FEATURES Section

Selecting Formats for Digital Preservation: Lessons Learned during the Archivematica Project
Evelyn Peters McLellan

The Archivematica project was launched a year ago as a follow-up to a report entitled *Towards an Open Source Repository and Preservation System* which was released by the UNESCO Memory of the World Sub-Committee on Technology in 2007. The report surveyed then existing open-source tools for digital preservation, concluding that although a wide variety of such tools were available, they were neither comprehensive nor integrated enough to provide a complete ingest-to-access environment for preserving digital records. The report recommended that UNESCO support "the aggregation and development of an open source archival system, building on and drawing together existing open source programs." This is the goal of Archivematica, a preservation system being developed with funding from UNESCO and in collaboration with the City of Vancouver Archives. Software development has thus far been led by Artefactual Systems, a Canadian open-source software company specializing in information management systems for archives and libraries.

Archivematica is an integrated environment of open-source tools which can be deployed from a single installation on any operating system. (The system is based on Ubuntu Linux but can be implemented in a virtualized environment, allowing it to be run on top of any number of host operating systems such as Microsoft Windows.) The software and the source code are freely available online under a GPL license. The system is based on a detailed use case analysis of the ISO Open archival information system (OAIS) functional model and supports best practice metadata standards such as PREMIS, METS, Dublin Core, EAD and ISAD(G). Detailed workflow documentation assists the user to move an Information Package through submission, integrity checking, identification and validation, normalization, packaging into an Archival Information Package, storage, and provision of access. One of the key components of this process is normalization, the conversion of digital objects into a small number of standard preservation-friendly formats on ingest. This is the part of the Preservation Planning function of OAIS which "receives archive approved standards and migration goals from Administration," including "format standards," the goal being to ensure that the preserved objects remain accessible and usable in the future despite issues of technological obsolescence and incompatibility.

Archivematica supports emulation preservation plans by preserving original bitstreams, and it supports migration preservation plans by monitoring at-risk file formats and providing a process to migrate them at a future date. Nevertheless, Archivematica's default preservation strategy is to normalize digital objects into preservation formats upon ingest in order to make best use of the limited time that organizations will have to process and monitor large, diverse collections of digital objects.

Building normalization paths into the software requires choosing target formats and integrating open-source tools to perform the migrations. The choice of formats is based on four basic criteria which will be familiar to many of those who have experience with digital preservation:

1. The specification must be freely available.
2. There must be no patents or licenses on the format.
3. Other established digital repositories should be using or have endorsed the format.
4. There should be a variety of writing and rendering tools available for the format.

Selection of formats has been an iterative process of researching best practices, testing normalization tools, and, as far as possible, comparing before and after results of conversions by measuring significant properties. During this process it was found that selecting target formats based on the first three criteria is not difficult, since a great deal of research has been done on the subject and de facto standards have been proposed and in some cases implemented. However, there are some significant challenges with the fourth criterion. Specifically, for this project there needs to be open-source tools available for conversion from original formats. This is to ensure that the tools can be integrated and distributed with the existing tools in the system, which must remain entirely free of software license restrictions and costs. Another important consideration is that they offer a Linux command-line interface to enable full ingest process automation. Thus far, it has been the Archivematica team's experience that the scarcity of some types of tools and inadequacy of others has made the process of selection considerably more difficult, illustrating the challenges that can arise when moving from the realm of the ideal to the realm of the practical.

Moving image files provide an example of some of the difficulties involved. A consensus seems to be building in the research community that Motion JPEG2000 is the desired target format because it provides mathematically lossless, wavelet compression. (See, for example, *Lossless Video Compression for Archives: Motion JPEG2k and Other Options*.) Motion JPEG2000 was adopted as an ISO standard in 2001 (ISO/IED 15444-3); however, during the nine years of its existence only a handful of tools have emerged to convert to it, and these are proprietary and not designed for use with Linux. Most heritage institutions that are converting to Motion JPEG2000 are converting directly from analog video using specialized hardware. There are a number of open-source Linux-based tools for converting moving image files from one digital format to another, most notably FFmpeg and Avidemux, but they do not currently encode to Motion JPEG2000. For these reasons, the default normalization path for moving image files in Archivematica is MPEG-2. MPEG-2 is a reasonably well-accepted preservation format; for example, the Library of Congress (with some reservations and qualifications) and Library and Archives Canada both recommend it. However, MPEG-2 compression is not entirely lossless. If and when an appropriate Motion JPEG2000 normalization tool becomes available, it will be added to Archivematica and users will then have the option to migrate the original moving images to Motion JPEG2000 and discard the existing MPEG-2 versions.

Even more problematic are Microsoft Office files. Theoretically, it is a simple task to choose the XML-based Open Document Format (ODF) and PDF/Archival (PDF/A), both ISO standards, as preservation formats. (The Archivematica team briefly considered using Office Open XML, the Microsoft XML format that was approved as an ISO Standard in 2008. However, there are no open-source tools that convert to the format at present, and because at over 6,000 pages the standard is so complex and lengthy it is unlikely that any such tools will emerge in the near future.) The user could choose to use one or the other of ODF or PDF/A or both. There are Linux-based tools to convert to ODF, including Xena, the National Archives of Australia's bulk normalization tool. To normalize office documents, Xena calls on OpenOffice to manage the conversion of a Microsoft Office document to ODF. Unfortunately, because OpenOffice has had to rely on reverse engineering of proprietary Microsoft specifications to map to ODF, the formatting of the converted document often differs from the original; the differences may be minor but they can change the overall look and feel of the document, which may call into question the authenticity of the conversion.

This problem becomes even more critical with PDF/A, since one of the most compelling reasons for using that format is to provide an accurate visual representation of the original. There are very few open-source bulk normalization tools to convert to PDF/A, and those that do (such as OpenOffice) must, once again, rely on reverse engineering of closed specifications in order to perform the

conversion. When OpenOffice opens a Microsoft document, the document renders with some changes to formatting; the PDF/A is created from this altered rendering.

Conversion using a plug-in that works directly with the native application is the most direct path to success with either ODF or PDF/A. The proprietary Adobe Acrobat Distiller, for example, works directly with Microsoft software to produce visually true conversions to PDF/A. Similarly, Sun Microsystems has produced a free (but not open-source) plug-in to convert documents to ODF from within Microsoft applications. However, the problem remains that there is no way to integrate these tools into any freely available open-source digital preservation system. Following the example set by the National Archives of Australia, the Archivematica team has chosen to build default normalization to ODF into its system and is currently testing conversions to determine which tool to use. As with moving image files, acceptable conversions might not be possible for the immediate future and bulk Microsoft Office migration processes may need to be run at a later time when better tools become available. PDF/A remains under consideration as a preservation format, but more time is needed to evaluate available tools before it can be built into the system as a default.

Fortunately, some types of files lend themselves more easily to normalization. Numerous raster image formats, for example, can be converted easily to uncompressed TIFF 6.0 using ImageMagick, and FFmpeg does a good job of converting audio files to uncompressed (LPCM) WAVE files. Both of these are well accepted preservation formats in the library and archives community. However, the lack of tools for other kinds of digital objects means that, with regard to normalization, any open-source integration of digital preservation tools must remain a work in progress. The driving goal behind the Archivematica project has been to lower the barriers to digital preservation for institutions which may have limited technical and financial resources. The best way to do this is to provide a complete system that can be freely downloaded, used, distributed, and modified by any individual or institution. At this time, Archivematica incorporates best-possible normalization paths, and the team has adopted an agile development process in which the system incorporates new tools as soon as they become available. This is the most effective way to work within current limitations, and it is the most realistic means of achieving UNESCO's goal of bringing open-source digital preservation capability to institutions all over the world.

Evelyn Peters McLellan (evelyn@artefactual.com) is Systems Archivist at Artefactual Systems Inc., Vancouver, Canada.

Relevant Links

Archivematica project
www.archivematica.org

Archivematica Media Type Preservation Plans
www.archivematica.org/wiki/index.php?title=Media_type_preservation_plans

Avidemux
fixounet.free.fr/avidemux/

Bradley, Kevin, et al. Towards an Open Source Repository and Preservation System: Recommendations on the Implementation of an Open source Digital Archival and Preservation System and on Related Software Development. UNESCO Memory of the World Sub-Committee

on Technology, June, 2007.
portal.unesco.org/ci/en/ev.php-URL_ID=24700&URL_DO=DO_TOPIC&URL_SECTION=201.html

Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A-1), ISO 19005-1:2005
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920

FFmpeg
ffmpeg.org/

Gilmour, Ian and R. Justin Dávila. Lossless Video Compression for Archives: Motion JPEG2k and Other Options. Media Matters, January 2006.
www.media-matters.net/docs/WhitePapers/WPMJ2k.pdf

Guidelines for Computer File Types, Interchange Formats and Information Standards. Library and Archives Canada, 2004.
www.collectionscanada.gc.ca/government/products-services/007002-3017-e.html

ImageMagick software
www.imagemagick.org/

Information technology - JPEG 2000 image coding system: Motion JPEG 2000, ISO/IED 15444-3:2007
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41570

Information technology - Open Document Format for Office Applications (OpenDocument) v1.0, ISO/IEC 26300:2006
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43485

Material Exchange Format (XMF), SMPTE 377-1-2009
en.wikipedia.org/wiki/MXF

MPEG-2, Generic coding of moving pictures and associated audio information, ISO/IEC 13818 (9 parts)
mpeg.chiariglione.org/standards/mpeg-2/mpeg-2.htm

MPEG-2 Video Encoding (H.262). In: Sustainability of Digital Formats: Planning for Library of Congress Collections.
www.digitalpreservation.gov/formats/fdd/fdd000028.shtml.

Space data and information transfer systems - Open archival information system - Reference model, ISO 14721:2003
www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683

Xena software
xena.sourceforge.net/