Educating for Digital Archiving: Digital Forensics

Patricia Galloway
School of Information
University of Texas at Austin

Digital Archiving education at UT

- Philosophy
  - Digital archivists must know their materials and how they are created, valued, and used
  - Where materials are digital, archivists must understand and use both the technology used to create the materials and that used to preserve and access them
- Teaching goals
  - Create a suite of courses to provide students with the skills to work independently with digital objects
  - This means that students must learn how to solve novel problems within a framework designed to protect original materials against damage
- Infrastructure goals
  - Build a repository and ancillary infrastructure (laboratory space, equipment) where both experimentation and real work can be done

2000-2002

- Course creation
  - 2000: Introduction to digital records
  - 2001: Digital Archiving
  - 2001: Lifecycle Metadata for Digital Objects
- Beginning infrastructure
  - First used LAMP (Linux, Apache, MySql, php)
  - Emergence of DSpace winter 2002-3
- Early experiments: email, text processing

2003-2004

- DSpace implemented, spring 2003
- Digital Archiving course, 2003 (4 students)
  - Website archive planning
  - Publication of infrastructure plan
  - Planning documents ingested
- Digital Archiving course, 2004 (6 students)
  - AAA Anthrosource (ASSC) experiments
  - Development of plan for archiving of AAA publications (eventual adoption by AAA of Portico)
  - Planning documents ingested

2005: Departmental IR, incubator for HRC

- Digital Archiving course (19 students)
  - Four faculty members: Davis, Dillon, Hallmark, Wyllys
    - Publications, Syllabi, Webpages, Other learning objects
  - School of Information website
    - Experiment with crawling (Heretrix)
  - Michael Joyce collection, HRC
    - Protocol for processing legacy formats
- Age of media: current
- Copying procedure
  - Write block, Virus scan, Copy, Message digest
- Students in class
  - Catherine Stollar [now Peters] (Joyce project)
  - Zach Vowell (Davis project)

2006

- Digital Archiving course (14 students)
  - ASSC: email attachments
  - Cochineal: multimedia documents
  - School of Information tutorials: complex objects
  - School of Information website: apply records schedule
- Age of media: current
- Copying procedure
  - Write block, Virus scan, Copy, Message digest
- Metadata course, 2006
  - Developed METS SIP profiles for DSpace object types
- Student in Introduction class
  - Gabriela Redwine (beginning of Mailer digital issues)

2007

- Digital Archiving course (18 students)
  - Wesker: naming conventions, privacy
  - Uris and Crowley: corrupt and unreadable files
  - Dillon redux: copyright
  - School of Information commencement videos: originals and derivatives
  - School of Information tutorials: complex objects
  - ANAGPIC publications: digitization
- Age of media: current
- Copying procedure
  - Write block, Virus scan, Copy, Message digest

2008

- Digital Archiving course (13 students)
  - Gracy: range of learning objects
  - Mailer: extracting from Iomega OneStep
  - George Sanger Videogame sound: IP issues, music formats
- Age of media: (relatively) current
- Copying procedure
  - Write block, Virus scan, Copy, Message digest

2009

- Digital Archiving course (23 students)
  - Paul Banks digital materials (3.5”, Zip)
  - Warren Spector email (Apple DVD-R)
  - Warren Spector design documents (Apple and Kaypro 5.25”)
  - George Sanger games: ATF, Putt-Putt (3.5”, Zip, DAT, SVHS)
  - Heather Kelley, Redbeard’s Pirate Quest (Jaz)
  - Terrence McNally (3.5”, CD-ROM)
- Age of media: Current, legacy
- Copying procedure
  - Write block, Virus scan
  - In 5.25” cases: Disk image, clone extraction (using original hardware at Goodwill Computer Museum)
  - Message digest

2010

- Digital Archiving course (24 students)
  - George Sanger email (DVD)
  - George Sanger ADAT (SVHS)
  - 1988 campaign interviews (3.5”)
  - Faculty members Lukenbill, Davis (Zip, USB key, 5.25”)
  - Tutorials (available on server)
  - Maya database materials (5.25”, 3.5”)
- GCM collaboration (Frankenstein I)
- Copying procedure (offline)
  - Write block
  - Media image (dd), message digest
  - Replicate image, clone extraction, message digests
  - Tar directory tree content where relevant

Frankenstein I

Current Interests: Technology

- Creating digital archaeology lab in new School of Information space (done in 2010)
- Establish collecting strategy for legacy drives, software, media, collaborating with IT staff and GCM
- Acquire, configure, and begin using forensic workstation for current media (shipped yesterday)
- Collaborate with Goodwill Computer Museum to build Frankenstein II
  - Processor array
  - RAID array for multiple OSs and workspaces
- Note that new requirements arise in the process of classwork-as-discovery; our experience shows that every collection is unique in the technological details, but students can build on what previous students did, hence reports are archived.

Research challenges

- Adequately informing creator of preservation and keeping procedures as currently understood
  - Need for understandable donor agreement including informed consent for preservation procedures
- Alternative collection presentations
  - Preserving a record of potential directory structures to preserve original represented order as potentially seen by creator (OS affordances)
  - Mirroring original order in some representation of collection
  - Create “archival order” articulating collection for use
  - Inviting guest scholar/curators to author different virtual orders
- Capture and representation of creation environment and process
  - Creator technology adoption and use histories
  - Arrangement and rearrangement of digital materials over time: versioning practices
  - Implications of (e.g.) desktop design, operating system for practice
- Creator behavior in providing collections to repository
  - How and why do they repackage?
  - How to recover creator actions after intervention of third parties (“dealer “grooming” of collections)?
  - How do archivists figure as third parties here?
  - Will secure online transfer, self-archiving by creators, and adoption of MPLP processing practices affect this problem?

Perspectives

- Digital materials are mediated by a complex sociotechnical infrastructure
- It will never be possible to know that infrastructure completely even as it applies to only one person
- We can, however, know what patterns of use practice manifest themselves in digital corpora and even partial corpora—if corpora are obtained and permission to investigate them thoroughly is secured
- Documentation and/or representation of these practices is important to support 21st century research into the “textuality” of the digital