Provenance, Journal of the Society of Georgia Archivists

Volume 24 | Number 1 Article 3

January 2006

When Not All Papers are Paper: A Case Study in Digital Archivy

Catherine Stollar Peters
University of Texas Austin

When Not All Papers Are Paper: A Case Study in Digital Archivy

Catherine Stollar Peters

INTRODUCTION

Hypertext poet Deena Larsen is worried about the potential loss of her digital poetry, but she has a plan to save it. In a 2004 article, "The Uncertain Fate of Scholarly Artifacts in a Digital Age," Larsen revealed her plans for preserving her hypertext work *Marble Springs*.[1] "Ms. Larsen started collecting old Macintosh computers so people will always be able to read *Marble Springs* in its original format. She has 100 computers in her two-bedroom apartment." Although Larsen's two-bedroom mausoleum of circa 1990s technology is one strategy for saving born-digital hypertext works, it is probably not the best. An armada of aging hardware will not protect digital objects from hard drive crashes, hardware failure, inoperable software, operating system malfunctions, unreadability, or natural disasters. Preservation of electronic records requires a commitment to active preservation practices including migration, refreshing, and

[1] Scott Carlson, "The Uncertain Fate of Scholarly Artifacts in a Digital Age," *Chronicle of Higher Education* 50, no. 4 (January 30, 2004) (online resource) <http://chronicle.com/weekly/v50/i21/21aa02501.htm> (accessed April 17 2006).

integrity and authenticity checks of stored digital records. Maintaining the status quo, regardless of the magnitude of hardware and software stockpiles, is not a viable preservation strategy. The Electronic Literature Organization (ELO) notes the inadequacy of just holding onto digital materials and advocates more active digital preservation strategies in their latest publication, *Born-Again Bits*: "The stakes are even higher when we consider that keeping works of electronic literature alive in their original form does not serve all present needs, let alone those of the future."[2]

DIGITAL PRESERVATION AT THE HARRY RANSOM CENTER

Like Larsen and ELO, the Harry Ransom Center is concerned with preserving digital literature. The Ransom Center, a collecting arts and humanities archives located at the University of Texas at Austin, recently acquired the archive of hypertext author and Vassar professor Michael Joyce. In addition to authoring perhaps the most influential hypertext novel, *Afternoon, a Story*, Michael Joyce wrote, along with Jay David Bolter and John B. Smith, the hypertext authoring and reading software Storyspace. The Michael Joyce Papers, composed of both paper-based and digital materials, contain his early linear fiction and other works, correspondence, personal papers, and writings by his contemporaries, including Deena Larsen. In acquiring the Michael Joyce archive, the Ransom Center has the opportunity to preserve rare and unique electronic files documenting the creation and evolution of hypertext fiction.

As hypertext has facilitated new relationships between narrative and technology, digital preservation strategies have forged new connections between traditional archival practice and technology. Technology provides tools that allow for new methods of archival practice, such as a flexible arrangement of electronic files compared to static arrangement of papers-based records and new methods of marking up information in and about files such as Encoded Archival Description (EAD), Qualified Dublin Core (QDC), and other metadata schemas. The innovative natures of hypertext and digital preservation make hypertext an ideal narrative form and Michael Joyce an appropriate author

[2] Electronic Literature Organization, *Born-Again Bits* (August 5, 2005): 1 (online resource) <www.eliterature.org/pad/bab.html> (accessed April 24, 2006).

with which to begin our program of digital preservation at the Ransom Center.

PROJECT DESCRIPTION

In January 2005 I participated in the first phase of a project to preserve the paper and digital records of Michael Joyce at the Ransom Center.[3] Along with fellow project participants Thomas Kiehne and Vivian Spoliansky, I enrolled in a digital preservation course taught by Dr. Patricia Galloway at the School of Information at The University of Texas at Austin. We spent five months preparing, arranging, describing, and ingesting the first accession of 371 3.5-inch floppy disks, totaling 211 megabytes, of Joyce's files into an institutional repository developed by the Massachusetts Institute of Technology and the Hewlett-Packard Company called DSpace, based on the Reference Model for Open Archival Information System (OAIS).[4] Currently, I am processing the second accession of the Joyce Papers, composed of twenty-six linear feet of papers and eight gigabytes of digital files, including the contents of two hard drives saved to two DVDs, three CD-ROMs, and files from one laptop.

There are programs that create and manage institutional repositories, but DSpace software met our needs best. The School of Information created a DSpace institutional repository, and we chose to use it for this project because it is open-source software, which can be modified by a programmer, has a large user community, is frequently updated, and handles files without damaging the original bitstream. DSpace wraps digital objects with a metadata file relative to the object instead of altering the original. DSpace also maintains the integrity of ingested files by creating a copy of the original file when downloaded and automatically creates an MD5 hash value for each file ingested. With our DSpace repository, we are able to preserve the original bitstream and metadata about the original bitstream of digital

---

[3] Thomas Kiehne, Vivian Spoliansky, and Catherine Stollar, "From Floppies to Repository: A Transition of Bits" (May 11, 2005) (online resource) <https://pacer.ischool.utexas.edu/handle/2081/941> (accessed April 18, 2006).

[4] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS) (online resource) <http://public.ccsds.org/publications/archive/650x0b1.pdf> (accessed October 17, 2006).

objects for refreshing, migration, and emulation of hardware or software components. Additionally, DSpace meets the needs of our scholars who can use file comparison and analytical utilities that reveal information about electronic literature and other digital works solely from comparing bitstreams maintained in a DSpace institutional repository.

DIGITAL ARCHEOLOGY AND BITSTREAM PRESERVATION

The advanced age of the first accession of the 3.5-inch floppy disks caused concern and required additional digital archeology to recover data from the disks. The earliest of Joyce's files were created in the mid-1980s, thereby necessitating the creation of a digital preservation strategy to prevent loss to media failure or software inoperability. Our digital preservation strategy was to remove the contents of the decaying disks to the hard drive of a processing computer, mainly a Macintosh running both OS X and Mac Classic (OS 9), and upload the files into a DSpace repository hosted on a server at the School of Information. These disks were created using "classic" era Macintosh software and hardware. During our exploratory tests using a Macintosh OS X computer with an external USB floppy drive we encountered some difficulty accessing the disks. This was not surprising as many of the floppies arrived at the Ransom Center labeled "unreadable." We knew that Joyce requested that a student assistant survey all of the disks before sending them to the Ransom Center and found most disks unreadable with hardware and software not contemporary with the earliest disks. Fortunately, older Macintosh hardware components with integrated floppy drives were readily available at the Ransom Center and allowed most of the content of the first accession of disks to be migrated from floppy disks to the hard drive. Only files created by Joyce or other electronic works were removed from floppy disks. Disk utilities and other programs on the disks were used to help recover files but were not migrated to a hard drive for preservation due to migration restrictions on the copyrighted third-party disk utilities and use issues of the third-party executable files.[5]

The age of disks in the first accession also caused concern due to potential viruses, disk errors caused by corroded or dirty surfaces on the disk and floppy drive, and unsupported, out-of-

[5] Kiehne, "From Floppies to Repository," 3.

date proprietary software. These concerns were readily addressed using software, usually open-source, and hardware contemporary to the disks. Surprisingly, few files were unrecoverable from even the oldest disks. Some files written in Microsoft Word 1.0 and WriteNow were recovered but were undecipherable when opened in plain text form. Fortunately, Michael Joyce retained copies of outdated software like HyperCard and a file compression/decompression utility called Compressor that allowed us to recover files which were otherwise inaccessible.

Most of the digital archeology tasks performed to recover digital files from the floppy disks were time-consuming due to limited functionality of the programs we used: no utility existed that would perform all the digital archeology tasks we desired at one time. One of the main results from the data-recovery portion of this project is a recommendation to use integrated, open-source utilities that would complete the tasks of virus checking, file recovery, file listing or catalog creation, duplicate recognition, and file integrity checks to automate and streamline digital archeology tasks necessary for preservation. Open-source tools are recommended because they are usually less expensive and can be easily modified to meet institutional needs by a staff member with computer programming skills.

ARRANGEMENT

After recovering most of the bitstreams from the first accession of 371 floppy disks, we began the process of archival arrangement. In the beginning, we asked ourselves some questions. Can and should digital files be arranged like paper-based records? Should we heed traditional archival arrangement practices or follow theories of arrangement based on item-level metadata? Do electronic records have a natural hierarchy that can be expressed in a traditional arrangement? Should physical housing for digital materials be kept? If so, where? Should we retain exact duplicates? Our answers to these questions are not definitive, but we came to a compromise incorporating basic tenets of archival theory with features of on-demand, flexible file arrangement using item-level metadata.

Analyzing the relationship between physical materials and digital materials with similar content within the Michael Joyce archive helped us determine an arrangement strategy. After accessioning the paper-based portion of Joyce's archive,

we noticed that a number of digital materials within the archive had a paper-based counterpart, demonstrating that Joyce created both digital and analog records while performing the same activities. For example, his paper drafts of the linear novel *Going the Distance* were written by hand or (if born digital) were printed. He created similar electronic counterparts to the paper documents as Microsoft Word and Storyspace drafts. Joyce created additional versions of *Going the Distance* in the reading and authoring software called TK3 published by Night Kitchen. One-to-one relationships also exist between some of his e-mail messages that exist as both electronic and printed copies. Both formats of records were created synchronously, and at an institution like the Ransom Center that preserves not only influential works but also maintains the context in which those works were created, an arrangement demonstrating that synchronicity would best describe the creation of Joyce's records. Although his electronic and paper materials would be housed separately, we chose to arrange all of his materials using the same functional series, as opposed to series based on format, to demonstrate the original order in which Michael Joyce created his papers.

Additionally, we mapped the arrangement of the Michael Joyce Papers to the DSpace environment. Institutional repository software like DSpace can facilitate digital object arrangement into functional groups by using the community, sub-community, collection, sub-collection, and item-level hierarchy in DSpace. We mapped DSpace's hierarchies to traditional archival hierarchical levels as follows: communities equate to archival *fonds*, sub-communities to series and sub-series, collections to other layers of granularity within a series, and item-level entries relate to digital objects. In an additional level of granularity, items composed of multiple sub-components (i.e. Web sites with multiple linked HTML files) can be ingested as bundled files.

Another instance of the relationship between physical and digital objects is the housing for digital files. Electronic media, like the original floppy disks and CD-ROMs, as well as jewel cases and paper folders housing published digital works written by Joyce or other hypertext authors, directly correspond to digital files. Previous policies and procedures at the Ransom Center dictated that electronic media should be physically housed in Hollinger boxes separate from the rest of the paper-based materials. This separation policy apparently arose out of concern for potential

damage to other materials caused by degrading electronic media. However, no studies on electronic media degradation have found any examples of off-gassing or other damaging effects of filing electronic media with paper-based materials.[6] Based on our research findings, we chose to interfile housing from digital objects, like jewel cases and magnetic disks, with the paper material we received in the second accession of Joyce's materials. Although we integrated all physical components contained in the second accession of Joyce's archive regardless of physical format, we kept the first accession of 371 floppy disks separate from the rest of the archive to maintain the original order in which we received the disks. We associated digital files ingested into DSpace with the numbers we assigned to each floppy disk and for the sake of convenience chose to maintain the numbered order we created for the first accession floppy disks.

Although we integrated Joyce's digital objects into a functional group arrangement similar to his paper-based records, we also took advantage of the flexible nature of digital object arrangement by enabling on-demand, user-controlled arrangement by item-level metadata. Metadata at the item-level reveals the entire contents of an archive as opposed to traditional series arrangements that only reveal higher levels of description (such as "Correspondence, 1964" or "Works, A-G"). Preservation of digital objects depends on item-level metadata used to document, migrate, emulate, authenticate, and preserve them. Item-level metadata recorded for preservation also enables flexible arrangement of digital objects. At the heart of DSpace, like most repositories based on the Open Archival Information System (OAIS) reference model, is a database populated by individual digital objects supported by content, context, and structure metadata, and the arrangement of those objects depends on the user interface for the database. Digital arrangement allows archivists and users multiple options for organizing objects depending on the parameters set by the user interface, such as file name, title, author, date created, subject, collection, or other metadata element. Arrangement is limited only by the skills of the programmer developing the user interface used to access the database and the precision of metadata recorded for each object.

[6] Fred R. Byers, *Care and Handling of CDs and DVDs--A Guide for Librarians and Archivists,* (Washington, DC: CLIR, 2003) (online resource) <www.clir.org/pubs/reports/pub121/pub121.pdf> (accessed April 15, 2005).

Arrangement was also affected by how we ingested objects into DSpace because the method of ingest affected what metadata fields were included. Although manual metadata assignment of all files within the Joyce archive was laborious, certain metadata fields were impossible to record automatically. Content metadata, such as "subject" and "title of work," had to be entered by hand because no automatic tools were available to extract content accurately. Eventually, the practice of entering subject metadata on an item level was abandoned and replaced by the assumption that arrangement into series and available subject metadata for the whole archive would address the needs of most users. It was difficult to use file names as titles because they were not specific or standardized; however, we found no other solutions for creating titles for files except by manual entry or automatic extraction of file name.

Not all digital fonds require such high levels of description that demand manual manipulation of metadata. Some smaller archives with shallow or no hierarchical organization, or archives with few digital objects or few one-to-one relationships between digital and analog materials could be arranged at a lower level of description. Less robust description equates to limited discovery, but for some archives that may suffice. For such archives, automated ingest and metadata assignment may speed the time spent processing digital objects.

We faced additional limitations for precise metadata due to the metadata standard used by DSpace and by the ingest form provided with the graphical user interface (GUI). Unfortunately, not all metadata recorded for individual digital objects were included in the Qualified Dublin Core (QDC) metadata wrapper supplied in DSpace for each object during ingest and in the item display. We recorded some data, like directory hierarchies and original path names, in a spreadsheet created by the shareware tool, CatFinder 3.0. We then ingested the spreadsheet into a DSpace collection called Project Documentation. We also ingested with records of our arrangement process for the Joyce Papers because there was no metadata field offered for path names during the GUI ingest. Using the bulk ingest method, which occurs at the command line, we added a QDC metadata element "description.uri" to the dublin_core.xml file to record the path name of the ingested object, although slightly different from the original path name after arrangement of the files.

Fortunately, DSpace version 1.4 allows the addition of other metadata elements from defined metadata schemas, but the web interface is designed to accept and record QDC only. Unfortunately, the DSpace version running on the School of Information server is DSpace 1.2. To address the limitations of QDC, we are uploading an additional metadata file for each item from the second accession created using a metadata harvesting tool developed by National Library of New Zealand which uses their metadata schema. Additionally, use of other metadata schemas within DSpace are the subject of ongoing research at the University of Texas at Austin's School of Information.

Duplicate files within the archive raised additional issues for arrangement. Michael Joyce often maintained the same file on all three of his hard drives. He created backups of important files in case of hardware failure on his laptop, home and office computers and made duplicate copies in order to work on the same file from different locations. While using the software zsCompare (a comparison and synchronization utility from Zizasoft) to find duplicate files we noticed a trend: files with the exact same content had creation and modification dates that were exactly twenty-three hours and three minutes apart. We attributed the differences in timestamps to an improperly set internal clock in one of Joyce's computers. After noticing a fair amount of duplicate files we had to make an appraisal decision: were we going to keep every file accessioned with the Michael Joyce Papers, or could some of the copies be discarded? Because we created a file catalog for each disk using the software CatFinder 3.0, we decided to note that duplicates existed, save them to a separate directory on the hard drive of the Macintosh computer used for processing the files, but not to migrate all copies to DSpace. Although weeding through the duplicate files was time consuming, recording the metadata for the additional files would have been even more so considering some of the preservation tasks for each file that needed to be performed by hand.

Although DSpace is best suited to uploading individual items into the repository, a number of file associations within directories needed to be maintained. Some hypertext works within the archive are composed of multiple HTML files linked with hyperlinks and maintained in one directory. Because hypertext is based on internal links and because those links are often demarcated by a local file path, retaining a hierarchical

relationship is key to a functional product for download from DSpace. Maintaining directory relationships requires files to be ingested into DSpace as a bundle of files composing one item or as items ingested within the same collection. Both methods of retaining relationships between files require additional steps in the ingest process but are necessary for retaining relationships between some files.

We adopted methods for traditional archival arrangement and strategies for on-demand item-level arrangement while processing digital objects within the Michael Joyce Papers. Together, both methods allow users to browse records according to functional series and create new arrangements based on item-level metadata available for individual objects.

PRESERVATION BEYOND THE BITSTREAM

Digital preservation of the hypertext works in our case study raised unique preservation concerns beyond the preservation of bitstream copies. In addition to concerns for migration, authenticity, storage, and use similar to those for other born-digital objects, hypertext works require dynamic links and guard fields (words within the text that enable dynamic links), which create new issues for digital preservation. As described by ELO, preservation "solutions (for example, The Text Encoding Initiative's TEI schema or the library METS metadata standard) are often better suited for print, or print-like static works that have been digitized than for born-digital artifacts of electronic literature with dynamic, interactive, or networked behaviors and other experimental features . . . ."[7] ELO's solution for preservation is the X-Literature initiative, which has two parts: creating emulators and interpreters that enable the experience of digital works in a simulation of their native environment and developing a schema for electronic literature that can preserve unique aspects of hypertext, like links and guard fields, otherwise missing from current metadata standards.

Emulators and interpreters would address concerns for the preservation of Storyspace and Hypercard records in our case study by recreating the software and hardware environments in which the hypertext work was written. Currently, Storyspace (partially written by Michael Joyce) only runs on Windows or

[7] ELO, *Born-Again Bits,* 3.

Macintosh operating systems, but the same program does not run on both nor does a file written in Storyspace 1.5 run properly in Storyspace 2.5. The most current version of the software runs on Windows XP and Macintosh OS X. Storyspace is not open-source software, but the Ransom Center holds a copy of the source code. Copyright concerns, continued distribution of Storyspace by Eastgate Systems, and a lack of programming staff and time have prevented any steps towards creating emulating software to run Storyspace documents on the next iteration of operating systems. Hypercard files, created by proprietary Macintosh software and no longer supported, are also present within the archive. We welcome collaboration with other institutions and organizations, like ELO, willing to focus on creating ways to access the files we are preserving in DSpace.

Other preservation issues concern how scholars will want to research hypertext works in the future. Some users may want to experience hypertext in an original format and will need emulators. Other users might be interested more in the content of hypertext works and will be satisfied with XML records of works. Still other users may be interested in the various layers of hypertext as it appeared on original storage media and would need disk images to analyze the works. Scholars interested in hypertext works archived at the Ransom Center will most likely have sophisticated technological skills and may want to employ methods of literary analysis that involve other types of technology. As archivists, it is impossible for us to predict how scholars will want to use digital objects. Instead, we must strive for a utilitarian approach to digital preservation. We must address how most users will want to access our digital objects and preserve as much metadata as possible to facilitate scholarly use.

CONCLUSIONS AND RECOMMENDATIONS

Processing both accessions of the Michael Joyce Papers helped us draw conclusions about digital archivy that can be summed up in the following recommendations.

Automated, open-source tools are essential for future digital preservation projects.

Whether items are ingested manually or automatically, comprehensive open-source disk utilities need to be created to streamline the digital archeology portion of digital preserva-

tion. One integrated tool should check for viruses, recover files, create file catalogs, and preserve item authenticity by creating MD5 hashes. Tools for arrangement and ingest are desperately needed as well. Initiatives for automated record processing and ingest are developing but usable tools are absent.[8] Wherever possible, processes that were performed separately in our case study should be integrated into one tool. Accurate content analysis and comparison tools should be developed and integrated into digital processing tools as well.

Although we recommend more open-source software, we realize a higher level of specialized staff will be needed to find, download, install, manipulate, and use open-source software as compared to off-the-shelf software with built-in help functions, graphical installation interfaces, and technical assistance helplines. With this in mind, we offer a second recommendation.

Digital preservation will require specialized knowledge and specialized staff.

Archives will have to employ specialized staff with experience in information technology. Digital preservation requires knowledge of hardware, software, file formats, systems, servers, programming languages, metadata schemas and standards, Web applications, databases, and other specialized knowledge that most archivists do not have. At a time when archives are suffering from severe budget cutbacks, creative approaches to employing specialized staff will have to be considered. Archives may be able to fill these openings with hybrid positions, as grant-funded employees, or with shared workers between consortiums and/or collective agencies.

Methods of archival processing, arrangement, and description should adapt to handle issues presented by electronic records.

Archival theory and practice will need to change in response to the presence of electronic records archives that individuals are producing right now. Methods for processing electronic records archives will depend on cost, staff time and knowledge required, users' needs, tools available, institutional

[8] Manuscripts and Archives, Yale University Library, and the Digital Collections and Archives, Tufts University, *Fedora and The Preservation of University Records*. (Medford, MA: Tufts University, 2006) (online resource) <http://dca.tufts.edu/features/nhprc/index.html> (accessed October 17, 2006).

repository, hardware availability, and status of collection and may rapidly change as the number and size of digital archives grow. Archivists will need to be even more flexible and creative in their methods of processing materials in the future.

Before starting a digital preservation project, clear policies and procedures must be determined.

The policies and procedures for any digital preservation project require a permanent commitment by the preserving institution to manage, maintain, and migrate digital content. Without an institutional commitment, files can be neglected and eventually lost, which negates the purpose of preservation. Policies and procedures must clearly define how digital objects will be recovered, processed, ingested, and preserved to prevent duplication of work or improperly ingested digital objects.

This case study in digital archivy addresses some procedures for preservation of electronic literary archives at the Ransom Center. Although our methods for preservation will undoubtedly change in the future, we feel time invested now to create policies and procedures for preserving digital objects will decrease the effort spent to resuscitate older electronic objects later when it may be too late.

Catherine Stollar Peters is an archivist specializing in electronic records preservation at the Harry Ransom Center in Austin, Texas. She earned her BA and MS in Information Studies from the University of Texas at Austin.