Protecting the Long-Term Viability of Digital Composite Objects through Format Migration

Elizabeth Roke
MARBL, Emory University
540 Asbury Circle
Atlanta, GA 30322-1006
+1 (404) 727-2345
elizabeth.roke@emory.edu

Dorothy Waugh
MARBL, Emory University
540 Asbury Circle
Atlanta, GA 30322-1006
+1 (404) 727-2471
dorothy.waugh@emory.edu

ABSTRACT

This poster documents work recently undertaken at Emory University's Manuscript, Archives, and Rare Book Library (MARBL) to review policy on disk image file formats used to capture and store digital content in our Fedora repository. Survey of the field and current best practices revealed waning support for the formats previously used and prompted collaborative efforts between Digital Archives staff and software engineers to migrate existing disk images to formats now deemed more suitable for long-term digital preservation.

General Terms

Preservation strategies and workflows

Keywords

Digital preservation; disk imaging; file format migration; PREMIS; digital archives; digital repositories.

1. THE PRESERVATION OF DISK IMAGES AT MARBL

The *Trusted Repositories Audit & Certification: Criteria and Checklist* requires that digital repositories monitor changes in technology that might impact preservation planning and maintain agile policy that can respond effectively to such changes [1]. This ongoing cycle of review and response is critical to the long-term preservation of digital objects and has been a key consideration in the development of policy at MARBL since functionality for the ingest of forensic and logical disk images was added in 2014 to our Fedora repository.

MARBL's collections include increasing numbers of digital media. A survey of the environment and best practices, conducted not long after the establishment of MARBL's Digital Archives unit, resulted in the decision to capture forensic disk images of this media using the open source Advanced Forensic Format (AFF), while logical disk images were captured using AccessData's AD1 file format. At that time, AFF offered a good solution for the capture of forensic disk images: unlike raw disk images, AFF files package disk image metadata with the image file. AFF's method of segmenting the disk image also made image compression possible [2]. That AFF is open source added to its appeal as a format for long-term preservation as it meant that we did not have to depend on limited proprietary formats, the viability of which often fluctuate in response to commercial markets. However, the development of Libewf by Joachim Metz, a library of tools supporting access to the proprietary Expert Witness Compression Formats, have decreased the need for an open source alternative like AFF. As a result, the creator of AFF, Simson Garfinkel, has stopped active development and no longer recommends AFF as a format for digital preservation [3]. In response to this shift in best practice, Digital Archives at MARBL recognized a need to update policy and workflow, which also presented a good opportunity to address the acquisition of logical disk images. Use of the AD1 file format to capture logical disk images had allowed us to generate and record fixity information as part of the imaging process. However, we were well aware that AD1's proprietary format left our data vulnerable, and we were keen to find an alternative better suited to our goals for long-term preservation.

Following conversation with colleagues across the field, we made the decision moving forward to acquire raw disk images or, where circumstances prevented complete forensic imaging, tar files. While this shift in policy did mean that we lost the benefits of the AFF format, we felt that data stored as raw disk image files was less vulnerable to obsolescence, as raw image formats are supported across platforms and their limited complexity results in a format that is, theoretically at least, more easily maintained over the long-term. Similarly, the ubiquity of tar, in addition to built-in functionality that preserves file metadata, presents a preferred alternative to the AD1 file format. While these changes are reflected in our current workflows, continued monitoring of the environment and best practices remains a key part of our policy. In response, we expect that our workflows will continue to develop and hope that they will continue to improve.

One of the challenges that resulted from this shift in workflow was how to migrate AFF and AD1 files already captured and stored in our Fedora repository.

2. THE MIGRATION PROCESS

This migration was Emory's first attempt to do file format migration in our Fedora digital preservation repository. It required us not only to develop methods for migrating the files but also to reconfigure the repository to accept new mimetypes for the newly ingested files.

The most straightforward objects to migrate were our AFF files. Software engineers performed the majority of the migration work. AFF, an open source format, has a library and toolkit that Emory used to automate the migration. Software developers obtained the AFF files from our digital repository, extracted the raw image from the AFF image, and uploaded the migrated disk image back into the repository. The fact that the AFF image itself includes the raw MD5 and SHA-1 checksums as part of the metadata enabled us to validate the migrated image upon conversion. Checksums were validated at each stage in this process to ensure the integrity of the digital object. A file containing these checksums has been stored with the object in the digital repository as a supplemental file. Once the migrated file was successfully ingested into the repository, the original AFF was deleted.

Migration of the AD1 images, a proprietary format, was a manual and much more complicated process. Fortunately, MARBL had captured only a limited number of these types of files, making this approach feasible. Software engineers obtained the AD1 files from the repository. Digital archivists loaded each AD1 file into Access Data's FTK Imager as an evidence item and extracted the files from the image. We also used FTK Imager to generate a file inventory with checksums for each file. Finally, the extracted files were packaged into a tar file using the Cygwin Console's tar utility with its built-in option to preserve file metadata.

Software engineers batch ingested the migrated files into the repository through a batch version of MARBL's normal ingest procedures. The tar files and the associated file inventories were packaged together using Python BagIt, which also generated fixity information. After the bags were ingested, the repository validated the digital object checksums and stored the file-level checksums as a supplemental file attached to the object. Once the migrated AD1 files were successfully ingested and validated, system administrators deleted the original AD1 images.

3. PRESERVATION METADATA AND SUPPORTING DOCUMENTATION

Since Emory began ingesting disk images into our preservation repository, we have relied upon the PREMIS metadata standard to encode the provenance of the original physical object. Technical metadata documenting the original environment (hardware and software) as well as forensic information about the imaging process are all recorded in PREMIS metadata. We also record events such as fixity checks. This structure is based in part on a model developed by the BitCurator project at the University of North Carolina that maps disk image metadata into PREMIS [4]. For the disk image file migration, we added a migration event to the object's PREMIS metadata that captured the details of the migration, including the software applications we used, migration dates, and other details.

Our use of PREMIS for disk images has also enabled us to capture file metadata no longer stored in the AFF or AD1 image. AFF and AD1 files natively package metadata about the original physical object and the imaging process within the disk image file. The raw disk image file format we now are using does not contain any of this valuable metadata. Instead, we are adding this information to PREMIS metadata for the object, ensuring that we are able to retain the metadata we need to preserve and access the object in the future.

4. IMPACTS

The migration of disk images from a proprietary or unsupported format to a raw file format has made it easier for us to manage and preserve these objects and mitigates the threat of obsolescence for the near term. The migration is not without long-term consequences, however. Although our extensive use of PREMIS preserves most of the metadata encoded in AD1 or AFF images, some system information captured as part of logical disk images has been lost as a result of the migration. We don't currently use any of this data, but it is a piece of forensic information about the object that we can no longer access. The deprecation of the AFF file format also means that we can no longer compress our disk images. This is not a concern now, but may be in the future as we continue to add large objects to our digital repository.

The greatest impact from migration has been on our imaging and processing workflows for composite objects. AFF and AD1 file formats, which automatically included system information and fixity information, guaranteed that that we preserved these types of objects in ways that were forensically sound. Going forward, we will be able to store the same metadata, but the process will be more complicated and require workflows that ensure we do so. Additionally, files will no longer contain any embedded metadata, meaning that we will consciously have to track that information along with the object.

The migration to a raw file format has made the digital file itself easier to preserve. The ongoing question is how easy it will be to preserve the original object it represents.

5. REFERENCES

[1] The Center for Research Libraries. 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist,* 31-32. http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf.

[2] Garfinkel, S., D. Malan, K. Dubec, C. Stevens, and C. Pham. Advanced forensic format: An open, extensible format for disk imaging. *Advances in Digital Forensics II,* M. Olivier and S. Shenoi, Eds. FIP International Conference on Digital Forensics (Orlando, FL, January 29-February 1, 2006). Springer, New York, NY, 17-31. http://dash.harvard.edu/bitstream/handle/1/2829932/Malan_AdvancedForensic.pdf?sequence=4.

[3] AFF format deprecated. January 15, 2014. Guymager wiki. http://sourceforge.net/p/guymager/wiki/AFF%20format%20deprecated/.

[4] Chassanoff, Alexandra, Kam Woods, and Christopher Lee. Mapping Digital Forensics Metadata to Preservation Events Using Bitcurator. SAA Research Forum (New Orleans, LA, August 13, 2013). http://files.archivists.org/pubs/proceedings/ResearchForum/2013/ChassanoffWoodsLee-ResearchForumPoster13.pdf