The "M" Word: What Works and What Doesn't in File Format Migration

Lisa Gregory
Digital Information Management Program
State Library of North Carolina

Abstract: As the official permanent depository for all North Carolina state publications, the State Library of North Carolina is concerned with preservation and access of these materials, regardless of format. This paper describes our efforts at file format migration. With a limited budget and programming resources, we investigated appropriate migration file formats that match our current and projected needs, as well as open source tools that would normalize and document that migration. We found that, although far from perfect, there are tools that can effectively migrate a number of prevalent formats on a case-by-case basis. Work still needs to be accomplished to scale migration up to production level.

Introduction

> *"It can be argued that unless an object is accessible, it cannot be said to be preserved, as an inaccessible chunk of zeroes and ones is of no use whatsoever. Thus, any talk of preserving digital objects must include ways to access the objects"* (*Clausen 2004, 3*).

The State Library of North Carolina (SLNC) is the official permanent depository for all North Carolina state publications. We currently use a number of third-party tools to manage short-term access to our digital collections and are working on refining our long-term access plan. The most recent installment in this plan is an investigation of file format options for migration as well as available tools that would match our workflow and our resources. Similar to the reasons mentioned in Lawrence et al. (2000), we decided on migration rather than emulation (the other often-proposed long-term file access option) because of our environment and the programming expertise at our disposal.

Before embarking on this process, the "m" word invoked a fair bit of trepidation – not quite an expletive, but one that caused a bit of a cringe. But between our legal mandate and our passion for maintaining access to government information, we can't relegate the process of digital preservation to polite company forever. With a very limited budget, we set out to determine whether or not open source tools are a viable option for migrating data as part of a program for long-term access and, if so, which ones function best for the needs of libraries with similar resources. This paper recounts our first foray into migration testing, and it turned out a lot better than we expected.

Methodology

Investigating file formats and mapping transformations

At the SLNC, our current digital collections are predominantly textual in nature with a mixture of born-digital and digitized content. To prepare for migration testing, we started by considering the file formats currently in our repository. At this point in time, we get our files from several locations: (1) digitized in house, (2) digitized at a local partner institution, (3) retrieved from state agency web sites, and (4) submitted by agencies through email or on disc. We also have a

significant amount of web content stored with the Internet Archive's Archive-It through an ongoing web harvesting program.[1] We also hope that state agency donations will increase in the future as we make more contacts and publicize our services. With this increase we will need to be prepared to migrate a wider range of file formats.

For testing, we chose a range of file formats currently in our repository storage, as well as a number of formats we considered ourselves likely to receive in the future from state employees. After identification, we loosely grouped the formats according to type. The bulk of our current and anticipated formats fell into the "Images and Structured Graphics" and "Document-Like" categories. We also included a spreadsheet, audio/video files, and a web archive file in the test. In addition to a variety of files, we tried to select candidates that represented a range of file attributes. We chose one of our oldest files (2001), .pdfs with security settings and embedded content, a range of file sizes, multi-page .tifs, and converted files (such as .pdfs created from images or text files). We also manipulated the code in two of our .pdf files to determine if corrupted files could be migrated and if the corruption would be brought to our attention during any part of the process.

After we chose the file formats we were interested in migrating, we began to review current literature as well as other institutions' recommendations for file format transformations to determine the recommended output format for each type. Regardless of the content of a file, formats are generally preferred for migration when they are uncompressed, encoded in an open standard, widely used, and interoperable (Brown 2003). We fully intend to keep the original file formats for all files we are preserving; we hope our migrated files will keep our content accessible. Table 1 shows the desired transformations for our target files and the supporting references from which we drew our decisions, if any.

Out of all of the file formats we chose for investigation, we could not find a recommended migration format for Microsoft Publisher (.pub) and Adobe Photoshop (.psd) files. Though each could be migrated to a newer version using proprietary software, we could find no open format that accommodates the content, look, and interactivity of these files. We decided to migrate them to more static formats (.pdf and uncompressed .tif, respectively) using proprietary software as the only solution on hand which we could incorporate into our current workflow.

After deciding on each file's migration equivalent, we identified characteristics of the original file format and then determined which significant properties we hoped to retain after migration. These mapped closely to those described in Clausen (2004), which are listed below in parentheses. For each file format, we desired

1. No visual loss of content (readability);
2. No loss of metadata;
3. Minimal degradation in quality (appearance, "look & feel");
4. Minimal degradation in structure (comprehensibility);
5. Minimal degradation in interactivity (functionality).

Partly because our expectations varied depending on file type and partly because we did not know what a reasonable expectation of loss would be during this process, we deferred quantifying exact characteristics for each format.

Table 1. Migration Formats and Associated Tools

---

[1] See http://webarchives.ncdcr.gov for more information.

Original Format	Desired Migration Result	Selected Tools	Sources (see References)
Document-Like			
.css	.txt	XENA, PLANETS	
.doc (all versions)	.odt	XENA, PLANETS	2, 11
.docx	.odt	XENA, PLANETS	2, 11
.html	.txt	XENA, PLANETS	
.pdf	.pdf/a	XENA, PLANETS	2, 3, 15
.ppt	.odp	XENA, PLANETS	2, 11
.pub	.pdf*	*Acrobat*	
.rtf	.pdf/a	PLANETS, *Convert Doc*	8
.txt	.txt	XENA, PLANETS	2, 11, 15
Images and Structured Graphics			
.ai	.svg	Inkscape	2, 11
.psd	.tif (uncompressed)*	XENA, PLANETS	
.gif	.tif (uncompressed)	XENA, PLANETS	2, 3, 8, 11
.jpg	.tif (uncompressed)	XENA, PLANETS	2, 3, 8, 11, 15
.tif (compressed)	.tif (uncompressed)	XENA, PLANETS, *AVS Image Converter*	2, 3, 8, 11, 15
Audio/Video			
.mov	.mpeg-2 + mxf wrapper	FFmpeg	2, 3, 8, 11, 15, 17
.mp3	.wav file + bwf header	FFmpeg	1, 2, 3, 5, 8
Spreadsheets			
.xls	.odf	XENA, PLANETS	2, 11
Web Archives			
.arc	.warc	*	3, 18

* No recommended format or tool found.

NOTE: Tools in italics were used, but do not fit into the requirements for our final workflow (see the "Tools" section below).

Choosing Tools

The other part of our preparations dealt with discovering tools we wanted to test. Our aim was to test tools that could be used in practice within our institution and be incorporated fairly quickly into our workflow with our current minimal level of resources, expertise, and technology support. With this in mind, we established the following requirements. The tools would need to:

1. Be free,
2. Be open source,
3. Be relatively well documented,
4. Be currently maintained,
5. Provide an audit trail and, of course,
6. Perform the required transformation(s) successfully.

We hoped to find tools that met the above requirements but could also:

1. Be used easily (preferably with a GUI) and
2. Transform multiple formats.

After surveying a range of options, FFmpeg, Inkscape, the PLANETS Testbed, and XENA appeared to be the best options meeting our requirements for the range of file formats we were interested in testing.[2] A brief description of what we considered to be the most relevant characteristics of these tools (in addition to the criteria mentioned above) can be found in the Appendix.

We reviewed a number of other tools, some of which are rolled into the tools we used (like ImageMagick, JHOVE, DROID) but chose to test the tools which could accommodate the largest range of file formats, with a few exceptions to fill in the gaps. There were two additional tools originally in our testing plan. The first was PLATO[3] (Planets Preservation Planning Tool), described on the project website as "a decision support tool that implements a solid preservation planning process and integrates services for content characterisation, preservation action and automatic object comparison in a service-oriented architecture to provide maximum support for preservation planning endeavours." We were unsuccessful in connecting to PLATO, and decided not to pursue access for this project. The second was warc-tools,[4] which was the only option we could find that fit our criteria and could convert .arc files to the .warc format. At this point in time, warc-tools is still being developed and the code has not been released for public use.

Testing Procedure

Before we began migrating files, we checksummed those we had identified for testing and then copied them to a single folder on a local server. After they were copied, we verified the checksums to be sure they were intact. Using a computer in our on-site digital lab, we installed all of the tools mentioned above (with the exception of the web-based PLANETS Testbed). It should be noted that XENA also requires installation of OpenOffice.[5]

During our migration testing, each file was run through its associated tool(s). Any particular difficulty using the tools was noted. We used the viewers in the associated tools as well as the output from FITS to compare the retention of significant properties and retention of metadata as discussed above.

Results

Following are the results of our migration tests. All files with mapped migration transformations were tested at least once through at least one of the tools. Our tests were done over the course of two days.

Document-Like Files and Spreadsheets

We tried migrating almost all of our document-like files using both XENA and the PLANETS Testbed, with the exception of Microsoft Publisher files. As might be expected due to their lack of interactivity and relational data, .txt and similar files (.css and .html) rendered best during this process. Both content and structure remained intact, and XENA packaged them nicely with a metadata wrapper.

---

[2] For several of our files, we also experimented with FITS (File Information Tool Set). Created by the Harvard University Library, FITS "identifies, validates, and extracts technical metadata for various file formats" and then outputs the results in an XML file. FITS does not migrate files, but is useful for verifying metadata or identifying the type of file you have on hand. For more information, see http://code.google.com/p/fits/.
[3] http://www.ifs.tuwien.ac.at/dp/plato/intro.html
[4] http://code.google.com/p/warc-tools/
[5] http://www.openoffice.org

Although the bulk of our born-digital document-like files are in .pdf format, we also recognize the ubiquity of Microsoft formats and want to be prepared to migrate them if needed. In XENA, both Microsoft Word .doc and .docx files converted successfully to .odt format. When viewing them in the XENA viewer, the content and fonts appeared intact but the tables and tabs (.doc) and the bullets (.docx) did not render exactly as in the original. If exported from XENA and then opened in OpenOffice, however, the contents and formatting in both file types appeared to be correct. The metadata also transferred correctly. When converting a PowerPoint (.ppt) file to .odp using XENA, page numbers were missing but other content seemed to convert well. The .odp file had more formatting issues than the .odt files, including problems with fonts and footers. Although we were skeptical about migrating the Microsoft Excel (.xls) file, however it came through surprisingly well as an .odf file. Content, tabs, formulas and a text box appeared to be the same as the original. The only drawback was the loss of properties (author, manager, company, etc.) metadata, which were absent in the .odf file.

The .pdf/a file type presented more challenges than we expected. XENA does not normalize .pdf to .pdf/a, instead simply wraps the .pdf file in XML. Of our two manually corrupted files, neither rendered in the XENA viewer and we could only identify one as corrupted from the XENA report. While .pdf/a is an output option in the PLANETS Testbed, .pdf is not one of its corresponding input options. As far as we could discern, the Testbed also does not allow you to put in a file as undefined or unidentified. In the end we did not have success converting .pdf to .pdf/a using any of the open source tools in this experiment.

The two suggested migration formats for Rich Text Format (.rtf) are.odt and .pdf/a. We decided to attempt conversion to .pdf/a as an example case for using .pdf/a as a general migration format for textual documents. Our first tool was the PLANETS Testbed, but we found that .rtf is not available as an input format for .pdf/a. After searching for a free tool that would make the conversion we required, we found Convert Doc by SoftInterface, Inc.[6] We successfully converted the file to .pdf/a and the content and formatting appeared to be correct, but there were multiple errors when we checked the output file for compliance to pdf/a-1a and pdf/a-1b, including the failure of the metadata to convert successfully. Although we don't anticipate using .pdf/a as a migration format for .rtf, this was a useful demonstration of some of the issues that would have to be resolved should we ever wish to adopt this workflow.

The last document-like file format we hoped to convert using an open source tool was a Microsoft Publisher (.pub) file. As mentioned above, we could not identify a recommended open preservation format or a tool that fit our criteria for converting these files. Acrobat is capable of converting .pub files to .pdf, however as with the .rtf files we would not consider .pdf to be the best option because it diminishes the processability of the original data in the files.

In general, both XENA and the PLANETS Testbed performed well for document-like file types. In some cases, using the XENA viewer presented a different viewable result than exporting and viewing in OpenOffice. XENA inconsistently migrated the metadata (properties) from the original file format. We were not successful in migrating .pdf to .pdf/a using XENA or PLANETS, and Microsoft Publisher files presented the most obstacles for migration.

Images and Structured Graphics

---

[6] It should be noted that Convert Doc does *not* meet the criteria for tools described in our methodology. A free trial is available, which we used for this test.

As with the document-like file formats, this category was also dominated by XENA and the PLANETS Testbed. While using the Testbed, we chose the ImageMagick migration service option each time.

For .gif formats, XENA can convert to .png (with a less crisp result than the original) but we were interested in a .tif transformation for .gif files, which the PLANETS Testbed did successfully offer. The Testbed also successfully migrated a .tif file with LZW compression to an uncompressed .tif. When it came to migrating a .jpg to .tif, the migrated content was correct, however the "modified date" metadata in the file header of the original object was changed to indicate the date of the experiment. We didn't notice this happening with any other image formats, and additional .jpg to .tif experiments resulted in the same thing. On a production level, this change to the original object would not be acceptable.

The PLANETS Testbed accommodates transforming an Adobe Photoshop (.psd) file to a .tif file. The resulting file is excellent (perhaps even better in visual quality) but any layers or objects in the .psd file are not retained, leaving a limited range of options for reuse or examination of the original structure of the file.

Neither XENA nor the PLANETS Testbed could accommodate migrating .ai files, so we used Inkscape to perform this migration. The content of the .ai file remained intact, however the font formatting and coloring seemed mildly different in the resulting .svg.

XENA, the Testbed, and Inkscape all successfully converted our image files. The primary issues were the incorrect metadata for the .jpg to .tif conversion and the difficulty in preserving structured data (layers, etc.) in the structured graphics formats. The latter we expected, but the former we did not.

Audio and Video Files

For conversion of our QuickTime (.mov) and .mp3 files, we had success with FFmpeg. Initially, we were going to try converting .mov to .mj2 as a preferred migration format, but after having trouble locating a tool that fit our criteria to perform this conversion and a further review of the literature decided to see if .mpeg-2 with an .mxf wrapper was an acceptable alternative. FFmpeg did successfully perform the conversion, however there was significant degradation in the quality of both the sound and video. FFmpeg did successfully convert our .mp3 file to .wav, and the quality in the resultant .wav file was much more comparable to the original than in our video file conversion.

FFmpeg worked well in both instances, although we do not believe that .mpeg-2 is an acceptable migration destination for .mov files due to the poor quality of the result. In dealing with migration of audio and video files, we became very aware of our lack of expertise with these types of files. It hampered us in determining the best migration alternatives. We also concluded that the quality of audio and video files, because they include so much data layered in the format, should not be tested solely using human perception.

Web Archives

The final transformation we wished to test during this migration experiment was .arc to .warc. Unfortunately, we could not find any tool that met our criteria that could perform such a transformation. While the warc-tools project looks promising, the product is not yet available for testing. The Internet Archive stewards our .arc files for the time being, but we hope to see a tool

that can perform this service at some point in the future to allow us to experiment on some of our archived files and ensure increased control of access to our content.

General Tool Observations

In addition to the success of each file format migration, we came away with general impressions of the tools we used. XENA, though limited in some of the migrations, worked well. We appreciated the GUI interface, which is not a given with open source tools. Our primary difficulty was with the XENA viewer, which did not always render things in the same way as OpenOffice.

FFmpeg can be intimidating for multiple reasons. If you're not used to working at the command line, it takes a bit to get up to speed. In addition, between compressions, frame rates, and codecs, repositories not predominantly dealing with audio visual material and who lack that expertise will find a definite learning curve. Lacinak ([2010](#)) is highly recommended as a get-started resource.

As for the Testbed, the extended documentation functions are detailed and flexible and offer the ability to comment and report often during the process. The user interface was straightforward, but we felt that requiring users to go through each step of the "experiment" might be overwhelming to someone interested in a very specific function. We found the structure of the FTP area (used for processing batches of files) somewhat confusing. During the experiment, we wondered what a user would do if he or she had not validated the input file format or if the exact format version was unknown. It seems like it would be easy to incorporate FITS or one of its constituent tools as an extra step to suggest an input file format. Finally, the Testbed is precisely that – it does accommodate batch migration, although if it's a large batch it may be scheduled to avoid overloading their resources. However all of the files in the batch must be of the same file type, which we don't feel would scale to a real-life situation.

Limitations

We freely acknowledge that the migration testing described here has a number of limitations. Some of these were self-imposed – if we opened up the arena to proprietary and for-fee software, we would have more and, possibly, more robust options. In selecting file formats, we limited ourselves to objects in our repository or those created recently. None of our files were older than 2001, and yet it is likely that we could receive older legacy publications from a state agency at some point in the future. By confining ourselves to real files at hand, we didn't have access to "maxed-out" versions of each file format, such as a spreadsheet replete with macros and higher order functions or a presentation full of videos and audio clips. We only used what we had, although we did try to find files with as many interesting characteristics as possible.

As mentioned above, we also felt limited in our lack of knowledge regarding the general components and structure of video and audio formats. While we tried to educate ourselves, we still feel that more information would be helpful to truly ensure we were losing the fewest significant properties of those formats during migration.

Related to this idea, for all of these file formats, we only visually inspected the output to determine attribute retention. We examined metadata, compared images on the same monitor, and listened as closely as possible. But we acknowledge that there are more exacting ways to determine the difference between two similar files. In most cases, however, we felt that a visual analysis was enough, or that the difference between the original and migrated format was so

marked that it was sufficient to make a determination about whether or not we'd be happy with the migration.

Finally, as is apparent from our description above, this was a very manual process. In general, we walked files through each program one by one. For any repository with a significant amount of content, and especially one that isn't planning to migrate on the fly, procedures and programs that not only ensured the integrity of the original file formats but also could be performed in batch would need to be developed.

Conclusion and Next Steps

We found this file format migration process informative and (pun intended) refreshing. We did encounter a number of issues related to retention of metadata and loss of format, and there were a few file formats we were unable to successfully transform. On the whole, both XENA and the PLANETS Testbed were easy for us to use, ensured retention of more file contents than anticipated, and supplied us with plenty of audit trail information about our procedures. FFmpeg and Inkscape were also viable alternatives, albeit not as multi-purpose for our collections as the other tools we tested. Going forward, we hope to put these tools through a second round of testing to verify our results and solidify the procedure. We'd also like to broaden our test batch of files to include a broader range, perhaps including a sample of legacy items already held by the State Archives of North Carolina. In addition to keeping tabs on new tools, we will need to find processes that can scale to production before we can incorporate them into our workflow on a regular basis. Perhaps the most valuable result of this testing was our increased comfort level with migrating files in general. Eventually, if we're all saying it, it might not be so bad after all.

Appendix: Tools Used

FFmpeg
http://ffmpeg.org/index.html
FFmpeg can be used to transform, create, and stream audio and video files. Although FFmpeg does not have an institution behind its development, it has been widely used and incorporated into a variety of projects.
Creation Date: Unknown – at least since 2004
Documentation: Web page with intermediate and advanced materials (these do not introduce you to the concepts behind video and audio encoding, but do give you examples and specific commands)
Platform(s): Windows, UNIX, Mac
Interface: Command line

Inkscape
http://inkscape.org/
We decided to use Inkscape to accommodate converting .ai files to .svg files, something that other multi-purpose tools we selected could not handle.
Creation Date: 2003
Documentation: Wiki, FAQ, tutorials blog, and robust user community
Platform(s): Windows, UNIX, Mac OS X
Interface: GUI

PLANETS Testbed
http://testbed.planets-project.eu/testbed/
According to the PLANETS Testbed website, the Testbed "provides a dedicated research environment where services and data can be experimented upon, results can be evaluated and outcomes shared with the wider community." Released for broader use in 2010, the site provides robust features for documenting preservation experiments.
Creation Date: 2007 internally, 2010 for broad use
Documentation: User guide, some documentation on SourceForge
Platform(s): Independent
Interface: GUI

XENA (Xml Electronic Normalizing for Archives)
http://xena.sourceforge.net/
XENA is a tool specifically designed for digital preservation activities. It can detect file formats and then convert files into open formats for preservation. The National Archives of Australia has been developing this tool for quite awhile (when compared with other open source tools in this area). The last release was in December of 2009.
Creation Date: 2003
Documentation: Wiki with introductory and advanced materials
Platform(s): Windows, Linux, Mac OS X
Interface: GUI

References

1. Addis, Matthew, and Richard Wright. 2010. Audiovisual preservation strategies, data models and value-chains. PRESTOPrime project. https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.1_preservationstrategies_R0_v1.00.pdf (accessed August 19, 2010).
2. Archivematica. 2010. Media type preservation plans. http://www.archivematica.org/wiki/index.php?title=Media_type_preservation_plans (accessed August 19, 2010).
3. Arms, Caroline R. and Carl Fleischhauer. Format descriptions. Sustainability of Digital Formats Planning for Library of Congress Collections. http://www.digitalpreservation.gov/formats/fdd/descriptions.shtml (accessed August 19, 2010).
4. Brown, Adrian. 2003. Digital preservation guidance note: Selecting file formats for long term preservation. United Kingdom: The National Archives. http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf (accessed August 19, 2010).
5. CARLI. 2009. Digitization best practices for audio. Consortium of Academic and Research Libraries in Illinois. http://www.carli.illinois.edu/mem-prod/contentdm/guidelines_for_audio.pdf (accessed August 19, 2010).
6. Clausen, Lars. 2004. Handling file formats. Århus, Denmark: State and University Library / Copenhagen, Denmark: Royal Library. http://netarchive.dk/publikationer/FileFormats-2004.pdf (accessed August 19, 2010).
7. Fleischhauer, Carl. Format considerations in audio-visual preservation reformatting: Snapshots from the Federal Agencies Digitization Guidelines Initiative. *Information Standards Quarterly* 22, no. 2 (2010): 34-40. http://www.digitizationguidelines.gov/audio-visual/documents/IP_Fleischhauer_AudioVisual_Reformatting_isqv22no2.pdf (accessed August 19, 2010).
8. Harold B. Lee Library. Digital formats. http://www.lib.byu.edu/departs/dp/guidelines/formats.html
9. Lacinak, Chris. 2010. Primer on codecs for moving image and sound archives. New York: AudioVisual Preservation Solutions. http://www.avpreserve.com/wp-content/uploads/2010/04/AVPS_Codec_Primer.pdf (accessed August 19, 2010).
10. Lawrence, Gregory W., William R. Kehoe, Oya Y. Rieger, William H. Walters, and Anne R. Kenney. 2000. Risk management of digital information: A file format investigation. Washington, D.C.: CLIR. http://www.clir.org/pubs/reports/pub93/pub93.pdf. (accessed August 19, 2010).
11. Prom, Chris. 6. Develop preservation/access plans. Practical E-Records. http://e-records.chrisprom.com/?page_id=581 (accessed August 19, 2010).
12. ---. May 7, 2010. Planets Testbed review. Practical E-Records. http://e-records.chrisprom.com/?p=1183 (accessed September 21, 2010).
13. ---. April 25, 2010. PLATO (Digital Preservation Planning) software review. Practical E-Records. http://e-records.chrisprom.com/?p=1082 (accessed September 21, 2010).
14. ---. April 22, 2010. Review of XENA normalization software. Practical E-Records. http://e-records.chrisprom.com/?p=1081 (accessed September 21, 2010).
15. Rohde-Enslin, Stefan and Keith R. Allen. 2009. Nothing lasts forever. nestor / Germany: Institute for Museum Research. http://files.d-nb.de/nestor/ratgeber/ratg01_2_en.pdf (accessed August 19, 2010).

16. Rosenthal, David S. H., Thomas Lipkis, Thomas S. Robertson, and Seth Morabito. Transparent format migration of preserved web content. *D-Lib Magazine*, January 2005. http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html. (accessed September 21, 2010).
17. Schüller, Dietrich. 2009. Video archiving and the dilemma of data compression. *Internet Preservation News* 47:5-7. http://www.ifla.org/files/pac/IPN_47_web.pdf (accessed August 19, 2010).
18. Strodl, Stephan, Peter Paul Beran, and Andreas Rauber. 2009. Migrating content in WARC files. Paper presented at the 9th International Web Archiving Workshop in Corfu, Greece. http://publik.tuwien.ac.at/files/PubDat_181115.pdf (accessed August 19, 2010).
19. Van Wijk, Caroline. 2006. KB and migration test plan. National Library of the Netherlands. http://www.kb.nl/hrd/dd/dd_projecten/KB%20and%20Migration%20Test%20Plan.pdf (accessed September 21, 2010).