Preservation & Storage Formats for Repositories

Overview

Formats control whether digital content is to be accessible now and in the longer term. Institutional Repositories (IRs), which provide access to and store digital objects produced by many creators, will need to manage a range of formats[1]. This briefing paper explains how formats affect preservation, considers which formats repositories should use for deposit and storage, and describes the practical steps repositories can take to produce an initial preservation plan.

How are formats used?

Digital objects are produced, viewed and reused using application programs such as a word processor or image editor. The files produced are encoded in some way to represent characters, layout and other features. The rules of the encoding are defined by the format of the object; more than one format may be involved for complex objects such as web pages or learning objects. Applications are often closely tied to formats – some are only produced by one application, some can only be read by a single application, whereas others have many applications which can read and write the format. For example, Microsoft Word can be used to produce the document (.doc) format but other applications can read this, and Word can read and write other formats Microsoft Publisher file, by contrast, are typically only readable by MS Publisher. JPEG images can be read and written by many applications.

Why are formats important for preservations?

Problems can arise with any format once applications no longer exist to easily access it. Application-specific formats can cause problems from the outset when users try to open an object without access to the application used to create it, or without the correct version of an application. This problem increases over time, that is, it becomes harder to open documents in their original format if the application has changed or no longer exists.

It follows that some objects risk becoming inaccessible – we have the bits but they no longer mean anything. This is why formats are a primary focus for preservation actions, and why repositories need to be aware of the formats of the digital objects they acquire, store and distribute. These formats do not need to be the same.

Which deposit formats should an IR allow?

There are many different types of digital objects (e.g. texts, images, videos), and many different applications for producing them. There are also different views on which formats are the most 'preservable'; a repository needs to be able to transform the objects given to it into such a preservable format, and ideally the deposit format is a preservable format. There is one format that an IR should always commit to obtaining: the author's source format, even if it is not an ideal preservable format. That is, the version produced by the author directly from the application used at the time of completion.

By requiring authors to submit the source format for *preservation*, the repository can then convert to its preferred *presentation* format, which could be PDF, if that is different from the source format. The repository should aim to automate this process (and record that it took place.)

Which formats should repositories commit to support in the long term?

The key phrases that describe the ideal longer-term storage format are *widely-supported* and *open standard,* the latter meaning the specification is freely available and implementable. Consequently it is more likely that applications to view and use such formats will be available at any given time, since viewers can be developed by the wider community of users with an interest in the format, and not just the original application developer. Open standard formats include PDF/A, JPEG and OpenDocument format (ODF) [2], an XML file format for electronic office documents. For repositories this approach is likely to prove over-simplistic because of the use of popular applications, which are not always open standards (but are widely-used), and the dependence of repositories on their authors for content. The need for content should come before placing extra requirements on the way authors produce and deliver what they have created.

There is no single answer to this question about which storage formats to support. The most flexible approach is to require deposit in the formats that authors produce, convert to preservation formats that are long-lived, use presentation formats as required, and produce an informed plan for long-term storage formats.

How can repositories plan preservation & storage formats?

Repositories need to take three steps to produce a plan for preservation & storage formats:

1. Accurately identify the formats of objects stored in the repository and those likely to be deposited

2. Adopt a trusted and current list of storage formats and their prospects for preservation

3. Develop a plan of action based on the findings of 1 and 2

For 1 and 2 you can find tools and services on the web and tools are emerging to help with 3. Format identification tools such as DROID[3] are open source and can be downloaded and used as part of the deposit process. Projects such as PRESERV and KEEPIT have explored providing such

services remotely. JISC have recently funded EDINA to deliver key shared infrastructure services[4]. In deciding which formats to support there are a number of reference sources, notably Library of Congress[5].

It is important to note that formats are always changing, so 1 needs to be revised fairly often, but 2 and 3 should be longer-lasting documents. The critical step is combining these sources to produce a viable action plan for the repository, and this is where specialist knowledge may play a role.

To ensure plans are up-to-date and properly applied repositories may want to seek preservation services from trusted sources. At minimum, these can alert them when stored formats are approaching obsolescence, and provide some years warning to allow plans to be executed. A number of projects are investigating a more complete provision of services and these projects involve prospective service providers[5]. In the meantime, repositories can plan for preservation, particularly by addressing preservation and format issues within the overall repository framework. In the widest sense, preservation is about preserving intellectual access to material – which is what repositories are all about.

Further information:

[1]**Wikipedia article on file formats**
http://en.wikipedia.org/wiki/File_formats

[2]**OpenDocument XML.org**
http://opendocument.xml.org/

[3]**DROID, The National Archives**
http://www.nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm

[4]**UK RepositoryNet+**
http://edina.ac.uk/projects/ukrnplus_summary.html

[5]**Sustainability of Digital Formats: Planning for Library of Congress Collections**
http://www.digitalpreservation.gov/formats/

Repositories Support Project
http://www.rsp.ac.uk/
The Repositories Support Project (RSP) aims to coordinate and deliver good practice and practical advice to HEIs to enable the implementation, management and development of digital institutional repositories.