Preparing Future Digital Curators

PART I: A Summary Report on the Digital Curation Curriculum (DigCCurr) Project
Christopher (Cal) Lee and Carolyn Hank

PART II: Four Perspectives on Applying Academic Understanding in a Practice Setting
John Blythe, Lisa Gregory, Samantha Guss, and Jennifer Mantooth

School of Information and Library Science
University of North Carolina at Chapel Hill
http://www.ils.unc.edu/digccurr/

LAUNC-CH Research Forum, Chapel Hill, NC: 21 May 2008
PART I:

A Summary Report on the Digital Curation Curriculum (DigCCurr) Project
Professional Evolution

• Advances in management, preservation & dissemination of digital resources

• Many streams of activity (e.g. computer scientists, archivists, records managers, librarians, scientific data engineers, museum curators, organizational IT staff)

• Increasing recognition in past decade of common challenges & opportunities

• Recent adoption of term “digital curation” – pulling together many previously distinct research communities
Digital Curation

• “The active management and preservation of digital resources over the life-cycle of scholarly and scientific interest, and over time for current and future generations of users.”

• Widely used by scientists & those responsible for data sets

• Seen by many as more inclusive – in disciplinary scope & coverage of lifecycle -- than “digital preservation”

*Digital Curation Centre. “What is Digital Curation?” http://www.dcc.ac.uk/about/what/
Education & Professional Development

• Many valuable components of a digital curation curriculum
  – Individual courses & components within graduate programs (most in LIS programs, but also e.g., computer science, business, public policy, history)
  – Professional workshops (usually 1-5 days)

• Training in specific disciplines generally doesn’t address issues such as long-term access, integrity, contextual information

• LIS students would benefit from more understanding of specific digital environments & resource types
DigCCurr Project

• IMLS Grant # RE-05-06-0044

• Collaboration of School of Information & Library Science (SILS), University of North Carolina at Chapel Hill (UNC-CH) & U.S. National Archives & Records Administration (NARA)

• Runs July 1, 2006 – June 30, 2009
DigCCurr Components/Goals

**Curriculum:** To prepare students for digital curation with a wide variety of organizations, contexts & types of resources:

- Graduate-level curricular framework
- Course modules
- Experiential components

**Two International Symposia:**

- First was held April 18-20, 2007 in Chapel Hill - [http://ils.unc.edu/digccurr2007/](http://ils.unc.edu/digccurr2007/)
- Second to take place early April 2009 (near end of project)

**Carolina Digital Curation Fellowship program**
Practical Field Experience

• Should engage in at least two different field experiences in different institutional contexts

• Should involve some hands-on work with digital objects with actual consequences, rather than just conceptual or policy work

• Importance of partnering with sites that already actively engage in digital curation
Carolina Digital Curation Fellows

- 5 Digital Curation Fellows pursuing degrees at SILS - began fall 2007
- UNC partners providing practical experience opportunities: ibiblio, ITS, Odum Institute, University Library
- Specialized introductory seminar held Fall 2007
- Overseeing & learning from their practical engagement work
- Advising on course selection
- Plan for future practical engagement opportunities
Matrix of Digital Curation Knowledge & Competencies

• Tool for thinking about, planning for, identifying & organizing curriculum

• Each unit of curriculum content can address one or more dimensions

• Helping to address fundamental issue: All digital curation students should get some aspects of curriculum, but other aspects only necessary for students planning to work in particular types of places or jobs (i.e. balancing core vs. specialized knowledge)
The Digital Curation Curriculum (DigCCurr) project (IMLS RE-05-06-0044) is developing a graduate-level curricular framework, course modules & experiential components to prepare students for digital curation professions in a variety of information environments. The two tables below summarize the contents of our two main conceptual products & tools for organizing our work. More detailed elaboration of the components & sources from which we’ve drawn will be available at the DigCCurr web site.

### Mandates, Values & Principles
- Trustworthiness
- Sustainability
- Innovative Reuse
- Accessibility & Interoperability
- Assessing & Managing of Risk
- Collection Care
- Content
- Context
- Integrity
- Longevity
- Appropriateness

### Legal Requirements
- NLM
- Copyright

### Functions & Skills

#### 3. Professional, Institutional or Organizational Context
- History of Professional Activities
- Roles and responsibilities of digital curators
- Legal and ethical issues for digital curators

#### 4. Type of Resource
- Level of aggregation
- General description
- Network infrastructure

#### 5. Prerequisite Knowledge
- Metadata
- Archival principles
- Technical skills

### Draft Matrix of Topics for Digital Curation Curriculum

#### 1. Objectives
- Digital Curation Overview
- Objectives & Principles
- Functions & Skills
- Mandates, Values & Principles

#### 2. Skills & Competencies
- Evaluation of Digital Collections
- Management of Digital Collections
- Preservation of Digital Collections

#### 3. Professional, Institutional, or Organizational Context
- History of Professional Activities
- Roles and Responsibilities of Digital Curators
- Legal and Ethical Issues for Digital Curators

#### 4. Type of Resource
- Level of Aggregation
- General Description
- Network Infrastructure

#### 5. Prerequisite Knowledge
- Metadata
- Archival Principles
- Technical Skills

### Draft High-Level Categories of Digital Curation Functions

#### 1. Overview of Digital Curatorial Responsibilities
- Planning of Digital Curatorial Activities
- Management of Digital Collections
- Preservation of Digital Collections

#### 2. Evaluation of Digital Collections
- Analysis of Digital Collections
- Comparison of Digital Collections
- Quality Assurance

#### 3. Management of Digital Collections
- Planning of Digital Collections
- Management of Digital Collections
- Preservation of Digital Collections

#### 4. Preservation of Digital Collections
- Planning of Digital Collections
- Management of Digital Collections
- Preservation of Digital Collections

### DigCCurr Research Activities:
- Extensive review & analysis of literature
- Interviews with 17 expert advisory board members
- Survey of participants in major international conference in April 2007
- Planned wider-scale survey of digital curation needs & expectations
- Analysis of syllabi
- Analysis of job postings

---

Association for Library and Information Science Education (ALISE) Annual Conference - January 8-11, 2008, Philadelphia, Pennsylvania
PART II

Four Perspectives on Applying Academic Understanding in a Practice Setting
Current Status of DocSouth CD Migration:
A Report from the Carolina Digital Library & Archives

John Blythe
Documenting the American South (DocSouth) is a digital publishing initiative that provides internet access to texts, images, and audio files related to southern history, literature, and culture. Currently DocSouth includes ten thematic collections of books, diaries, posters, artifacts, letters, oral history interviews, and songs.

- The Church in the Southern Black Community
- The Colonial and State Records of North Carolina
- The First Century of the First State University
- First-Person Narratives of the American South
- Library of Southern Literature
- North American Slave Narratives
- The North Carolina Experience
- North Carolinians and the Great War
- Oral Histories of the American South
- The Southern Homefront, 1861-1865
- True and Candid Compositions: The Lives and Writings of Antebellum Students at the University of North Carolina
FROM CD to DARK ARCHIVE

- MF Digital Scribe 9000 allows data extraction from two CDs simultaneously
- 1,642 CD’s
- 44,787 files
- 3,296 folders
- 903 GB
- Approx 100 hours
• *Discs 11-49b* = group of discs extracted during one session
• *Disc 16* = individual disc. Named manually
• *Uncle Johnson, the Pilgrim of Six Score Years* = name of folder on CD
• *unclejcv.jpg* = filename
Gustavus L. Foster (Gustavus Lemuel), 1818-1876
Uncle Johnson, the Pilgrim of Six Score Years.
Philadelphia: Presbyterian Publication Committee, 1867.
Image cataloged in DocSouth database

- Source publication and each image assigned \textit{Item\_ID}
- \textit{Item\_ID} for Uncle Johnson, the Pilgrim... = 39
- \textit{Item\_ID} for cover image = 53698
The Ideal: Programming Magic

• Script matches file name in dark archive with \textit{item\_id} in DocSouth database
• Script rebuilds dark archive directory to match database structure.
The Reality: Auto-matching is not perfect

- 16,847 file names matched and assigned appropriate item_id
- 1,257 file names matched more than one item_id
- 14,800 file names did not match
Why no match?

- *small.jpg* (a thumbnail) not in database
- Additional info added to end of file name (i.e. `-thumb, -1, -50,-75,-150, _100, _150, _at_75, _at_150`)
- Slight variations in file names used on website (database) and archived on CD (dark archive)
- Typos in file names - mistyped letter, extra space, extra letter (i.e. *circltp.jpg* becomes *circlrtp.jpg*)
- Images not in database
More examples

• Variance in one number between file name on website (database) and on CD (dark archive), i.e. *hicks23.jpg* in database is *hicks22.jpg* in dark archive

• Files have two different names: *unclecv.jpg* in dark archive is *fostecv.jpg* in database
Auto-matching: Take 2

- Establish Parent_ID by searching Item table for name of publication
- Using Parent_ID search Item_child for range of Item_IDS that belong to Parent
- Use Item_ID of children to search Illustration_item for specific file name associated with that ID.
The Results: 550 matches

The Reasons

• Variations in names assigned to folders in dark archive (I.e. Uncle Johnson, the Pilgrim...I) and names listed in Item table of database

• Slight variation in punctuation
Next Steps

• Strip punctuation from folder names in dark archive and names in sort_by field of Item table

• Reduce multiple Item_ID possibilities by creating auto process to evaluate Item_IDs based on their numerical proximity to each other.

• Hire a student
The Intersection of Image Needs and Expectations:
A Look from the Carolina Digital Library & Archive's Digital Production Center

Lisa Gregory
VIRGINIAE
Item et
FLORIDAE
Amricee Provinciarum, nova
DESCRIPTIO.
Metadata Issues at a Large Social Science Data Repository: A Snapshot of Digital Curator Needs in the Practice Setting

Samantha Guss
Data Archive Services

Data Catalog

Welcome to our new search engine.

The Odum Institute maintains one of the oldest and largest catalog of machine-readable data in the U.S. It has an extensive collection of U.S. Census data, including one of the most complete holdings for 1970 Census files. Other major sources of data include the North Carolina State Data Center, which distributes North Carolina census data; and the National Center for Health Statistics.

Public Opinion Poll Database

The Odum Institute’s Public Opinion Poll Database allows any researcher to search for specific poll questions among the more than 230,000 questions in the Institute’s archive by key words, date, study number, study title, or state.

NC Vital Statistics
Determining Archival Value in Orphaned WebSpaces: An Investigation within the ibiblio.org Domain

Jennifer Mantooth
Websites and Open source software
Sample View of Context-less Files
Re-Appraisal Questions

1. Does the website contain content?
2. Is the material accessible?
3. Is the material easy to negotiate?
   a. Are the file names meaningful?
   b. Do the files have value on their own?
Ongoing Issues

• Migrating abandoned database driven websites

• Alerting the users of defunct links/outdated sites
  – Many users don’t go through ibiblio’s collection index (Google)
Questions?

Thank you
http://ils.unc.edu/digccurr