Preserving and Emulating Digital Art Objects

Dianne Dietrich, Cornell University Library, Ithaca NY USA; Madeleine Casad, Cornell University Library, Ithaca NY USA; Jason Kovari, Cornell University Library, Ithaca NY USA

Abstract

In February 2013, Cornell University received a $300,000 grant from the National Endowment for the Humanities to develop PAFDAO: preservation and access frameworks for complex digital media art objects. This research and development project was undertaken in cooperation with the Rose Goldsen Archive of New Media Art, part of Cornell University Library's Division of Rare and Manuscript Collections. This paper provides an overview of PAFDAO's outcomes, focusing on technical and also curatorial components of the project that might be applied to comparable collections at other institutions. The authors outline their imaging workflow and metadata framework, and describe their methods for addressing questions of cultural authenticity and considerations related to their decision to adopt emulation as an access strategy.

Introduction

From 2013–2015, a seven-person project team at Cornell University worked on PAFDAO,[1] a project to develop preservation and access frameworks for complex, interactive, born-digital media art objects housed in the Rose Goldsen Archive of New Media Art, a part of Cornell University Librarys Division of Rare and Manuscript Collections. The project was supported by a $300,000 grant from the National Endowment for the Humanities' Division of Preservation and Access, with the aim of creating workflows and processes that would assist other institutions with similar collections of born-digital material.

Preserving such assets poses a serious challenge to today's archiving institutions. There are, as yet, no established best practices for how to preserve such materials, even though they represent a vitally important segment of our cultural heritage that is already subject to serious risks. Many complex born-digital assets are housed on fragile storage media like optical discs; they face degradation (data decay or "bit rot") themselves, and, moreover, often cannot be accessed without the use of obsolete legacy hardware and software. Such assets are furthermore notoriously complex, involving many different kinds of digital media files and dependencies on software and operating systems that may themselves be obsolete. A relatively small glitch in the system, for example an obsolete media-player plugin or a single unreadable media file, can have a powerful effect on the way an asset renders. Especially in the case of artworks, such rendering problems can negatively impact the transmission of the asset's meaning.

The test bed for the PAFDAO project included approximately one hundred interactive born-digital artworks created for CD-ROM, many of which date back to the early 1990s. In most cases, these CD-ROMs containing interactive new media artworks were created for exhibition on small-screen monitors in both private and public spaces. The goal of the grant was to explore methodologies for the characterization, classification, preservation, and access provisions for these works, and to create scalable strategies and workflows for the preservation of interactive digital assets more generally. The PAFDAO project team has undertaken an in-depth analysis of different disk imaging applications, explored access options and determined that emulation is indeed feasible at scale, and developed a multi-part strategy for preservation, technical, and descriptive metadata.

The remainder of this paper provides an overview of activities that happened more or less concurrently and iteratively over the two-year project period, undertaken by different sub-groups of the project team in consultation with one another. The authors have attempted to present these activities in a logical way. The section that follows details our efforts to contact and characterize the needs of a broad community of users of media art archives. The section following that outlines the project team's approach to bit-level preservation of test collection materials through disk imaging. Next, the paper documents the team's investigations of emulation as an access strategy. Description of the varied metadata produced follows, as well as considerations for deposit of these assets into the Cornell University Library Archival Repository (CULAR).

User Community and Access

From the beginning, the PAFDAO project team aimed to develop both metadata frameworks and access strategies that would genuinely reflect the needs of the media archives user community. Toward this end, the team developed a questionnaire to assess collection users' practices and preferences. The questionnaire was designed to elicit open-ended, qualitative responses from a wide spectrum of respondents. Questions inquired about respondents' position in the media archives community and the kinds of queries that drove their archival work. Respondents were asked, among other things, to describe their purposes, preferences, and any frustrations they experienced when working with collections of digital media objects. The questionnaire specifically inquired about respondents' preferred strategies for access and exploration of media art collections.

The questionnaire was disseminated by email to curatorial, art historical, media arts, digital libraries, and digital humanities listservs. A preliminary report of survey findings is available at Cornell University Library's Digital Scholarship and Preservation Services blog;[1] a more comprehensive report will be published under separate title and publicly available in the PAFDAO white paper.

The initial goal of the questionnaire was to develop "personas," or user profiles that would encapsulate the needs and preferences of different kinds of media archives patrons. Responses to

[1] Project wiki: `https://confluence.cornell.edu/display/pafdao/Home`

the questionnaire were rich and incredibly informative; however, no coherent user profiles emerged from the results. Rather, results pointed to a wide variety of perspectives and preferences. Individual respondents often expressed their views strongly, but individual positions equally often proved to be contradictory when responses were considered in aggregate. Because of this, questionnaire results were not as directly prescriptive as the PAFDAO team might have hoped at the start of the project.

Nevertheless, the questionnaire and its results had a significant impact on project development and outcomes. First of all, it allowed the project team to greatly streamline an initial metadata approach that, though thorough, would have been somewhat unwieldy to implement at scale. The questionnaire also allowed the team to rule out specific access use cases, such as providing access to artworks' source code, which, though invaluable for conservation purposes and select research foci,[2] seems to be a far less prevalent researcher need based on respondents' feedback, and will not require systemic access provisions.

Analyzing questionnaire results, the project team was also able to identify a clear concern for "authenticity" among media art researchers, which was not yet adequately addressed in our initial project development plans. In response to this concern, the team developed a conservation-oriented artist questionnaire and interview process. This is a measure undertaken by most institutions of cultural heritage with active conservation programs, but one that PAFDAO had not properly considered before the start of the project. Once considered, however, the wisdom, of opening such a conversation with artists was immediately clear. It would ensure that, even if they must be imperfect renderings, the versions of an artwork presented to researchers in the future would still be driven by the archiving institution's best efforts to identify and respect the creators' visions, aspirations, and priorities for the work. The PAFDAO artist questionnaire was informed by members of the project advisory board[2] and created with reference to two especially powerful and salient models from the media arts community: the Variable Media Questionnaire[3] and the artist questionnaire used by Turbulence.org.[4]

The PAFDAO team opted for a significantly pared-down questionnaire that reflected the unique position of the test collection as part of an academic research library, rather than a museum or gallery or similar cultural institution that would involve a broader range of media types and be guided by a more active exhibition program.

Developing the artist questionnaire toward the end of the two-year project period, the PAFDAO team was able to incorporate the findings of emulation investigations into the questions posed to artists. In addition to enlisting essential input from artists about artworks' significant properties and most important experiential characteristics, the artist questionnaire enables early discussion of the project team's intent to use emulation as an access strategy for born-digital works, and to disclose in a systematic way some of the known rendering shortcomings associated with different emulators. The questionnaire and interview invite artists' instructions for how best to present their work under varied rendering conditions. The questionnaire also broaches a basic inquiry about whether or not artists still hold the source code or working files used to create the artworks in question, and whether they would consider depositing their working files with the Goldsen collection for the purpose of future conservation work.

Disk Imaging & Visual Documentation

The works in the PAFDAO test collection consisted mainly of CD-ROMs, including retail quality CD-Rs that were burned more than a decade ago. These are fragile media with a limited lifespan, and it was of utmost importance to migrate their content while also preserving the underlying structure of the discs. The project team elected to create bit-by-bit disk images for each of the items, rather than producing logical images, which would simply copy file contents. There are a few reasons for this choice. First, many of the works had been cross-compiled for Windows and Macintosh systems: therefore, there were often multiple file systems present on each disc. Disk images preserve the exact structure of original source media, and in this case, would preserve the multiple file systems that existed for a work on disc. Additionally, many works were designed for use with older Apple hardware and had HFS file systems. Files in an HFS file system can include both a "data fork" and a "resource fork." Copying operations can sometimes fail to include the resource fork; this may render some files inoperable. Disk images ensure that all necessary components for individual files have been preserved.

The project team investigated two pieces of software to create disk images. IsoBuster[5] is a Windows-based program, and Guymager[6] is Linux-based software included in the BitCurator[7] suite of utilities. IsoBuster provides an intuitive and friendly interface for viewing the layout and structure of a disc, which proved helpful for identifying hybrid audio/data discs. The project team ultimately opted to use Guymager to create disk images, however. Guymager produces an extensive information file for each disk image created, which includes the sector addresses of any unreadable sectors and critical preservation metadata such as hardware specifications for the CD drive used in the imaging process. Guymager also optionally provides a verification step, and can provide confirmation that the same exact data was read during the acquisition and verification stages and that what was read off the disc matched what was written to the disk image file. The project team concluded that this information would be critical to future curators and users in verifying that the disk images produced were the most faithful representations of the original items.

In concert with the disk imaging process, the project team photographed and scanned all physical materials in the test collection, including the discs themselves as well as unique CD cases, booklets, or other associated materials. These image files will be deposited along with disk images in the Cornell University Archival Repository, with filetypes and metadata described in Metadata.

[2] For an excellent discussion of some of the issues at stake in such curatorial and conservation efforts, see Richard Rinehart and Jon Ippolito, Re-Collection: Art, New Media, and Social Memory, Leonardo. (Cambridge, Massachusetts: The MIT Press, 2014.)

[3] http://variablemediaquestionnaire.net/

[4] http://www.turbulence.org/

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

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

[7] http://www.bitcurator.net/

Analysis & Access

When writing the initial grant proposal for the PAFDAO project, the project team presumed that emulation would not be a scalable access strategy for the test collection artworks; the initial proposed work plan accordingly promised little in the way of emulation investigation or support. Once the grant was awarded, however, project advisors strongly advocated that the team explore emulation environments as a feasible strategy. At the outset of the project, the project team did some preliminary analysis to prepare for exploring emulation more thoroughly. First, the team analyzed the stated system requirements for the works in the test collection, drawing from information provided in the works' respective catalog records. The team simultaneously tested the works outside of emulation environment: Macintosh-compatible works were tested in a legacy hardware environment[8] and Windows-based works were tested on current PC hardware and Windows 7. The team documented any rendering problems or glitches that emerged; these notes would later serve as one kind of "control" and baseline rendering for future emulation testing and artwork classification.

Across the collection, the project team found that system requirements ranged from Windows systems as early as 3.1 and Macintosh systems as early as System 7. While Windows operating systems have historically supported running executable files built for earlier versions (e.g., 32-bit executables built for Windows 95), Macintosh operating systems typically have had a limited window of support for legacy applications. While this might suggest a dual strategy for access—modern hardware for Windows-based works and emulation for Macintosh-based works—the project team has decided to pursue emulation as a strategy for access for all works.[9] The project team found several instances where third-party plugins required for a Windows-compatible artwork no longer reliably functioned on a contemporary Windows 7 system (or conflicted with other applications on that system). Further, since the look and feel of operating systems and web browsers has evolved considerably since the 1990s, viewing a work in an emulated system is closer to an authentic experience (as suggested by the artist through stated system requirements) than viewing the work in a modern system.

Emulation is not a perfect access solution, however, and the project team has had to explore and document a number of artistic and technical considerations throughout the exploration process. The changes introduced by emulation can be dramatic: processor speeds have increased considerably, and even setting a virtual machine to a relatively "slow" setting can still result in images in a work refreshing far faster than originally intended. Running an older system on a new machine often still means using newer hardware. Here, too, change has a significant effect: CRT monitors with 4:3 aspect ratios and mechanical mice have given way to LCD monitors with aspect ratios of 16:9, and optical and laser mice and trackpads. For select works, the project team has compared the work in emulation with its performance in a legacy machine (i.e., an iMac from around 2002) and noted emulation artifacts along with possible workarounds. One example of a suggested workaround involves the recommendation to increase the the proportion of red tones on a newer LCD monitor so they more closely match the warmer tones of an older CRT monitor.

The project team tested a number of emulation and virtual machine software with the artworks in the collection, and considered various criteria when suggesting their preferred software for access. These criteria included ease of use and community support, as represented by, for example, easy to find and up to date troubleshooting forums. Ultimately, the project team recommended Basilisk II[10] and SheepShaver[11] for emulating mid-1990s Macintosh systems and QEMU[12] for emulating PC systems. All of the team's emulator testing was performed on a workstation running BitCurator: the team compiled Basilisk II and SheepShaver from source to optimize their performance on this workstation, but found that the pre-built binary for QEMU in the Ubuntu repository was stable enough for use and testing. A number of "pre-set" systems (e.g., Mac OS 8.1 for Basilisk II; Windows 95 for QEMU, Mac OS 9.0 for SheepShaver, etc.) were created for preservation purposes and to support an access workstation for users to view the works.

Emulation software is often driven by volunteers and enthusiasts, and the software used today may not be supported ten or even five years from now. Thoroughly documenting known limitations of emulators used now (see Emulation Documentation for further detail on this documentation) and continued monitoring of all newly available tools will be critical to ensure ongoing access to the Goldsen materials.

Metadata

For description of the artworks' descriptive elements, technical details, and preservation actions, the project team used a complex framework of structured and unstructured metadata. Creating a metadata framework for the project was an ongoing and iterative process, undertaken with an eye to all of PAFDAOs other "moving parts": media archives user feedback, artist feedback, disk imaging tools and workflow, emulation data, as well as the architectural requirements of the Cornell University Library Archival Repository (CULAR). Metadata collected will aid future stakeholders to assess necessary preservation actions as well as identify methods of rendering the artworks.

Descriptive Metadata

The majority of artworks in PAFDAO's testbed were cataloged prior to this project, natively available as MARC metadata and exportable as MARCXML. The MARC metadata includes expected values (e.g., artist names, titles, dates of creation, etc.) as well as basic systems requirements as these appear on inserts packaged with CD-ROMs in the test collection. While system requirements on disk inserts are often limited, this data nonetheless provides a starting point for understanding artists intentions concerning the expected appearance and behavior for individual artworks. CULAR's current structure does not facilitate referencing metadata in other repositories or databases; thus, MARCXML ingested to CULAR by the PAFDAO project team is static and intended primarily to provide a reference for collection stewards

[8] This was a G3 (PowerPC) Apple Macintosh with OS 10.2 running "Classic Environment."

[9] The project team stresses here that legacy support in Windows might not continue indefinitely, strongly suggesting the rationale to explore emulation as a strategy for Windows-based works.

[10] `http://basilisk.cebix.net/`

[11] `http://sheepshaver.cebix.net/`

[12] `http://wiki.qemu.org/Main_Page`

when they engage with materials in CULAR.

Technical Metadata

For each disk image, the project team formatted technical metadata as Digital Forensics XML (DFXML).[13] Since The Sleuth Kit (TSK)[14] digital forensic utilities do not handle HFS file systems, the project team could not solely rely on them to produce DFXML (e.g., by using only fiwalk to produce technical metadata). The team used additional utilities to capture HFS file system details and metadata (including HFS-specific details such as entry type, creator code, resource fork size, and more). Python scripts connected these various command line utilities (e.g., TSK utilities, hfsutils, etc.) to create the complete technical metadata in DFXML. The project team used a MODS note with type attributes to represent HFS-specific information within the appropriate DFXML <fileobject>. These artworks are being preserved at the disk-image level rather than at the file-level; thus, the project team is concerned with basic file identification within the disk image. While technical metadata for these disk images do not include file validation nor did the project team extract embedded metadata from files within the disk images, staff can validate files or extract metadata if future stakeholders develop use cases that warrant either file validation or the extraction of embedded metadata. In cases where file-level metadata was ambiguous or appeared erroneous, the project team viewed the work in emulation to confirm file-level details in a suitable rendering environment. Emulation thus occasionally served as an analysis tool, and not only an access strategy.

Preservation Metadata

As described in Disk Imaging, Guymager output includes details for the hardware used to create the disk images, three hashes (MD5, SHA1, SHA256), and additional metadata. The metadata contained in this output form the core of the preservation metadata and are extracted using a Python script, then formatted as PREMIS XML. In addition to Guymager-derived content, the PREMIS files include significant properties (i.e., file systems and classifications that the project team determined for each work, as described in Artwork Classifications as well as information concerning recommended rendering environments (i.e., emulators tested with the artwork) for select key works that the project team analyzed more carefully. In a few instances, conservation efforts demanded the creation of derivative disk images. For example, the project team created an ISO9660-compliant disk image for an artwork where system requirements indicated that the artist intended for the artwork to function on a Windows machine but the original disc was a Macintosh-formatted disc. For these derivative disk images, the project team created a PREMIS XML for the derivative disk images, which document details concerning rationale and process for the creation of the derivative.

Artwork Classifications

To aid in use of these artworks, the project team created narrative artwork classifications and classification-level documentation. The documentation includes descriptions of the classifications' properties, implications and strategies for accessing and rendering works (e.g., recommendations and caveats about running works in various emulators), and information concerning their restoration potential.[15] Initial classifications were determined by file systems and file types present on the disc, as well as by stated technical requirements of the work. The project team found that artworks' significant properties were most often non-restrictive classifications, and that a single work often required referencing multiple overlapping classifications, each with its own preservation and access implications. During the classification process, the project team relied on notes from the original review of the works. The artworks' PREMIS files include references to these classifications.

Emulation Documentation

The project team tested out various emulation software. Three of the emulators tested—Basilisk II, SheepShaver, and QEMU—are open source and ongoing development is largely community-driven. Given this, the project team felt it prudent to fully understand the capabilities and limitations of the proposed access and emulation strategy. The project team documented the configuration and setup used, including configuration flags for compiling Basilisk II and SheepShaver from source, and noted any issues that surfaced while testing the software. It is possible that future access to these may rely on alternate emulation software, and it will be helpful to know how potential future strategies compare to current and previous ones.

Sector Notes Documentation

Given the fact that the collection included older CD-ROMs and retail quality CD-Rs that were burned over a decade ago, the project team anticipated that some of the media may be partially unreadable. Since the project team completed two rounds of testing with different software (i.e., IsoBuster and Guymager) and hardware (i.e., an internal CD-ROM drive and one that connected to the workstation using USB), they were able to compare the results from both rounds of imaging when either piece of software reported that there were bad or unreadable sectors on a disc. When this happened, the project team carefully analyzed each disk image to determine which scan best represented a faithful copy of the original media. The project team documented all steps taken during this process so that future curators know why a particular scan had been chosen for preservation.

Artwork Visual Documentation

To enhance documentation of the artworks, the project team captured images of the containers and inserts packaged alongside the CD-ROMs, as described in the Disk Imaging section, and saved these images as TIFFs. Metadata for these TIFF files are formatted as VRA core metadata.

Ingest

As part of the preservation workflow, disk images, metadata and other digital files are ingested to the Cornell University Library Archival Repository (CULAR), a home-grown archival repository based on Fedora, which structures deposits as Aggregates, Resources, and Metadata. During the ingest, each high-

[13] `https://github.com/simsong/dfxml`

[14] `http://www.sleuthkit.org/`

[15] Restoration options are projected by the teams technical understanding of the technologies used for the artworks.

level aggregate receives EAD-XML collection-level description. Within these aggregates, each resource undergoes JHOVE validation upon ingest; while JHOVE will not validate many of the filetypes in this collection, validation is a routinized part of the ingest process. The JHOVE XML is discoverable alongside the MARCXML, DFXML and PREMIS XML for each artwork.

The PAFDAO ingest aligns with the structure preferred by the Division of Rare and Manuscript Collections, under which the Goldsen Archive and these materials are administered. At the top-level, PAFDAO materials are aggregated according to the archival collection from which they derive; at this aggregate, collection-level EAD-XML metadata describes the archival collection. An aggregate for documentation and aggregates for the artworks are located within each archival collection aggregate; documentation includes documentation about the ingest process as well as any documentation such as the Artwork Classifications. For each artwork, an aggregate wraps aggregates for disk image(s) and for the TIFFs of the artwork visual documentations. Affixed to the appropriate aggregate are the PREMIS, DFXML, and MARCXML files.

Conclusions

While PAFDAO represents a significant leap in the institutional preservation of interactive new media art, the team recognizes that is merely the first-phase of a long-term preservation strategy. Digital preservation is an ongoing process of curation, management, and provisioning for access, and the preservation of digital content must go hand-in-hand with active support for its use. The successes of this project have already facilitated broader access to interactive born-digital content held in the Goldsen Archive within research and pedagogical contexts at Cornell University.

A major goal of the PAFDAO project was to develop a scalable workflow for the preservation of born-digital interactive artworks, with the intention that this framework could be implemented at other institutions that hold similar materials. Concluding project reports will include clear documentation and recommendations of the teams findings in these areas.

The project was also framed as a research endeavor, however, and many of the team's findings may be of value even if they do not address directly the implementation needs or archival situation of comparable institutions. In many areas, PAFDAO project efforts exceeded the original scope of the work plan. The need for these workplan expansions stemmed from discoveries made along the way—for example, the team's adaptation of initial preservation and access frameworks and workflows when it became clear that emulation would be a viable access strategy, or the development of artists interviews in order to address the concern for cultural authenticity that emerged from our survey of media art researchers' needs.

In some cases, these activities reflected the uniqueness of the Rose Goldsen Archive's position intersecting the institutional contexts of media art conservation and research library. From the start of the project, the PAFDAO team recognized the need to match preservation and access goals to institutional mission and capacity as well as to patron needs. Offering both general recommendations and well-documented case studies, the final PAFDAO report will, we hope, be of value to archiving institutions from all quarters of the cultural heritage preservation community.

Acknowledgments

The authors gratefully acknowledge the support of the National Endowment for the Humanities, the PAFDAO project advisory board, consultants AVPreserve, and the entire project team.

References

[1] "Interactive Digital Media Art Survey: Key Findings and Observations: DSPS Press," accessed March 18, 2015, http://blogs.cornell.edu/dsps/2014/07/30/interactive-digital-media-art-survey-key-findings-and-observations/.

[2] Deena Engel and Glenn Wharton, "Reading between the Lines: Source Code Documentation as a Conservation Strategy for Software-Based Art," Studies in Conservation 59, no. 6 (November 2014): 40415, doi:10.1179/2047058413Y.0000000115.

Author Biography

Dianne Dietrich was a fellow in the Digital Preservation and Scholarship Services unit at Cornell University Library from 2013–2015. During that time, she served as technical lead and Digital Forensic Analyst on the NEH-funded grant, Preservation and Access Framework for Digital Art Objects. She holds a degree in library science from the University of Michigan and a BA from Wesleyan University.

Madeleine Casad is Associate Curator of the Rose Goldsen Archive of New Media Art and Curator for Digital Scholarship in the Digital Scholarship and Preservation Services unit of Cornell University Library. Her curatorial work emphasizes the cultural, aesthetic, and historical dimensions of interactive digital interfaces, and the challenge of supporting research into the history of digital culture by preserving access to such digital artifacts. She holds a PhD from Cornell University and BA from the University of Iowa.

Jason Kovari is Head of Metadata Services and Web Archivist at Cornell University Library (CUL). In addition to providing direction for the Metadata Services unit, Kovari participates on a variety of digital collection building and digital preservation efforts across Cornell, including his service on the Cornell University Library Archival Repository (CULAR) steering group. He holds an MLS from SUNY Buffalo and BA from Binghamton University, SUNY.