Guarding the Guards:
Archiving the Electronic Records of Hypertext Author Michael Joyce

Catherine Stollar
Archivist
Harry Ransom Center
Austin, TX
cstollar@mail.utexas.edu

Thomas Kiehne
Web Applications Developer
Fosforus, Inc.
Seattle, WA
thomas.peter@kiehnefamily.us

ABSTRACT

In 2005, the Harry Ransom Center at the University at Austin acquired the *fonds* of hypertext author Michael Joyce. The major emphasis of the Ransom Center's collections is the study of literature and culture in the late 20th and early 21st century of the United States, Great Britain, and France. Michael Joyce's groundbreaking work in hypertext poetry and fiction make his papers a desirable addition to the Ransom Center holdings.

The Michael Joyce Papers are mostly composed of electronic records with an additional 60 manuscript boxes of paper -based materials This is the first mostly electronic archive the Ransom Center has acquired and new strategies for preserving digital content were employed. This case study discusses the techniques and skills utilized to preserve the electronic records of Michael Joyce as a model for processing future digital manuscripts at the Ransom Center.

SCENARIO

Established in 1957 by University of Texas Vice President and Provost Harry Huntt Ransom, the Harry Ransom Humanities Research Center at The University of Texas at Austin incorporated a strategy for collecting older rare books and manuscript collections with a new initiative to collect literary, photographic, and theatrical works by modern artists. Some of the authors whose works are included in the Ransom Center's collections are Norman Bel Geddes, Don DeLillo, T.S. Eliot, James Joyce, Ernest Hemingway, Norman Mailer, D.H. Lawrence, Ezra Pound, Anne Sexton, Isaac Bashevis Singer, and Tennessee Williams. Michael Joyce's work as perhaps the most influential hypertext poet and author fits nicely into the Ransom Center's contemporary author collecting policy.

Our case study to preserve Michael Joyce's digital manuscripts resulted from collaboration between the School of Information at the University of Texas at Austin and the Harry Ransom Center. Three students, Thomas Kiehne, Vivian Spoliansky, and Catherine Stollar, from Dr. Patricia Galloway's Problems in Permanent Retention of Electronic Records course offered at the School of Information undertook a semester long project to develop a strategy for archiving an initial accession of electronic materials saved on 371 3.5" floppy disks (totaling 211 KB) from author Michael Joyce. Upon completion of the project, a second accession of electronic and paper-based materials, including the contents of three hard drives (totaling 8.38 GB) and 60 manuscript boxes, was acquired by the Harry Ransom Center and is currently being processed by staff archivist Catherine Stollar according to the strategy developed during the class project. Our case study discusses strategies for file recovery, migration, preservation, arrangement, and description developed working with both accessions of Joyce's materials. The electronic records are currently maintained in a DSpace repository administered by the School of Information, however, in the future the Joyce records will move to a DSpace repository controlled by the Ransom Center and the General Libraries of the University of Texas.

The lack of Ransom Center staff with skills in digital archivy provided the impetus for the Ransom Center to partner with the School of Information on the Michael Joyce Papers. Although the Ransom Center employs talented archivists and IT professionals, no staff member possessed skills necessary for archiving digital manuscripts. The Ransom Center sought advice from Professor Galloway and agreed to use the Joyce materials as a case study in the Problems in Permanent Retention of Electronic Records course.

Some audio and video migration preservation projects were already in progress at the Ransom center in the Department of Photography and Visual Collections to preserve audio and video works, but there were no concerted efforts to preserve born digital manuscripts. Policies and Procedures for migration of audio and video content to new media were unsuited for born digital manuscript preservation and policies for preserving digital manuscripts were inadequate to capture the complete behavior of the original digital record. Previously, the few electronic manuscripts and correspondence already in the Ransom Center's manuscript collections were printed and organized in boxes like paper records. Because digital records are entirely unlike paper-based records, a preservation strategy based in printing records preserved very little of the original document. Electronic media were saved, but researchers were prevented from viewing original disks and no access copies were created.

The main component of our preservation strategy is to ingest electronic records and associated metadata into an institutional repository. DSpace, created from a joint project between MIT and Hewlett-Packard, is the institutional repository we used and will continue to use for electronic record preservation. At the heart of DSpace, like most open archival information systems (OAIS), is a database populated by individual digital objects supported by content, context, and structure metadata. We used DSpace, instead of FEDORA or another institutional repository, because it was already the established as the repository of choice for the School of Information. Although we had issues with the web user-interface for ingesting, viewing, and accessing materials within the repository, we plan to work with a talented ISchool student with Java programming skills to make our installation of DSpace more user-friendly.

Partnerships were key components to our case study's productivity and success. The initial group of students participating in the first part of the case study each represented a different background. Thomas Kiehne brought a wealth of Information Technology skills to the project, including programming and operating systems knowledge. Catherine Stollar shared her knowledge of archival theory and practice during the case study. Vivian Spoliansky viewed the case study through the lens of preservation and shed light on aspects of authenticity and desired levels of service for object preservation. Working with a variety of subject specialists on the project enabled participants to learn key skills from the others that will be useful on future digital record preservation projects.

Processing as Digital Archeology

One of the more unique aspects of this project involved the processing of 371 3.5” floppy discs that contained the digital objects of the first accession. The provided floppy discs were mostly from the Macintosh “classic” era, some of which date as far back as the mid to late 1980s. The assumption at this stage is that the original storage media is not stable or reliable and the information that they hold must be moved quickly and efficiently. Otherwise, little was known about what to expect in terms of specific technological issues or challenges.

At the outset of the project, we had only a general idea of the process of moving the digital files from the source media to a repository, and as such, we could not express specific requirements for software tools and utilities that might be needed. In order to minimize project overhead in terms of time and resources, we desired to use only open source, shareware, or freeware tools that are readily available in order to assist with the extraction process. This approach allowed us to assess the suitability of tools that are currently available and their ability to interoperate. In the absence of suitable free tools, we intended to find commercial software or create our own programs or scripts to perform the required tasks as we identified them. In the course of processing the first accession of discs, we quickly elucidated a more detailed procedural framework that can be abstracted and applied to future projects.

The general process implemented during the processing of the discs is as follows:

1. Receive and identify physical media
2. Catalog the physical media
3. Copy files to newer physical media
4. Perform initial file processing
5. Create an item-level index of all recovered files
6. Create and process working copies of all files while retaining the original bitstream copies

Technical metadata is collected at each step in the process not only to facilitate the work in progress, but to support provenance and authenticity. Each operation performed on the bitstream – every copy and access – provides the opportunity for inadvertent loss or alteration, so careful recordkeeping is as essential as careful handling. Additionally, all personnel involved in processing must thoroughly understand the procedures involved in order to prevent duplication of effort or discontinuities in results. In many cases, software can automate these processes, thus reducing the chance of errors, but the extent to which software can mitigate such risks is limited by the assumptions made by the creators of the software and how well the personnel making use of this software understand these limitations.

Given that time was of the essence, we opted to use text entries in Microsoft Excel spreadsheets to create the initial disc catalog and the associated metadata. This approach allowed us to leverage existing proficiency with spreadsheets and the availability of the software to eliminate the time needed to create a custom database application or to learn project management software. Unfortunately, the conspicuous absence of relational or workflow aspects in the spreadsheet format made us vulnerable to recordkeeping errors, making quality control a primary concern.

The copy functionality of the computer operating systems involved were sufficient to perform the movement of digital files from floppies to hard rives and removable media. Unfortunately, the differences between Macintosh and Windows in the management of file system metadata became significant. Creation dates are handled differently between these two operating systems such that a copy made in Windows takes on the date of the copy operation, not the creation date of the original from which it was made. Additionally, file system metadata for Macintosh files are stored as separate, invisible resource forks that are notorious for becoming corrupted. As a result, we often could not trust the dates ascribed by the operating system and had to refer to external resources, such as Michael Joyce's curriculum vitae, to confirm or provide date metadata at a later time. Issues with Macintosh resource forks also affected file downloads from DSpace after ingest.

At many points during the processing, we encountered technical difficulties in the form of file or disc errors. These errors can occur for a number of reasons, including damaged media, exposure to magnetic or other hazards, dirty data surface areas, and so on. In the case of dirty surface areas, several attempts were needed to overcome a copy error. It is suggested to have a drive cleaning kit available and use it periodically to prevent build up of debris on the drive head. For other errors, it was necessary to have available software utilities that can attempt to recover from file copying errors. Windows provides such capabilities within the operating system (e.g.: Scandisk), but Macintosh does not. For our purposes we were able to discover an older version of a commercial program, Norton Utilities, which allowed us to recover many files that could not be copied initially. Virus checking was also a preeminent concern. Errors and crashes must be met with persistence as they are often surmountable, which implies at least a minimum degree of technical knowledge.

In moving the digital files to other media, we created a filesystem hierarchy that mimicked the physical arrangement of the discs. Such hierarchical arrangement allowed us to use file system tools to generate some of the metadata automatically. There are a number of freeware, shareware, and commercial applications for Macintosh that will catalog a file volume and produce reports. We used a shareware utility called CatFinder[1] to index the copied files and export a report to a delimited format that was imported into Excel. This report formed the basis of our item-level metadata, including fields for filename, file size, kind (document or folder), Macintosh file type (analogous to the Windows file extension), Macintosh creator code, creation date, and modification date. To this basic report we added a comments field for use during appraisal and to collect technical notes.

MD5 file hashes were also generated for each file. Having an MD5 hash for each file allowed us to do two important things: to identify and/or eliminate redundant files, and to support provenance auditing during the repository ingest process. A freeware PERL application called Integrity[2] automatically created MD5 hash calculations and exported the results to a delimited text file. Unfortunately, integrating the MD5 hashes into the CatFinder index was not trivial due to differences between the two applications in file name recursion and handling of hidden files.

[1]http://www.mindspring.com/~shdtree/newsite/id9.html
[2]http://therockquarry.com/integrity.htm

Having created a unified index of filesystem metadata, augmented with processing notes and MD5 hashes, we were able to more accurately assess the extent of the digital files and facilitate arrangement and appraisal. Unfortunately, the index was in no way tied to the digital files and presented us with a significant information management problem. For example, any movement of files was not automatically noted in the index, nor was any change or deletion in the index reflected in the filesystem. We can envision a workflow-oriented system that stands between the filesystem and a metadata database that would greatly increase the speed and reliability of processing large bodies of digital documents.

Arrangement[3]

After recovering most of the unique digital content from the first accession of 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 newer 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? 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.

A number of digital materials, including emails and published articles, within the archive had a paper-based counterpart, demonstrating that Michael Joyce created both digital and analog records while performing the same activities. Both formats of records were created synchronously, and at an institution like the Ransom Center that preserves not only works that have influenced the arts and humanities fields, but also preserves the context in which those works were created, we determined it would be desirable to reflect synchronous creation in the arrangement. We did not originally understand relationships between Joyce's digital and paper materials because our first portion of the case study only dealt with electronic records from the first accession of floppy disks. We initially arranged the files into 5 series: Works, Academic Career, Correspondence, Storyspace, Third-party Works, and Personal. After surveying the paper-based materials and the second accession of electronic materials, we had to alter our originally arrangement to include the newly accessioned materials. The final arrangement we created is Works and Related Materials, Academic Career, Correspondence, Storyspace, Journals and Appointment Books, Personal, Works by Other Authors, and Published Materials.

Institutional repositories like DSpace can facilitate digital object arrangement into our specified series by using the community, sub-community, collection, sub-collection, and item level hierarchies. DSpace's hierarchies relate to traditional archival hierarchical levels: communities equate to archival *fonds*, sub-communities to series and sub-series, collections as 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 or related files, i.e. websites with multiple linked HTML files, can be ingested as bundled files.

After determining how to arrange the paper and digital materials, we decided how to arrange the physical housing (jewel cases, magnetic media, paper holders, plastic cases, etc.) from Joyce's electronic works. 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 and to limit access to the electronic materials by researchers. No studies on electronic media degradation have found any instances of off-gassing or other damaging effects of filing electronic media with paper-based materials, so we determined physically integrating paper-based material and digital media would be the best policy for physically arranging the Michael Joyce Papers. The Ransom Center will still limit access to files saved on original media because researchers will have access to the files via DSpace.

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, non-linear nature of digital object arrangement by enabling on-demand, user-controlled arrangement by item-level metadata. Preservation of digital objects depends on item-level

[3] More detail about our project can be found in a forthcoming article about processing the Michael Joyce Papers in *Provenance*.

metadata used to document, migrate, emulate, and preserve the objects. Item-level metadata recorded for preservation in DSpace's database also enables flexible arrangement of digital objects. 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, or other metadata element. Arrangement is limited only by the skills of the programmer developing the user interface used to access the OAIS database and the precision of metadata recorded for each object.

Arrangement is also affected by how we ingested objects into DSpace because our method of ingest affected what metadata fields we 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 automatic tools to accurately extract content were not available.[4] We found it difficult to use file names within the archive to associate files with published titles because the file names were not specific or standardized.

We incorporated 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 any metadata available for individual objects.

Challenges

In addition to the challenges we encountered developing a strategy to preserve Joyce's text and graphic files, we faced unique challenges associated with preserving Joyce's most influential creative works written using specialized software called Storyspace. Storyspace, created by Michael Joyce, Jay David Bolter, and John B. Smith, as a format type presented (and continues to present) challenges for media migration, ingest, and file use. Hypertext works written in Storyspace are composed of multi-faceted texts linked by guard fields (words within texts that enable direct links to other nodes, usually under specific conditions) and can only be viewed using Storyspace software. To complicate matters, we originally thought the latest version of Storyspace was backwards compatible and could read works written in the first version of Storyspace. Unfortunately, this is not entirely the case as older Storyspace documents do not degrade gracefully. For example, the text from files written in Storyspace 1.5 can be read in Storyspace 2, but the individual nodes and links are missing, making the Storyspace 2 rendering of a older work vastly different from the original.

NEW SKILLS

- A thorough grounding in the various operating systems. The profusion of technical difficulties and operating system inconsistencies required an intuition about the various platforms that can only be gained by direct experience. While processing digital files it is essential to have an understanding of the environment in which they were created. Computer literacy in more than one platform and with networked environments will be ideal traits for archivists of the future.
- A basic understanding of the structure of digital documents. Knowing how a digital file is created and stored, including such basics as the difference between binary file formats and textual formats such as ACSII and UTF helps provide an understanding of what happens during processing. Furthermore, an intimate knowledge of the types of formats and how they might be identified (e.g.: file type extensions or creator codes), accessed, and converted is essential. Understanding digital formats offers clues to where to

[4] Literary text comparison tools designed for use with small numbers of digital works were not sufficient for our large collection of files. Apparently text-mining tools could serve our purposes to compare large bodies of records with each other. We have not utilized any text-mining tools to date.

find item-level metadata (e.g.: document properties embedded in word processing files, ID3 tags embedded in MP3 audio files, etc.) and suggests migration paths for long-term preservation.

- Proficiency with and trust in new tools. Familiar means of handling physical documents are not present with digital documents. Software tools and operating systems augment the functions of our senses in the digital world and mitigate some of this loss, but not completely. Integrated toolkits and processing systems are needed and must be developed so that they can be trusted to conform to the expectations of archival practice.
- Establishment of new workflows and procedures. The intangible nature of digital information makes documentary evidence crucial to processing. Many institutions have established procedures for document processing, including audio/visual materials, but these cannot be assumed to be sufficient for digital objects. Operating systems alone cannot document processes, so new systems that function according to sound processing policies are necessary.
- Ability to monitor current trends in digital preservation including metadata standards, crosswalks between encoding standards, available tools, storage systems, file format repositories, national and international research initiatives, user expectations, and published best practices guides.
- A thorough understanding of traditional archival theory and practice. Archivists who work with digital records should be able to extrapolate traditional theory and apply it to electronic record preservation, but must be flexible enough to create new standards for archival practice. What we do as archivists will change (practice), but why we do it will not (theory).

DISCUSSION QUESTIONS

- Would automatic content management be more time consuming or less time consuming that manually arranging the digital manuscripts? Would it result in a better arrangement?
- Is it feasible to devise a one-size fits all processing toolkit?
- How can authors, who may deposit their materials in an institution like the Ransom Center, implement a digital preservation strategy at home, closer to the point of document creation?
- How desirable is it to keep most files in proprietary formats that are the current de facto standard? (i.e. Microsoft Word, Adobe PDF, etc.)
- Should files be arranged at all or should they be indexed and sorted using search engines using item-level metadata?
- Is DSpace a viable option for smaller repositories and organizations?
- How will DSpace integrate into existing points of access? (i.e. OPACs, website, EAD consortium sites)
- How do archivists best obtain the skills we are advocating they have? Classes? Projects? Workshops? Conferences?