The Electronic Media Review

Electronic Media Group

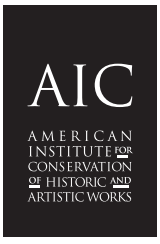
Volume Three 2015

Papers presented at the Electronic Media Group session of the 41st AIC Annual Meeting, Indianapolis, Indiana, 2013, and the 42nd AIC Annual Meeting, San Francisco, California, 2014.

Jeffery Warda and Briana Feston-Brunet, Managing Editors

Edited by Helen Bailey, Briana Feston-Brunet, Karen Pavelka, and Jeffrey Warda

Layout by Amber Hares
(Original design by Jon Rosenthal, JonRosenthalDesign.com)
Typeset in Trade Gothic LT and Myriad Pro



American Institute for Conservation of Historic and Artistic Works
Washington DC

The Electronic Media Review was published once every two years in print format by the Electronic Media Group (EMG), a specialty group of the American Institute for Conservation of Historic and Artistic Works (AIC), until 2013 and published online only thereafter. *The Electronic Media Review* is distributed as a benefit to members of EMG who held membership during the year of the issue. Additional copies or back issues are available from AIC. All correspondence concerning subscriptions, membership, back issues, and address changes should be addressed to:

American Institute for Conservation of Historic and Artistic Works
727 15th Street NW, Ste. 500
Washington, DC 20005
info@conservation-us.org
http://www.conservation-us.org

The Electronic Media Review is a non-juried publication. Papers presented at the EMG session of the AIC Annual Meeting are selected by committee based on abstracts. After presentation, authors have the opportunity to revise their papers before submitting them for publication in *The Electronic Media Review*. There is no further selection review of these papers. Independent submissions are published at the discretion of the EMG Publications Committee. Authors are responsible for the content and accuracy of their submissions and for the methods and materials they present. Publication in *The Electronic Media Review* does not constitute official statements or endorsement by the EMG or by the AIC.

Presented at the Electronic Media Group Session, AIC 42nd Annual Meeting, May 28–31, 2014, San Francisco, CA.

IMAGING DIGITAL MEDIA FOR PRESERVATION WITH LAMMP

MATTHEW MCKINLEY

ABSTRACT

Hardware/software obsolescence and bit-level corruption pose a serious threat to the ongoing accessibility of digital media storage devices, both magnetic (floppy and ZIP disks) and optical (CD/DVD, USB drives). The Legacy Archival Media Migration Platform (LAMMP) was developed to rescue potentially valuable digital content from rapidly aging digital media formats. LAMMP automates the generation of bit-for-bit disk images suitable for preservation as well as disk and file-level metadata for future appraisal and access. Beyond basic preservation, there is also archival value in reliably capturing the work environments of digital content creators. Migrating media to a digital image format and ingesting it into a digital repository ensures its continued authenticity for future investigation, emulation, and access. Built on open-source software, digital forensics tools, and a combination of modern and legacy hardware, LAMMP is low-cost and highly customizable. Along with disk imaging and metadata generation, LAMMP performs essential digital preservation workflow tasks such as virus checking, hash checking, and file extraction. Finally, automation via Linux command line tools and shell scripting allows for non-technical staff to perform the LAMMP preservation procedure for most common digital media formats.

INTRODUCTION

In late 2011, the Special Collections and Archives department (SCA) at The University of California, Irvine (UC Irvine) had a pressing problem without an obvious solution. Like many collecting institutions, SCA had for a number of years been acquiring hybrid collections containing both paper-based materials such as records, books and correspondence as well as digital media such as floppy disks, CDs and USB thumb drives. Whether used for transfer, versioning, or storage, these digital media objects often formed a crucial part of the collection creator's work environment and output; they were therefore just as worthy of appraisal, arrangement, description and eventual access as their more "traditional" counterparts. However, unlike paper-based materials, SCA had no established workflow for processing digital media, so disks and drives remained on the shelf with no more than a guess at the potentially valuable content that lay within.

Unfortunately, digital media storage is basically a ticking time bomb from a preservation perspective. One issue is access: interpreting legacy content on legacy media requires legacy hardware and software. Legacy hardware means acquiring and maintaining specialized tools that are often no longer being manufactured or supported by vendors. Assuming the media loads, there is no guarantee that its technical structure or the content within can be identified and read. Potentially valuable information may have been saved in an obscure and proprietary format by programs that fell out of use 20 years ago.

Beyond the access issues posed by hardware and software obsolescence, digital media carriers themselves have a tendency to degrade over a relatively short span of time. Such degradation can be caused by user error, material disintegration and damage, or "the silent corruption of data on disk or tape," a phenomenon known as "bit rot" (Salter 2014). To put this in perspective, well-preserved paper materials under normal use and storage conditions have the capacity to last several hundred years (ANSI/NISO 2009), while recordable and rewritable DVD media in certain conditions may become unreadable in less than 15 years (Zheng and Slattery 2007, 17). Magnetic mass storage devices such as USB thumb drives or even the hard drive within a computer are also at risk; when it comes to hardware failure, it is never a question of *if*, but *when.*

Thus, SCA needed a system for processing digital media objects that could address these technological and temporal concerns and produce a verifiably authentic version of the object while still adhering to archival concepts of acquisition, appraisal, arrangement, and description. Further, to ensure that digital media collections were processed in a timely manner, the system needed to be straightforward enough to be operated by a staff member who may have never even seen a computer command line. Finally, to have any hope of being implemented in the near future, this system needed to be budget-friendly, ideally costing less than $500.

The solution became the Legacy Archival Media Migration Platform (LAMMP), designed and developed within the UCI Libraries' IT department. LAMMP uses tools and techniques borrowed from the digital forensics community to rescue valuable content from digital media carriers. By capturing an exact binary copy of the digital media and extracting content files soon after the media collection is acquired, LAMMP lessens the risk of obsolescence and bit rot and increases the potential for accessing content stored within. Because it is built entirely on free or low-cost hardware and open-source software, LAMMP was inexpensive to build and is easily customizable. Best of all, using LAMMP is easy: just follow the manual, run a script, enter basic metadata and the system will do the rest.

USING DIGITAL FORENSICS IN LIBRARIES & ARCHIVES

Capturing and authenticating digital media content is extremely important in the legal and law enforcement communities. Hard drives, CDs, USB drives and more are routinely seized as evidence; and the content within must be extracted, tested, and held to strict standards in order to be admissible in the court of law. Much like

a traditional crime lab analyzes physical evidence, practitioners of digital forensics must capture and analyze a suspect's digital media in order to recover as much data as possible, then verify this data to serve as evidence. The digital preservation community centered around libraries, archives, and museums shares this concern for capturing and authenticating born-digital content, and has adapted several digital forensics concepts and tools for its own work.

The first step is capture. The easiest and most complete method for capturing digital media content is imaging, which involves generating an exact bit-for-bit copy of a disk or other piece of digital source media. A disk image can be thought of as a purely digital clone of the physical media carrier: the exact technical structure, metadata, and content of the media is stored in a single file which may then be analyzed and accessed by technicians or future researchers, without the risk of physical degradation or failure to read that comes with loading the original media carrier. Released by Access Data as a free but more limited alternative to their full Forensic Tool Kit (FTK) suite, FTK Imager is widely used software for generating and verifying disk images as well as extracting technical metadata (available at Access Data Product Downloads: http://accessdata.com/product-download). FTK Imager allows for exploration of disk image content and provides a wealth of technical metadata for in-depth investigation. The tool generates disk images in two of the more popular image formats, RAW (.001) and Advanced Forensic Format (AFF).

To verify both the fidelity of a disk image as an exact copy of a digital media carrier and its continued authenticity over time, a cryptographic hash function known as a checksum algorithm is used. Running this algorithm on a file or other discrete digital object produces a string of characters known as a checksum hash that uniquely represent the content and state of an object. If the object is changed in any way, whether due to human intervention or data corruption via bit rot, a different checksum will be produced by running the checksum algorithm. In this way, a checksum functions as a sort of digital fingerprint, verifying that digital content remains unchanged during imaging, transfer, and preservation. Different checksum algorithms use their own methods to generate different hash values; two checksum algorithms widely used in the digital preservation community are MD5 and SHA-1.

A write blocker is another tool used to ensure authenticity while accessing or imaging digital media. Its basic function is self-explanatory: to prevent the imaging workstation from compromising the original authenticity of the digital media by blocking the workstation from writing to the media while under investigation. A hardware write blocker is a physical device that connects via USB or other standard interface and blocks all write activity from the imaging workstation to the digital media. A software write blocker is a program that runs on the imaging workstation's operating system, monitoring and disabling all potential write activity to the digital media. While some argue that write-blocking via software is more flexible and efficient, hardware write blockers are generally seen as more reliable, owing to the basic nature of hardware (device) vs. software (system): "systems that are designed to write but rely on some type of control system to prevent a write can experience a failure of the controlling system…media protection devices are systems that are designed not to write and thus have no controlling system to fail" (Menz and Bress 2004, 6).

Combining these concepts, we arrive at the basic process for generating and verifying a purely digital surrogate of a source digital media object:

- Attach appropriate media drive via hardware write blocker to the imaging workstation
- Insert media into drive
- Generate checksum of source media
- Generate disk image from source media
- Generate checksum of disk image
- Compare checksum of disk image to checksum of source media to ensure exact match

The checksum of the disk image file can then be used to ensure the file remains unchanged after any digital forensic analysis or future access, and to verify future versions of the disk image file as authentic representations of the original digital media object.

INSPIRATION FOR LAMMP

The main source of inspiration for LAMMP was the *Frankenstein* machine, a purpose-built digital forensics workstation found within the Digital Archeology Lab at the University of Texas, Austin School of Information (UT iSchool).[1] While attending UT iSchool, the author used Frankenstein to image both 3.5" and 5.25" floppy disks as part of a course on preserving electronic records. The knowledge gained and connections made with Frankenstein's developers were then used to develop a similar digital forensics workstation at UC Irvine.

Beyond the general idea of constructing a low-cost and purpose-built digital forensics workstation instead of purchasing a more expensive commercial solution, the Frankenstein machine directly influenced LAMMP in its use of the Linux Ubuntu operating system and its output of basic metadata file formats.

In his seminar "Digital Forensics using Linux and Open Source Tools," forensics expert Bruce Nikkel outlines the many advantages of using Linux for digital forensics work, including ease of automation and scripting, an active support community, open and vendor-neutral standards, and a wide range of supported media and hardware, as well as free and open source software tools useful for forensic analysis such as dd, dd_rescue, sleuthkit, md5sum, and more (Nikkel 2005). Linux is also much less susceptible to any computer viruses that may be encountered while imaging and analyzing digital media. The Ubuntu Linux distribution was chosen for its graphic user interface, which is more immediately familiar to users accustomed to the graphic interfaces of Windows and Mac operating systems. Usability was also the deciding factor in choosing to generate simple text (.txt) and comma-separated value (.csv) files to store metadata. Widely used and small in size, both formats can be easily edited from many command line tools and can be accessed by nearly any word processing or spreadsheet software.

WALKTHROUGH OF LAMMP PROCESS

The high level LAMMP process involves four steps, each of which generates one or more image or metadata files (fig. 1).

First, a photograph is taken of the digital media object to capture label markings, branding, and other physical aspects. Next, the appropriate drive is attached to the LAMMP workstation, using a Tableau USB Forensic Bridge write blocker wherever possible. The remaining steps are performed by a script file named imageScript written for the Bash shell Linux Command Line Interface (CLI).

Before any imaging begins, imageScript connects to the campus Ethernet network to update the virus definitions of the open source ClamAV software used later in the LAMMP process for virus scanning. The script then disconnects from the network so that LAMMP is safely quarantined during imaging, preventing any viruses or other malicious content found on digital media from spreading to other machines on the network. After disconnecting, imageScript asks the user to enter their name, the digital collection number and digital media object number, and digital media object type (3.5" floppy, 5.25" floppy, data CD, DVD, etc.). The digital media type determines which tools are used to generate image and metadata files.

For 3.5"/5.25" floppy disks, the command line version of FTK Imager is used to generate a full-size raw disk image file. Ideally, the LAMMP process would generate a full-size disk image file for all media types, as it is the most authentic digital object for future access and investigation of the source media. However, media such as data CD/DVDs can often hold several hundred megabytes of data, and many USB drives now come with a capacity of 16 gigabytes or more. It was decided early on in the development of LAMMP that UC Irvine Libraries

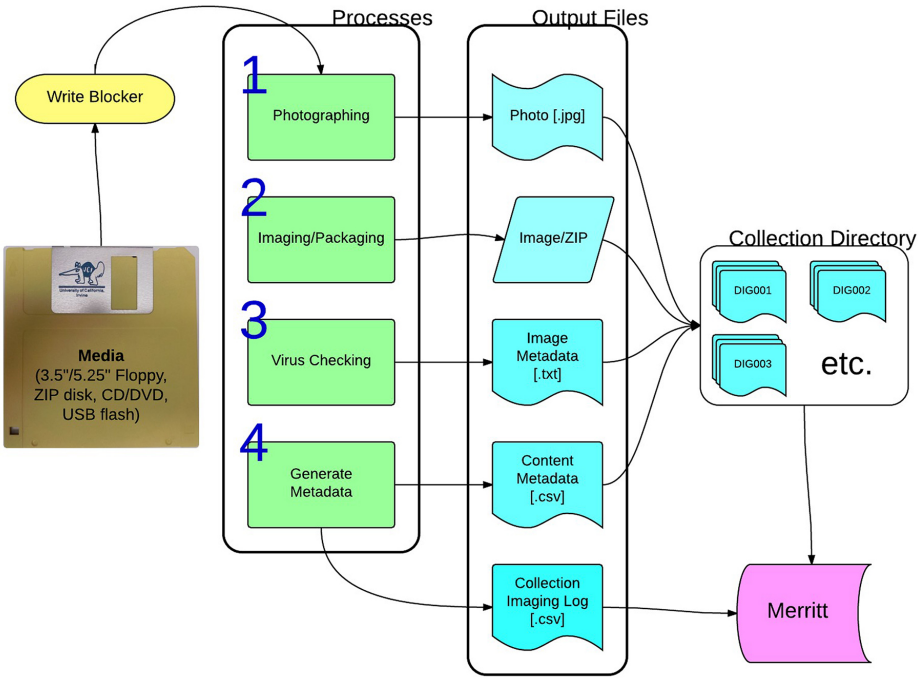


Fig. 1. A high-level overview of the LAMMP process.

did not wish to allot storage for exact images of these larger media objects, especially since many of the media contained allocated content comprising less than 50% of total capacity. Thus, for all digital media objects with a capacity of more than 100 megabytes, a .zip package containing only allocated content is generated in lieu of an exact disk image.

After imaging, the image file or .zip package is loaded in a forensically sound read-only way by passing specific parameters to the Linux *mount* command. This mounted directory is then scanned by the previously mentioned FTK Imager for viruses or other infected content. If infected content is found, the files are removed and their names noted in the disk metadata file below. Another virus scan runs on a digital media object's boot sector to look for infected data there. If no infected content is found, the "clean" result is printed to the disk metadata file.

Next, imageScript generates two metadata files. The first is a disk image metadata text file containing:

- Basic metadata of agent name, collection number, and object number entered above
- Technical metadata on the size and logical mapping of the disk
- Technical metadata generated by the Linux *disktype* command
- MD5 checksums of both the source media and resulting image to ensure a match

The second is a content metadata .csv file with a single line of metadata for each file found on the media including:

- Filename with extension
- Full file directory path
- File size in kilobytes
- Date created, if possible (this information is not saved on the Mac or Linux OS)
- Date last modified
- MD5 checksum of file

Finally, a line containing the following information for each media is added to an overall collection imaging log in .csv format:

- Image/.zip filename
- Date imaged
- Imaging agent
- Earliest date last modified of content found on media
- Image successful?
- Virus found?
- Notes

After examining the disk metadata file, the user records in the collection imaging log whether the image was successfully generated, whether infected content was found and removed, and any further notes on the media or imaging process. Unsuccessful images are later revisited to attempt a successful read via manual forensic investigation.

These steps are repeated for each media item in the collection. To link each image or .zip file with its associated metadata, all files are labeled according to a naming convention derived from the collection and object number. When all media for a collection have been processed, the files are grouped in a collection directory and ingested into Merritt, a preservation service provided by the California Digital Library (UC3 Merritt 2009).

The hardware and software required for accessing and imaging depends entirely upon the type of media being processed. PC-formatted 3.5" floppy disks are accessed via a USB floppy drive and imaged with FTK Imager. Although the Tableau USB Write Blocker is not compatible with the USB floppy drive, 3.5" floppy disks can be easily write-blocked by setting a small tab on the disk itself from read/write to read-only.

Accessing Mac-formatted 3.5" floppy disks is more complicated. This is because older Macintosh computers used a variable-speed system to write to floppies, which allowed more data to fit on each disk but results in a physically distinct magnetic encoding on the disk that cannot be interpreted in a PC-compatible floppy drive. Because of this, the LAMMP process uses a Powermac G3, the latest Mac model with an included Apple SuperDrive, to access these Mac-formatted 3.5" floppy disks and generate an .img disk image file. The .img files are then transferred to the main LAMMP workstation via a closed Ethernet network and converted to the raw disk image format with FTK Imager.

5 ¼" floppy disks are accessed via a TEAC floppy disk drive connected to LAMMP via the FC5025, a read-only USB floppy disk controller developed specifically for disk imaging and preservation.[2] Though the LAMMP process uses a software tool included with FC5025 to read and image the disk, the raw command line output is still piped to FTK Imager to generate a raw disk image.

CD/DVD data discs that contain logical directories and content (i.e., not primarily audiovisual) are processed in the same way as USB drives. For audio CDs, the open-source CDParanoia tool is used to convert audio tracks to .wav files, which are then stored in a .zip package. For audiovisual DVDs, another open source tool called IMGBurn is used to generate an ISO image of the entire disc.

ENHANCEMENTS

Since 2011, a number of enhancements have been made to the LAMMP environment to improve reliability and allow for appraisal and eventual access of disk image contents. First, the Ubuntu Builder tool was used to create a customized Ubuntu OS environment that mirrors LAMMP's hardware, software and customization. This customized OS was then stored in an .ISO package and loaded onto a USB thumb drive to create a portable, bootable version of the LAMMP environment. This USB drive can then be used to re-create LAMMP if the original workstation gets permanently corrupted due to a virus or hardware/software failure. The drive could also be used to easily clone the LAMMP environment to another workstation set up with the proper hardware.

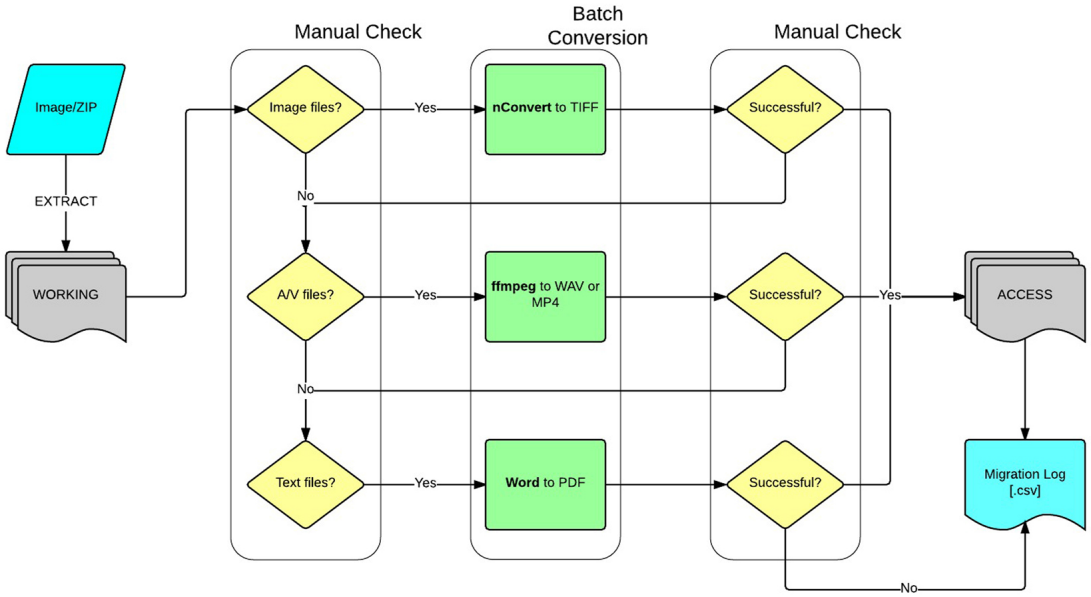


Fig. 2. LAMMP content migration process.

A virtual appraisal environment was also developed based on the open source BitCurator suite of digital forensics tools (http://wiki.bitcurator.net/index.php?title=Main_Page). BitCurator is a customized Ubuntu OS environment that comes packaged with tools for imaging, analyzing, searching and extracting metadata from disk images. The Basilisk II emulator (http://basilisk.cebix.net/) was also installed to enable viewing of legacy Mac images in their original environment, as many early Mac files cannot be interpreted in modern PC or Linux environments. Using these tools, the archivist may identify and appraise the value of the content found within each disk image.

After appraisal, disk images identified for further processing and eventual access are transferred to the content migration environment. This is a Windows OS environment with a series of Windows CLI scripts that, like the LAMMP process, have been designed to automate the extraction of disk image content and its conversion to modern preservation formats. Figure 2 gives an overview of this process.

After running the script to extract each disk image's content to a WORKING directory, a user examines this directory to identify file types. If image files are found, a batch conversion process is run using a file-type specific open source conversion tool and output files are saved in an ACCESS directory. The user then checks each file to ensure the conversion resulted in legible content. If so, the files remain in the ACCESS directory and are recorded in a .csv migration log. If not, files are removed from ACCESS and remain in the WORKING directory for further conversion. A similar process takes place for any audio, video, or text files found within the WORKING directory. Any files not converted at the end of the process are recorded as such in the migration log for further investigation. Although there are many old and esoteric file formats that cannot be converted via this batch approach, the process can successfully automate the conversion of most well-known file formats.

CONCLUSION

The initial collection processed by LAMMP contained 437 digital media objects, 396 of which were accessed and imaged, resulting in a success rate of 92%. A documented workflow has been implemented to deliver, track, and preserve SCA digital media objects via LAMMP. LAMMP itself is built mainly from donated hardware and open-source software—the only necessary purchases were the FC5025 5 ¼" floppy controller, the TEAC 5 ¼"

floppy drive, and the Tableau USB Write Blocker, which cost less than $500 total. By all accounts, UC Irvine Libraries succeeded in building a system to access and preserve legacy digital media as well as integrating archival concepts of acquisition, appraisal, arrangement, and description into the process.

That said, an important byproduct of LAMMP development was the learning process involved. The SCA and IT staff members who worked on the project now know much more about the realm of digital forensics and its application to the world of collecting institutions. In addition, identifying the weak points of this solution will help in asking the right questions about its eventual replacement. There are pros and cons of developing a home-grown technological solution using open source software instead of purchasing a commercial off-the-shelf solution. The gains in monetary savings and freedom to customize with open-source software come with a much heavier investment of time. However, the time spent installing, coding, and debugging develops local expertise and results in a much better understanding of the final tool, as well as ways of fixing and improving upon it.

The most important lesson: start dealing with digital media now! As stated previously, this media has a limited shelf life, and an automated solution such as LAMMP isn't necessary to perform meaningful digital forensics work. Potentially important content can be imaged with just a USB floppy drive and a free copy of FTK Imager, and active projects such as BitCurator are consistently finding new ways to bring the tools and methods of digital forensics to the world of collecting institutions. As relics of the recent past, digital media objects can form an important part of the research corpus of the future—but only if they are sought out, captured, and properly preserved.

NOTES

1. Information on the development and use of the Frankenstein machine (including manuals, tools and oral histories with its creators): http://pacer.ischool.utexas.edu/handle/2081/21808.
2. Information on the FC5025 can be found at Device Side Data – FC5025: http://www.deviceside.com/fc5025.html.

REFERENCES

Menz, M. and S. Bress. 2004. The Fallacy of Software Write Protection in Computer Forensics. www.mykeytech.com/SoftwareWriteBlocking2-4.pdf (accessed 08/03/14)

Nikkel, B. 2005. Digital Forensics using Linux and Open Source Tools. http://www.digitalforensics.ch/nikkel05b.pdf (accessed 08/10/14)

NISO. 2009. Permanence of Paper for Publications and Documents in Libraries and Archives, Z39.48-1992 (R2009). Baltimore: National Information Standards Organization.

Salter, J. 2014. Bitrot and atomic COWs: Inside "next-gen" filesystems. http://arstechnica.com/information-technology/2014/01/bitrot-and-atomic-cows-inside-next-gen-filesystems/ (accessed 07/31/14).

Sonic Purity. 2008. Working with Macintosh Floppy Disks in the New Millennium. http://siber-sonic.com/mac/newmillfloppy.html (accessed 08/23/14).

UC3 Merritt. 2009. University of California Curation Center, California Digital Library. https://merritt.cdlib.org/ (accessed 08/23/2014).

Zheng, J. and O. Slattery. 2007. NIST/Library of Congress Optical Disc Longevity Study: Final Report. www.loc.gov/preservation/resources/rt/NIST_LC_OpticalDiscLongevity.pdf (accessed 08/01/14).

Matthew McKinley
Digital Project Specialist
University of California Irvine
matthewjamesmckinley@gmail.com