing. For example, conversion of a word-processed document into PDF format effectively removes its processability. The requirement to maintain a processable version of the record must therefore be considered.

2.10 Presentation

The formatting of a digital record may have significant information value. If the authenticity of an electronic record requires preservation of its original 'look and feel' (for example, fonts, colours and layout), then the ability of a file format to support this through migration will be a crucial consideration.

3 Conclusion

There are many issues to be considered when selecting file formats, which extend beyond the immediate and obvious requirements of the situation. It may not be possible to select formats which meet all these 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 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 in this guidance note will greatly enhance the sustainability of the records created.