Archival Science, Digital Forensics and New Media Art

By

Dianne Dietrich and Frank Adelstein

From the proceedings of

The Digital Forensic Research Conference

DFRWS 2015 USA

Philadelphia, PA (Aug 9th - 13th)

Digital Investigation 14 (2015) S137–S145

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS 2015 US

Archival science, digital forensics, and new media art

Dianne Dietrich [a], *, Frank Adelstein [b]

[a] Cornell University Library, Ithaca, NY, USA
[b] Cayuga Networks, Ithaca, NY, USA

Keywords:
Archives
Archival forensics
New media art
Complex born-digital material
Case study
Intent vs. integrity

ABSTRACT

Digital archivists and traditional digital forensics practitioners have significant points of convergence as well as notable differences between their work. This paper provides an overview of how digital archivists use digital forensics tools and techniques to approach their work, comparing and contrasting archival with traditional computer forensics. Archives encounter a wide range of digital materials. This paper details a specific example within archival forensics—the analysis of complex, interactive, new media digital artworks. From this, the paper concludes with considerations for future directions and recommendations to the traditional forensics community to support the needs of cultural heritage institutions.

Introduction

Digital forensic analysts conduct digital investigations using various tools and techniques following the principles of Forensic Science. Digital archivists also use many of the same tools and techniques to conduct digital investigations as part of archival activities following the principles of Archival Science. A large overlap exists between these two fields. Both seek to understand the intent behind the artifacts they find, although the interpretations of intent as well as interactions with properties such as bitwise fidelity can be very different. This paper compares the commonalities and differences between archival and traditional forensics approaches to handling digital material, and considers these in light of a case study focusing on analysis of new media digital artworks.

The paper is organized as follows. The next section, Archival science, describes the essential principles of archival science, its goals, and the tools and technology used by digital archivists and where these converge and diverge with digital forensics. Following that, we present a case study from the analysis of a collection of digital New Media Digital Art from the mid 1990s to early 2000s, focusing on the analysis of three specific works, highlighting the challenges these works presented. The final section concludes the paper with a discussion of recommendations for tool developers and potential future work.

Archival science

The phrase "digital forensics" invokes an image of law enforcement officers conducting criminal investigations. The breadth of digital forensics practices goes far beyond this narrow definition. Civil cases use forensic analysis. Large corporations and organizations use their own forensics groups to investigate internal issues, compliance, and insider threats that are rarely publicly released. Governments have forensic resources that are applied in many areas, such as military intelligence.

In addition, a well-established area of forensic investigation that is rarely considered or mentioned by other forensics groups involves the use of digital forensics practices by digital archivists. There is a significant overlap between the goals and approaches of digital archivists and traditional forensics practitioners; further, archivists working

* Corresponding author.
E-mail addresses: dd388@cornell.edu (D. Dietrich), frank@notfrank.com (F. Adelstein).

http://dx.doi.org/10.1016/j.diin.2015.05.004

with digital materials often use utilities developed from traditional forensics fields (Kirschenbaum et al., 2010). (In this paper, we will use the term "traditional forensics" to denote non-archival forensics.) In this section, we introduce archival science, and then compare and contrast it to traditional forensics groups, considering high-level goals and objectives, as well as lower-level use of specific forensics technologies and techniques.

Archival science and archivists

In order to understand the work that digital archivists do, one must understand the framework that underpins their work—that is, the goals and aims of the archival profession as a whole. The Society of American Archivists defines archival science as a "systematic body of theory that supports the practice of appraising, *acquiring*, *authenticating*, *preserving*, *and providing access* [emphasis added] to recorded materials" (Pearce-Moses, 2005). This has many similarities to McKemmish's definition of *forensic computing* as the "process of identifying, preserving, analyzing and presenting digital evidence" (McKemmish, 1999). The above definition of archival science serves to support the creation and curation of archives. Archives generally contain primary source documentary materials, or records, that have been "preserved because of the enduring value contained in the information they contain or as evidence of the functions and responsibilities of their creator (Pearce-Moses, 2005)." Types of archives range widely and include university archives, government archives, corporate archives, and others. Not all archives house records only: some archives also collect rare materials (e.g., first editions of important novels or political ephemera) that are of interest to the institution or its user community. In general, though, archival practice draws from the core principles of archival science.

Archival science goals and objectives

Archivists provide access to trustworthy records, irrespective of their original format. Trustworthiness depends on a number of factors, including reliability and authenticity. In considering how archivists draw from forensic practice to approach handling digital material, we highlight two key characteristics of archival materials, as identified by the International Council on Archives.

- Records must have *integrity*, meaning they are complete and free from corruption. And,
- Records must be *usable*, stored in a way that allows others to retrieve, examine, and analyze them.[1]

Ensuring the *integrity* of digital materials means that archivists must have the appropriate tools and policies to prove that digital material has not been corrupted or inadvertently altered, either through decay or transfer to other storage environments or repositories.

Like all materials, the physical media containing the digital material is subject to decay. For example, manufacturers of so-called archival CD-Rs purport that this media can last up to 100 years, but the true lifespan of the media can be dependent on a variety of factors (Iraci, 2005) and research on optical media longevity is still ongoing (Library of Congress and National Institute of Standards and Technology, 2007). Unlike physical material, exact copies of digital materials can be produced (e.g., backups of files). Unless archivists take care when copying digital material, this process has the potential to introduce subtle changes that might go undetected, such as altering metadata (e.g., timestamps) or altering the data itself (e.g., inadvertently copying a file into a lossy format or failing to copy both forks of a file on an HFS file system). Archivists often try to avoid actions that change the material in any way, but if this is not possible (e.g., a degrading VHS tape needs to be digitized, or a rare book needs to be rebound), it is important to fully document what conservation actions were done in case these changes have implications for future users of the material.

In order to properly manage digital materials, archivists must define metadata that sufficiently describes the creation and context of complex digital material and the digital material itself. Long-term preservation ensures the ongoing accessibility and usability of records by users. In the following sections, we describe how archivists maintain record integrity and accessibility, highlighting where these activities and goals parallel those of and diverge from those of digital forensic investigators.

Ensuring integrity of materials

Archivists need to ensure that digital material has integrity, meaning it has not been inadvertently altered or changed in any way from acquisition through preservation actions, including transfer to and from storage environments and repositories. The following describes how archivists ensure material integrity at various stages in processing, with comparisons to similar activities in traditional forensics.

Integrity is closely related to, though not the same as, the archival concept of authenticity: the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) project defines an authentic record as "a record that is what it purports to be and is free from tampering or corruption" (MacNeil et al., 2001). The topic of authenticating data—for example, verifying an email has been sent by the person identified in the header—is out of scope for this paper. It was not needed in the work described in our examples because the artworks were either provided by the original artists or purchased from vendors who supplied credible provenance information.

Ensuring that records have not been inadvertently altered or corrupted begins with *accessioning* (Pearce-Moses, 2005), the process by which the archives assumes control and responsibility for materials, and acquisition, and continues through all subsequent processing steps. Archivists keep records regarding the details of the acquisition process. During acquisition, as well as afterwards, archivists must ensure that no inadvertent changes have been made to digital material or its respective metadata.

[1] http://www.ica.org/125/about-records-archives-and-the-profession/discover-archives-and-our-profession.html

Best practice suggests using physical write-blockers as standard practice for transferring material from one storage media to another, in order to prevent changing the original media, and storing hashes for digital materials (Lee et al., 2013; Erway, 2012).

In traditional forensics, maintaining data integrity is essential. The process begins on the scene. Data can be physically taken to the lab or imaged on-site. In either case, investigators gather metadata, such as time, location, device properties (e.g., disk type, capacity, etc.) and who performed the actions. They may also take pictures of the physical installation, wiring, power cords, connections, and other aspects of the scene. Data is placed in a tamper evident bag and taken to a forensics lab (Casey, 2000).

When the data is imaged, typically the imaging software produces one or more cryptographically secure hashes. These hashes are then recorded and stored along with the data. Investigators use hashes to support the argument that the data has not changed from the time it was imaged or acquired. The evidence is stored in a locked, secured location, and investigators maintain a record of every time the evidence is removed or replaced in the storage facility. While out, it is under the direct supervision of whoever signed it out, maintaining the Chain of Custody (Brezinski and Killalea, 2002) of evidence and copies. Traditional forensic investigators use standardized policies on acquiring, handling, and analyzing evidence to preserve integrity. This parallels the work of archivists.

Within archives, it is crucial that archivists can verify that files have not been inadvertently altered or corrupted in any way, especially since digital material may be transferred to and from multiple systems. An archivist may transfer digital material from fragile or obsolete hardware to more stable storage; digital material may be stored in a repository for long-term preservation; digital material may also be transferred to a system specifically designated for user access, such as a kiosk in a museum or a room with dedicated computer terminals, if networked access cannot be provided. Archivists verify that files have not been corrupted by calculating and storing hashes. This ensures that the integrity of the digital material has been maintained. Hashes also allow users to confirm they are working with an exact copy of the material the archives has supplied for use.

Archivists are more likely than traditional forensics professionals to work with older digital materials and may have to handle file formats that are no longer in use or readable using current software. In this case, they might need to convert files into a different format in order to determine their content or allow users to access the material. Given the importance of ensuring the integrity and usability of a record, archivists are often concerned with ensuring that the “significant properties” of digital material have been preserved (Grace et al., 2009), though determining what makes an altered record fundamentally the same as the original is not trivial (Yeo, 2010). For example, in some cases the layout and format of a text document may be critical to understanding its function and meaning as a record; in other cases, the text itself may be the only critical component of a file and may be formatted for reading in any way, with no significant loss of meaning. Original files and media may still be kept, depending on the archives' policies.

By contrast, in traditional forensics, once a copy or image has been created of the original media, the original media is generally never used again. In fact, investigators will make working copies from the copy, each time using the hash(es) to verify that the result is an exact bit-for-bit copy of the source.

Data migration is common in traditional forensics, but more as a functional necessity. Investigators work on a second-generation copy of the evidence, sometimes using their copy on a dedicated forensic workstation and sometimes using it from within a dedicated virtual machine. For evidence not originating from a file on disk, such as a memory dump, a process list, a list of active network connections, or other live data, investigators must migrate it from the native form, such as an in-memory OS data structure, into a file (Adelstein, 2006).

The migration, however, is performed as a matter of operational necessity, in order to import the data into a system for analysis. Once a trial has been completed, the evidence is generally of less importance. Because of the large case backlog, limited disk space, and the expense and workload of maintenance, case information is not stored online indefinitely. Instead, the old data is stored offline, as a box of tapes, DVDs, or disk drives and is rarely, if ever, used again. Most criminal forensic organizations have no long-term data preservation and maintenance policy beyond physical storage.

Also, if data is copied and the hashes do *not* match, investigators have limited options. The most likely outcome is that the investigator will examine the damaged evidence and attempt to argue that the evidence should be admitted because the damage does not impact the claims supported by the evidence, and that other evidence corroborates these claims.

Ensuring records are usable, accessible, and preserved

Providing access to users is a core function of archives. The specifics often vary, depending on factors such as institutional policy or donor agreements, and can range from on-site access, such as designated reading rooms where users must remain and register in order to work with archival material, to online access to digitized and “born-digital” material. In contrast, traditional forensics generally does not provide public access to forensic material, such as murder weapons or intelligence data.

Digital material can pose additional challenges to archivists who need to provide access to users. Archivists may need to redact sensitive or confidential information (e.g., phone numbers, email addresses, etc.) from a large corpus of digital material. The archivist may not immediately know the nature of the digital material collected at the time of acquisition, and analyzing hard drives for potentially sensitive or confidential material may be a complex task. Further, a donor agreement may specify that the archives can accession a complete disk image, but users may only access copies of select files, and the archives must ensure that the technical infrastructure is in place to handle user requests in a way that complies with donor agreements.

Archives provide descriptive information for materials to provide context for users. This is generally in the form of *finding aids* (Pearce-Moses, 2005), which, broadly speaking, include any documentation that facilitates the use and understanding of materials and helps users locate specific information within records. There are widely used standards for structuring this information about archival materials (e.g., Encoded Archival Description[2]). Traditional forensics investigators create reports and departments maintain cross-links for materials, such as case numbers, but these are typically for internal use; documentation is not intended to provide context for outside users.

While traditional forensic analysts may be using digital materials to support a claim, such as a suspect's involvement in criminal activity, archivists may try to avoid making any assumptions about a user's potential research question. When practical, an archivist may try to avoid imposing an order and minimize their interpretation of archival material, since it is impossible to predict how others may make use of it. For example, one user may be interested in the content of document files found on a famous scientist's hard drive, while another user may be interested in the history and progression of the various file formats found on that same hard drive.

Given this, the description for digital material needs to be structured to preserve as much information about its original state as possible. (This is tied to the archival concept of "original order" (Pearce-Moses, 2005).) There is some traction in the use of Digital Forensics XML (DFXML) (Garfinkel, 2012) to supplement technical metadata for digital materials (Lee et al., 2013) since it captures metadata about the structure and layout of digital media.

Records chosen for inclusion in archives are often chosen because they are of "enduring value" (Pearce-Moses, 2005). Thus, the act of preserving material for the long-term is a key function for archives. Without proper preservation, archival material is inaccessible for users. There is a wide range of ongoing maintenance that digital archivists perform to ensure that all archival material is properly preserved. Many of the same activities that support assurance of digital materials' integrity, authenticity, and usability also support their ongoing long-term preservation. *Fixity checks*—ensuring that files have not been corrupted at the bit level—are just one component of long-term digital preservation. Archival systems need to preserve associated administrative metadata as well.

Additionally, archives often encounter older digital material, on obsolete hardware and storage formats, and need to transfer data to newer storage platforms in order to preserve it. Here too, documentation is important; archivists are aware that there is "no preservation without loss"[3] and that preservation functions, like transferring data from one medium to another, converting to newer formats, or viewing files in emulation, all can effect change that needs to be recorded. The archival community has developed metadata standards for digital objects to support their preservation (i.e., PREMIS[4]).

In traditional forensics, however, the useful lifespan of data is closely tied to the case. Once that has been resolved, the likelihood that the data will be used drops very low. The data will be retained, but typically in an offline, unmaintained storage facility, with no regular fixity checks performed. In the event of an appeal, the investigator will attempt to recover the data from storage.

Archival science tools and techniques

The archival and digital preservation communities continue to develop tools and strategies to handle complex digital materials, such as the Duke Data Accessioner for migrating data off of disks,[5] various utilities for identifying, validating, and extracting metadata from files, such as FITS,[6] fido,[7] and DROID.[8] Of note here is BitCurator (Lee et al., 2013; Lee and Woods, 2014), an environment that adapts computer forensics utilities to meet the needs of those working in archives, libraries, and museums, staying mindful of those who may not be experts in computer forensics techniques. This environment includes multiple tools for report generation; imaging and analyzing media, such as Guymager, dcfldd, cdrdao, libewf, afflib, and bulk_extractor; generating DFXML, including fiwalk; file system forensics using The Sleuth Kit; and other utilities for antivirus, reading Outlook PST files, and HFSViewer for older Macintosh-formatted material. Ongoing development is focused on creating an environment that facilitates access to digital materials. In addition to freely available tools, archivists do also draw from commercial software, including free tools, such as FTK Imager, and non-free tools, such as EnCase.

The focus of many of the tools archivists use is to understand the nature of, and properly describe, digital materials so that they can be preserved and others can access them. For older material that may need obsolete software to render properly—such as an older or proprietary format—a virtual machine or emulator is one strategy to provide such access.

Traditional forensic analysts often use VMs because it is easy to create new systems that are in a known, clean state, and have a standard set of tools installed. In addition, sometimes key evidence that contained some item of interest, such as an email address or URL, is a data file for a relatively unknown program in an unknown format. VMs provide a repeatable, high fidelity execution environment that limits the risks of running unknown and possibly malicious code. Also, by restoring a VM's state to that of an earlier snapshot, a program can be repeatedly run to see how it uses data or how it attempted to erase data, and what artifacts it leaves behind. In more complicated cases when programs must be reverse engineered, VMs can serve

[2] http://www.loc.gov/ead

[3] http://www.slate.com/articles/arts/culturebox/2013/07/how_will_historians_of_the_future_run_ms_word_97_how_can_we_save_it_for.single.html

[4] http://www.loc.gov/standards/premis

[5] http://dataaccessioner.org/

[6] http://fitstool.org

[7] http://openpreservation.org/technology/products/fido/

[8] http://digitalpreservation.github.io/droid/

as an ideal platform for an execution and analysis environment.

Virtual machines and emulation were an essential part of the analysis and investigation of the artworks described in the following case study.

Case study: new media art

In this section, we present the highlights from processing a large collection of complex digital art. We first provide background information on the collection, then describe the overall approach used by the project team, and then present details from three works. A traditional forensic analogy to analyzing and archiving older digital artwork may be a case where investigators must re-open a previously-closed case in light of new evidence to find twenty year old digital data.

Background

The term "new media art" describes artwork created using so-called new media (i.e., a medium not previously used by artists at the time it was created), and includes "digital art, computer graphics, computer animation, visual art, Internet art, interactive art, [and] video games…".[9] Various archives and cultural heritage organizations have a stake in preserving and restoring this culturally significant material, which poses distinct challenges that differ from artwork in more traditional formats. Perhaps the most high-profile institution involved in the preservation and analysis of new media art is the Museum of Modern Art and its curation of a video game collection.[10] Another organization, Rhizome, helps fund artists working in new media art and hosts the ArtBase,[11] a collection of over two thousand new media artworks; the Transmediale CDROM Art Archive[12] includes a collection of several hundred new media artworks on optical media.

The Rose Goldsen Archive of New Media Art includes a collection of over 12,000 titles in video art, born-digital, complex interactive artworks on CD-ROM and DVD-ROM, Internet art, digital imagery, and research materials created from the mid 1960s to the present. This collection is currently housed within the archives of the Division of Rare and Manuscript Collections within Cornell University Library.

In 2012, Cornell University Library received a grant from the National Endowment for the Humanities (Casad, 2013) to develop a scalable preservation and access framework for a test bed of approximately three hundred artworks in the Rose Goldsen Archive of New Media Art. During that time, the project team has made extensive use of computer forensics tools to support the technical analysis of these artworks. The following section describes the overall approach of the project team in meeting its objectives to preserve this material and provision for its future access so that this rich history can be preserved for future scholars.

Overall approach

Since the works in the test bed were primarily housed on fragile media with a limited lifespan—including retail quality CD-Rs burned more than a decade ago—one of the primary tasks was to preserve the content of the discs by making an exact copy of the information contained on them by making disk images. Since the project team anticipated issues reading the discs (given their age), they wanted to note any errors during the imaging process in an automated way. They decided that if a disc was partially unreadable, they would capture the best scan that they could (i.e., using the drive and speed that produced the fewest unreadable sectors) and ensure that these processes, along with any notes about errors, were documented in scan logs. While there are numerous utilities for creating disc images, the project team mainly worked with the following.

Guymager[13] is a Linux utility that creates sector-by-sector copies of discs and produces an information file that includes a list of unreadable sectors, the hardware used to make the image, hashes for the source and image (for verification purposes) and other important administrative metadata.

IsoBuster,[14] which is Windows-based software, can read discs in raw format, which was especially important for analyzing mixed-mode discs (i.e., having both audio and data tracks concurrently). For further reading on working with mixed-mode CD-ROMs, see Brown (2012).

The artworks in the test bed collection were typically created for use on personal computers and consist of software, audiovisual files, and web files to create an interactive experience for the user. Since the project team determined that maintaining and supporting legacy hardware was not a reliable or sustainable strategy for providing access to this material, future access will rely on running the artwork on modern systems. While some operating systems do have some support for running legacy programs (i.e., Windows) it too is often not a reliable strategy for providing access for multiple reasons.

First, some works require third-party plugins or additional software to run, and the project team found it was not always possible to install these on modern browsers. Even in cases where installation was possible, they potentially conflicted with newer plugins. Second, the look and feel of operating systems and web browsers has changed dramatically over time, and running a work in a modern system is a different experience than interacting with it on a contemporary system setup.

The project team investigated emulation as a strategy for providing access to the artworks. It is far easier to meet the stated system requirements for an artwork through emulation. Emulation is not, however, a perfect solution: the process of running an artwork through an emulator can

[9] http://en.wikipedia.org/wiki/New_media_art

[10] http://www.moma.org/explore/inside_out/2012/11/29/videogames-14-in-the-collection-for-starters/

[11] http://rhizome.org/artbase

[12] http://bw-fla.unifreiburg.de/demo-transmediale.html

[13] http://guymager.sourceforge.net/

[14] http://www.isobuster.com/

introduce slight changes to the experience. Simply transferring the data from its original optical disc format to a disk image changes the overall physicality of a work; that is, a user no longer needs to load a physical disc into a drive on a computer to access it. Moreover, changes in the look and feel of peripheral hardware over time, such as keyboards and mice, can have an effect on a user's overall experience of a work (Hedstrom et al., 2006). For example, many artworks in the Goldsen collection place great emphasis on the physical, embodied experience of the user as he or she engages with an interactive interface. The material object of a computer mouse may be significant in such works as the thing or tool a user must manipulate in order to interact. This aspect of the user's experience may be altered in unintended and potentially detrimental ways when the work is viewed in emulation using a modern hardware setup: a trackpad may retain all the interactive functionality of a classic mouse, for example, but not the important quality of being a handheld object. By the same token, a mouse with a scroll wheel invites interactive gestures from the user that might not have been anticipated, or even possible, at the time of the artwork's creation. Such changes can significantly reshape the user's overall experience of an artwork.

Without knowing an artist's intent through direct conversation, or having detailed descriptions that can serve as reference points for evaluating the work, it can be difficult to know which emulation rendering infelicities can be tolerated and which negatively affect the work. One of the project team's strategies for dealing with this situation is thorough documentation of all apparent issues with running a work in emulation. For example, the color on a newer LCD monitor may not render a subtle red shade quite as well as a CRT monitor. Screen size, aspect ratio, and resolution are all somewhat different on modern LCD screens. Moreover, even on its slowest setting, a work might cycle through images far faster in emulation than it ever did on the original intended hardware. Whenever possible, the project team has documented strategies for ameliorating negative effects from emulation artifacts such as these.

Further, system requirements for the materials in the collection vary by artwork and can range anywhere from Windows 3.1 through XP and Macintosh System 7 through OS X. Many works were cross-compiled for Windows as well as Macintosh computers, and their documentation often referenced a diversity of system configurations that were capable of viewing the work. Again, without direct conversations or specific reference material, it can be difficult to identify the canonical standard experience to compare against when testing the work in various emulation environments.

It was also important for the project team to provide technical metadata for the artwork. This technical metadata needed to be thorough, yet not so information dense that future users or archivists would be overwhelmed by it. Building from the results of a user survey asking both artists and curators how they envisioned interacting with these materials, the project team determined what metadata was necessary for future users and archivists to successfully interact with and preserve the works. Emulation seemed like a viable access strategy, but nonetheless, it was especially important to provide descriptions in a general way. Strategies for access, such as emulation, and their supporting technologies are all likely to evolve over time. What emerged as crucial metadata included file system identification, file listings (for each file system), creation and access dates, file size, hashes, and basic file identification.

Some additional file system attributes for discs that included HFS partitions, like the size of the resource fork, creator, and type, were also included. Once the project team identified the desired set of metadata elements, they then determined what utilities were needed to gather all of the information. The project team was adamant that no single tool should drive the decision about what to include or exclude in the metadata, and carefully reviewed the capabilities and limitations of a number of utilities. Through this review, the team discovered, for example, fiwalk cannot produce metadata for HFS formatted discs. By using a custom script and a range of tools—including The Sleuth Kit suite of utilities, hfsutils, and others—the project team was able to generate various outputs to feed into another script that would structure the information in valid DFXML, a well-known standard in the community.

Investigation of specific works

The following section provides three examples of analysis done on select artworks from the collection, focusing on the challenges and how the project team addressed them.

#FFFFFF by Art Jones (2001)

#FFFFFF is an interactive multimedia collage that explores themes such as race and masculinity in consumer culture. This work presented a curious challenge: there were discrepancies between the artist's intent for the work and the technical capabilities of the disk that contained the work. First, the system requirements stated that the work functioned on either a Windows or Macintosh system, but the disc only had an HFS file system present, meaning it was only Macintosh compatible. During testing, the team noticed that the work occasionally froze when running on an emulated Macintosh system—which consisted of a Mac OS 9 installation running within SheepShaver[15]—so they wanted to test it an emulated Windows system.

In order to do this, the project team needed to create an ISO-9660[16] formatted disk image from the files contained on the original HFS-formatted disc. Once this derivative disk image was made, it was loaded into an emulated Windows system, which was a Windows 2000 installation running within QEMU.[17] Once the emulated system was running with the artwork loaded, the project team noticed that Shockwave 7 was required to view the work. The version included on the original media was for Shockwave's web installer; the final steps of that installation launched a web browser to download the remaining files from a

[15] http://www.cebix.net/sheepshaver

[16] http://en.wikipedia.org/wiki/ISO_9660

[17] http://www.qemu.org

website that no longer exists. The project team found the original full installer for Shockwave 7 (on a software repository online) that contained all of the files needed to complete the installation (and did not require fetching additional files from the web). They included this version of the installer on the new ISO-9660-formatted disk image created for Windows access to the artwork.

While authenticity is a key concern for archivists, in considering this artwork, it can be argued that the work's "authenticity" may best be understood in terms of fidelity to an artist's vision. The project team inferred the artist's intent through the work's documentation and began drafting interview protocols for further investigative conversations with the works' creators. While, in this instance, they will preserve the original disk image that represents the exact digital material on the physical CD-ROM the artist produced, they will also preserve the derived disk image with the alternate file system and replacement Shockwave 7 installer that allows a user to interact with the work in a Windows environment.

Beyond Manzanar by Tamiko Thiel and Zara Houhsmand (2002)

Beyond Manzanar uses 3D-rendering browser plugins to create an experience that places the user in an interactive, immersive environment set against the backdrop of the Japanese-American internment camp at Manzanar. This work provides a compelling case for using emulation to access a work that is meant to be viewed entirely within a web browser and consists of file formats still in use today, such as HTML and JPG, GIF, and PNG image formats. Often "browser-based" works, such as this one, require third-party plugins that can no longer reliably work on a modern system. In the case of Beyond Manzanar, the work included virtual reality components that the artist stated could only render properly using Blaxxun Contact VRML Browser (also included on the disc). The project team found that this work functioned best in a virtual machine running an older version of Windows. The project team tested VirtualBox with a Windows 2000 installation for this artwork.

After that, the main challenge for this was providing an experience that matched the artist's vision. The artist originally intended the work to be installed in a room with images projected on three walls to provide a fully immersive experience for the viewer; additionally, the work's stated system requirements indicated that a powerful graphics card was key. Since the project team could not provision for the original intended environment, they consulted the artist's website and looked at reference images to determine how closely they could approximate the artist's original vision, running a virtual machine with Windows 2000.

Once the project team configured the VRML browser to the artist's exact specifications, they noticed a significant improvement in quality and rendering of the work. For example, the rendering of textures improved. However, they also noticed that in some cases, the graphics in the emulated system were simply nowhere near the quality of those on the artist's website. Specifically, the text overlays on several images in the local version were fuzzy while the artist's version was not. By investigating and finding the exact PNG files contained on the disk image, the project team noticed the archived version included anti-aliased text with drop shadows, where the artist's version did not. The project team ultimately determined that the reference images on the web were fundamentally different than those provided on the Goldsen's copy, and, as such, the apparent reduction in quality was not an artifact of emulation or hardware (see Fig. 1). The project team could only support intent with the images in the work, so in this case the image quality could not be improved without further follow-up with the artist.

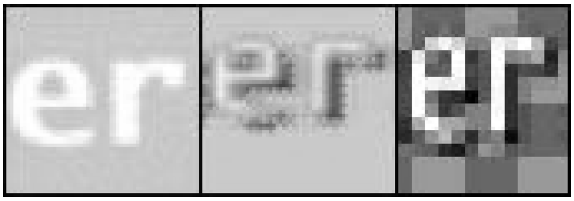


Fig. 1. Left: Reference image from the artist's website. Middle: Image from work in emulation. Right: Showing image transparency and drop shadow.

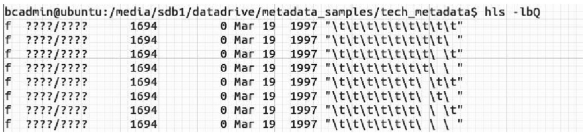
Just from Cynthia, by Albert Sorbelli (2001)

Just from Cynthia (2001), produced by Albert Sorbelli, is a compilation of artworks from the X/Y exhibition at the Centre Georges Pompidou. Investigating this work prompted the team to consider emulation as a key strategy for the analysis of a work, in addition to a method for providing user access. While reviewing the list of files included on this HFS-formatted disc, it emerged that there were approximately twenty files that appeared to have no name at all. Further investigation—by setting the tool used to list HFS files (hls) to escape special characters—revealed that the mysterious files each had a distinct name consisting of a combination of tabs and spaces, and that the size of each of the data forks was zero bytes (see Fig. 2).

The project team had encountered, on other HFS-formatted discs, instances of desktop icons whose sole purpose lay in their screen position, allowing them to function like a context cue for users. In a file listing, these files often appeared out of order but their names often revealed their purpose. In this case, without any obvious filenames, the project team decided to view the work in emulation to determine what might be happening, because none of the disk image analysis tools could confirm a full explanation.

Once the project team viewed the work in an emulated Macintosh system—running an installation of Mac OS 8 within Basilisk II[18]—it was clear that these were indeed icon files, arranged in a large mosaic graphic that became visible when viewing the contents of the work in a Finder Window (see Fig. 3). The many filenames that consisted of whitespace characters appeared as a solid block of colorin the larger mosaic. This becomes apparent when one of the tiles in the mosaic is moved elsewhere (see Fig. 4). Though they added to the interactive experience of the work, it became clear that the files were more of a decorative

[18] http://www.cebix.net/basiliskii



```
bcadmin@ubuntu:/media/sdb1/datadrive/metadata_samples/tech_metadata$ hls -lbQ
f  ????/????      1694          0 Mar 19  1997 "\t\t\t\t\t\t\t\t\t"
f  ????/????      1694          0 Mar 19  1997 "\t\t\t\t\t\t\t\t\ "
f  ????/????      1694          0 Mar 19  1997 "\t\t\t\t\t\t\t\ \t"
f  ????/????      1694          0 Mar 19  1997 "\t\t\t\t\t\t\t\ \ "
f  ????/????      1694          0 Mar 19  1997 "\t\t\t\t\t\t\ \t\t"
f  ????/????      1694          0 Mar 19  1997 "\t\t\t\t\t\t\ \t\ "
f  ????/????      1694          0 Mar 19  1997 "\t\t\t\t\t\t\ \ \t"
f  ????/????      1694          0 Mar 19  1997 "\t\t\t\t\t\t\ \ \ "
```

Fig. 2. Portion of the hls listing of "whitespace-named" characters in Just from Cynthia. File names are in the rightmost column.

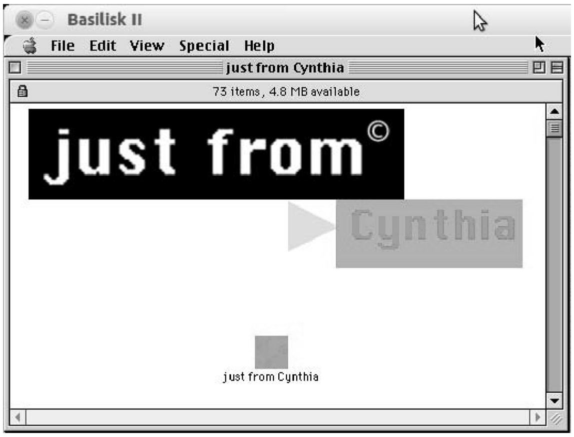
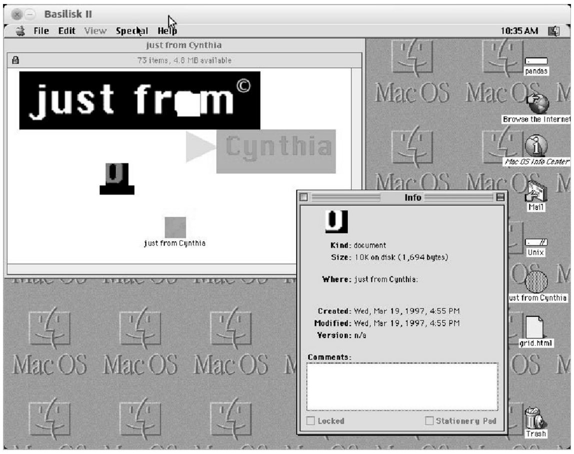


Fig. 3. Graphical mosaic of icon files in *Just from Cynthia*.

element within the work, rather than critical to its functioning within the operating system. Since the existence of these files in the DFXML metadata may be confusing, the project team has documented their investigation to provide the context necessary so that the purpose of these files is clear. These annotations may inform future archivists trying to understand digital artwork such as this.

Conclusion

In each of the cases presented, bitwise fidelity (integrity) could be seen to be at odds with the artists' intent: the project team had to analyze an anomaly (i.e., obsolete plugin installers, embedded Windows executable files on a Macintosh-formatted disc, discrepancies in quality of image files, icon files with confusing filenames) and determine how the discrepancy affected the work and what implications this had for preservation and description. While the goals of the project team differed from those of a traditional forensic investigator, similar tools and methodologies were used. The three works presented in this paper are highlights of the discoveries found within the test bed of the Goldsen collection discs. The project team reviewed all works in the test bed and performed detailed analysis and investigation of approximately twenty to thirty key works.



Kind: document
Size: 10K on disk (1,694 bytes)

Where: just from Cynthia:

Created: Wed, Mar 19, 1997, 4:55 PM
Modified: Wed, Mar 19, 1997, 4:55 PM
Version: n/a

Comments:

□ Locked □ Stationery Pad

Fig. 4. Graphical mosaic with one moved icon file and file info display screen for same icon file.

Future directions

Given the amount of older material that archives encounter with their mission to provide access to materials, the community continues to investigate whether emulation is a viable strategy for preservation of access. There is current research on various emulation access options, including the development of Emulation as a Service (Von Suchodoletz et al., 2013; Valizada et al., 2013), which aims to provide the technological framework to serve up emulated systems. For digital artwork, where the context (e.g., an older operating system) can be critical to an authentic experience of a work, this line of research is especially of interest to curators and archivists, including the project team.

Development of best practices for the accession of digital materials is also important to the archival community. For digital artworks, this can include artist interviews that address hardware and software requirements, providing for the preservation of source code, and planning for the ongoing preservation and access of the work. (See the Variable Media Questionnaire,[19] for further reading on the topic.) This is also ongoing work for the project team.

Recommendations

Some archivists may be working in environments where they do not have complete control over their systems, and some tool developers from the digital preservation community have structured their tools accordingly (i.e., the AVPreserve tool Fixity[20] does not require elevated or administrator privileges in Windows or MacOS; BitCurator can also be run in a virtual machine for users who cannot have a standalone, dedicated Linux machine). Tool developers should be mindful of the fact that while some users face such limitations, others do not; tools should target a range of expertise, system access, and support but not require a lower level of control.

The project team also found that existing forensics tools needed extensive adaptation to provide the technical information determined critical for all discs. Given the age of the collection, there were a number of CD-ROMs that included HFS file system data. Since HFS file systems are

[19] http://variablemediaquestionnaire.net/
[20] http://www.avpreserve.com/tools/fixity/

not supported by The Sleuth Kit, which drives select reporting and metadata creation tools within BitCurator, the project team put in considerable effort writing scripts that could pull the output from multiple utilities so that all file system metadata could be included in a single DFXML file. While keeping up with new technological developments is certainly of interest to archivists, there is also a strong need for developing tools to support analysis of older technologies.

Finally, archivists often receive digital material on storage media that can be fragile and in obsolete formats. Since current forensics tools focus more on current technologies, it can be difficult to work with older materials. For example, some archives are trying to rescue data from 5.25″ and 3.5″ floppy disks whose drives have long since disappeared from computer systems; the UltraBlock SCSI, a writeblocker for SCSI hard drives, has been discontinued.[21] Archivists are pursuing multiple strategies for handling older media, including sourcing hardware from eBay (or similar sites), and custom building new systems (Durno and Trofimchuk, 2015). In this context, sharing information on how to work with potentially 20–30 year old hardware and rescue data in a forensically sound way is vital[22] because older tutorials and walkthroughs may not be maintained by their creators.[23] Work done by the forensics community to understand and reverse engineer current hardware in software may be of use to archivists long after the forensics community has need for it. Saving as much information as possible will likely have benefits to archivists working decades from now.

Acknowledgments

The analysis of the artworks referenced in the case study section of this paper has been made possible by a grant from the National Endowment for the Humanities (PR-50182-13); the authors acknowledge the support of Cornell University Library, in particular Digital Scholarship and Preservation Services, the Division of Rare and Manuscript Collections, and the Rose Goldsen Archive of New Media Art. The authors also wish to extend their appreciation to Madeleine Casad and Steven Romig who provided helpful feedback on early drafts of this paper.

References

Adelstein F. Live forensics: diagnosing your system without killing it first. Commun ACM 2006;49:63–6.

Brezinski D, Killalea T. Guidelines for evidence collection and archiving. 2002.

Brown G. Developing virtual CD-ROM collections: the voyager company publications. Int J Digit Curation Dec. 2012;7(2):3–20.

Casad M. D-Lib magazine. In: Brief and in the news; 2013.

Casey E. Digital evidence and computer crime. Academic Press; 2000.

Durno J, Trofimchuk J. Digital forensics on a shoestring: a case study from the University of Victoria. Code4Lib J 2015;(27).

Erway R. First steps for managing born-digital content received on physical media: you've got to walk before you can run. 2012.

Garfinkel S. Digital forensics XML and the DFXML toolset. Digit Investig 2012;8:161–74.

Grace Stephen, Knight Gareth, Montague Lynne. Investigating the significant properties of electronic content over time: Final report. 2009.

Hedstrom ML, Lee CA, Olson JS, Lampe CA. “The old version Flickers more”: digital preservation from the user's perspective. Am Arch 2006;69(1):159–87.

Iraci J. The relative stabilities of optical disc formats. Restaurator 2005; 26(2):134–50.

Kirschenbaum M, Ovenden R, Redwine G. Digital forensics and born-digital content in cultural heritage collections. 2010.

Lee C, Woods K. Enabling digital forensics practices in libraries, archives and museums: the BitCurator experience. In: Digital forensics research workshop 2014; 2014.

Lee C, Woods K, Kirschenbaum M, Chassanoff A. From bitstreams to heritage: putting digital forensics into practice in collecting institutions. 2013.

Library of Congress and National Institute of Standards and Technology. Final report: NIST/LC optical disc longevity study. 2007.

MacNeil H, Wei C, Duranti L. Authenticity task force report. 2001.

McKemmish R. What is forensic computing? Trends Issues Crime Crim Justice 1999;(118).

Pearce-Moses R. A glossary of archival and records terminology. Society of American Archivists; 2005.

Valizada I, Rechert K, Meier K. Cloudy emulation—efficient and scalable emulation—based services. In: iPres 2013 10th international conference on preservation of digital objects; 2013.

Von Suchodoletz D, Rechert K, Valizada I. Towards emulation-as-a-service: cloud services for versatile digital object access. Int J Digit Curation Jun. 2013;8(1):131–42.

Yeo G. “Nothing is the same as something else”: significant properties and notions of identity and originality. Arch Sci Jun. 2010;10(2):85–116.

[21] http://www.digitalintelligence.com/products/ultrablock_scsi/

[22] http://www.nypl.org/blog/2012/07/23/digital-archaeologyrecovering-your-digital-history

[23] http://web.archive.org/web/20140119144844/http://mith.umd.edu/vintage-computers/fc5025-operation-instructions