REPOSITORIES SUPPORT PROJECT

JISC RepositoryNet

Preservation & Storage Formats for Repositories

support@rsp.ac.uk

Overview

Formats matter if digital content is to be accessible now and preservable 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. 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 documents are produced, in one form or another, using an application program such as a word processor. These documents are encoded with information to represent characters, layout and other features. The rules of the encoding are defined by the chosen format of the document. Applications are often closely tied to formats, e.g. Microsoft Word can be used to produce the document (.doc) format, Adobe Acrobat produces the portable document format (.pdf). These may not be the only formats that an application can produce, e.g. Word can also produce Rich Text Format (.rtf), and formats may not be exclusive to one application.

Why are formats important for preservation?

Problems with application-specific formats can arise when users try to open a digital document without access to the application that was used to create it, or without the correct version of an application. This is most likely to happen when opening a document created by someone else. 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.

If applications and formats can change over time, it follows that some risk becoming obsolete. 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 store.

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', and therefore which formats a repository should allow to be deposited. There is one format that an IR should always commit to obtaining: the author's source format. That is, the version produced by the author directly from the application used at the time of completion.

The most common example of deviating from this approach is a requirement for authors to deposit PDF, which is not an authored format - it is created by converting from another format. 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. It is likely this conversion can be automated and, in the process, documented.

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

The key phrase that describes the ideal longer-term storage format is *open standard*, 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 OpenDocument format (ODF)[1], 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, 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 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
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. Format identification tools such as DROID[2] are open source and can be downloaded and used as part of the deposit process. Alternatively, a repository registry service, ROAR, has format profiles in development for over 200 repositories[3]. In deciding which formats to support there are a number of reference sources, notably Library of Congress[4].

It is important to note that formats are always changing, so 1 and 2 really need to be dynamic sources that keep up with these changes. 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. Although the services currently on offer will not fulfil all three steps, 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 policy framework. Repository policy should not begin with preservation, but when preservation policy emerges it will invariably include analysis of formats.

References & further information:

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

[2] **DROID, The National Archives**
http://droid.sourceforge.net/wiki/index.php/Introduction/

[3] **Example format profile, ROAR**
http://roar.eprints.org/index.php?action=profile&url=http://dspace.anu.edu.au/

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

[5] **Hitchcock, S., et al. (2007) Digital Preservation Service Provider Models for Institutional Repositories,** *D-Lib Magazine,* Volume 13 Number 5/6, May/June 2007
http://www.dlib.org/dlib/may07/hitchcock/05hitchcock.html

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.