DIGITAL AND DATA CURATION

Dr. Helen R. Tibbo,
Alumni Distinguished Professor
School of Information and Library Science
University of North Carolina at Chapel Hill
December 2014
THANK YOU!

- To Sam Oh for inviting me to speak today
- To SKKU for supporting this trip
- To all of you for being attentive through a long presentation!
TODAY’S AGENDA

- A bit about me
- A bit about data
- What are digital and data curation?
- Data curation models
- Why are digital/data curation important?
- Major curation challenges
- What Lies Ahead?
- Digital and data curation at SILS
MY BIO

- Dr. Tibbo is an Alumni Distinguished Professor at the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill (UNC-CH), and teaches in the areas of archives and records management, digital preservation and access, data curation, appraisal, and archival reference and outreach. She is also a Fellow of the Society of American Archivists (SAA) and was SAA President 2010-2011. She has directed numerous grant projects.
MY BABIES: NICK
MY BABIES: NORA
WHAT ARE DATA?

- Not just scientific data
- Data come in a wide variety of forms and formats
- DCC definitions:
  - **Data**: “any information in binary digital form”
  - **Databases**: “structured collections of records or data stored in a computer system.”
  - **Simple Digital Objects** “are discrete digital items; such as textual files, images or sound files, along with their related identifiers and metadata.”
  - **Complex Digital Objects** are “discrete digital objects, made by combining a number of other digital objects, such as Web sites.”

- Ross Harvey, p. 3 and DCC website
NSF’S DATA DEFINITION

• “Any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret the data.”

DATA AS DIGITAL HERITAGE

• UNESCO’s Guidelines for the Preservation of Digital Heritage (2003) includes:
  • Electronic publications
  • Semi-published materials
  • Organizational and personal records
  • Data sets
  • Learning objects
  • Software tools
  • Unique unpublished materials
  • Electronic “manuscripts”
  • Digitally generated artworks and documentary photographs

BORN DIGITAL VS. DIGITIZED

• Born Digital Materials
  • Originally created by a computer
  • Could have an analog equivalent (e.g., printed Word document) or not (e.g., database)

• Digitized materials are the result of a digitization process.

• Once digital, issues and curation are the same.
DATA COLLECTIONS

- **Research Collections:**
  - E.g., local data generated in a lab or research project

- **Community Collections:**
  - E.g., genome databases such as MGI-Mouse Genome Informatics
    - International; scientists contribute data from around the world
    - Scientists from around the world use the data in this collection

- **Reference Collections:**
  - E.g., Protein Data Bank; “information about experimentally-determined structures of proteins”
DATA VARIETY

- For example:
  - Observational
  - Simulations
  - Multimedia
  - Surveys
  - Performances

- Computational
  - Software
  - Visualizations
  - Web pages
  - Social media content
METADATA

- NSF definition:
  - Metadata “summarize data content, structure, interrelationships, and provenance.”
WHAT’S THE BIG DEAL?

• Today’s science, medicine, and commerce are creating massive amounts of data.
  • “Deluge,” “waves,” and “knowledge overload” are some of the terms used to describe the situation (Carlson & Anderson, 2007).
  • New science such as genomics and those that collect data via sensors have changed data types and amounts.
  • 4th Paradigm – data-driven science
  • Data grids and greater computing capacity
  • Massive analysis and federation of disparate datasets.
  • Funders want researchers to create sound data that can be re-used, thus saving funding.
LIFE STAGES OF DATA -1

• Carlson & Anderson:
  • Data Collection: Born Digital and Heterogeneous Legacy Data
    • Much data is still collected in analog form.
  • Data Formatting: Codified Abstract Forms and Tacit Knowledge
    • “For collected materials to become data that can be used and mainly reused, they need to be rendered disseminative, that is to be rendered at the same time transportable in concise abstract forms and intelligible.”
    • Data must be normalized, cleaned, have adequate metadata that labels and explains the data. (pp. 307)
LIFE STAGES OF DATA - 2

• Data Release: Ownership, Consent, and Moral Rights
  • IRB, ethical consideration in sharing data

• Data Re-Use: Trust, Provenance, and “Cookery”
  • Data re-users need to trust data creators
  • Methodology must be explicit and documented
LIFE STAGES OF DATA - 3

- “Across all case studies it was clear that this disconnection required not only visualizing data in intelligible forms, but more importantly, making explicit their context of production and setting up appropriate systems of quality checks and assessment. “ - Carlson & Anderson, p. 309-310.

- Best practices help to instill trust

- Research CONTEXT must be captured and articulated to users.
Carlson & Anderson concluded that “two key assumptions that appear to underpin a number of discourses on e-science are not supported in practice”:

- That knowledge can easily and straightforwardly be disembedded from its producers and original contexts to become explicit data for temporally and geographically distributed re-users.
- That there is a binary divide between the “quantitative” and “qualitative” sciences in their approach to, and ability to benefit from, e-science tools and practices, especially in terms of data re-use.
WHAT IS DIGITAL CURATION?
WE’VE MADE IT – BUT NO ONE KNOW WHAT WE DO!

Dilbert, Sunday, October 30, 2011
WHAT IS A DIGITAL CURATION?

First - what it is not:

- Not Digital Preservation (but preservation is part of curation)
- Digital Archiving (but archiving is part of curation and archiving principles are essential to successful curation)
- Not digitization
- Not a Digital Library
- Not records management
- Not a software system like a digital repository
WHAT DIGITAL CURATION IS:

- A cross disciplinary field of study
- An emerging profession in libraries, archives, and computing
- A long-term commitment to the ongoing maintenance of digital content
- A complex set of tasks and workflows across the digital life-cycle
“Digital curation involves maintaining and adding value to digital research data throughout its lifecycle.”

-DCC Website

http://www.dcc.ac.uk/digital-curation/what-digital-curation

We can substitute “digital content” for “research data”.
According to Beagrie, the term “digital curation” was first used at the "Digital Curation: digital archives, libraries and e-science seminar" sponsored by the Digital Preservation Coalition and the British National Space Centre held in London on the 19th October 2001.

Curation involves adding value – collection building, adding metadata, providing search mechanisms, etc.

“Digital Curation” transfers physical curatorial practices to digital materials.

“Digital curation” and subsequently “data curation” became associated with e-science in the UK and Europe although much support came from the idea of preserving cultural heritage.
WHAT IS DIGITAL CURATION?

- “maintaining and adding value to a trusted body of digital research data for current and future use; it encompasses the active management throughout the research lifecycle.
  - Digital Curation Center, 2010

- “actions needed to maintain digital research data and other digital materials over their entire life-cycle and over time for current and future generations of users…”
  - Giaretta, 2005; Joint Information Systems Committee, 2003; Beagrie, 2006
KEY POINTS OF DEFINITIONS

- Preserved for future use – future users must be able to open, perform, and understand preserved data.
- Entails the data lifecycle; from standards setting (as in file formats) and data creation through active storage, archiving, reuse, and ultimate disposition of data.
- Preservation over time.
KEY POINTS OF DEFINITIONS

- Active management – not just keeping paper in boxes or bits in storage; digital curation requires ongoing effort (and cost).
- Not just “digital archiving” or “digital preservation”
- Much of the digital curation workflow involves working with data creators and data users; technology is just a small part.
MANY STAKEHOLDERS

“The foundation of the DCC reflects the belief that long term stewardship of digital assets is the responsibility of everyone in the digital information value chain”.

Chris Rushbridge et al. (2005).
“Digital curation is the active involvement of information professionals in the management, including the preservation, of digital data for future use.”

Elizabeth Yakel (2007)
YAKEL - CORE CONCEPTS AND ACTIVITIES

- Life cycle/continuum management of the materials perhaps even reaching back to the creation of the record keeping system.
- Active involvement over time of both the records creators and potentially digital curators.
- Appraisal and selection of materials.
- Development and provision of access.
- Ensuring preservation (usability and accessibility) of the objects.

Elizabeth Yakel (2007)
**Latina** cūrō: *I arrange, see to, attend to, take care of, ensure; I heal, cure; I govern, command; I undertake, procure…. from ProtoIndoEuropean *kʷeis ("to heed").*


- “Digital curation is the active involvement in the management, including the preservation, of digital resources for future use.”
DCI POINTS

- Not just research data
- Not just trusted data
- Not just information professionals.
Short Definition

- Digital preservation combines policies, strategies and actions that ensure access to digital content over time.

http://www.pla.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.pdf
Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time.
DIGITAL PRESERVATION / CURATION

Long Definition

- Digital preservation combines policies, strategies and actions to ensure the accurate rendering of authenticated content over time, regardless of the challenges of media failure and technological change.
- Digital preservation applies to both born digital and reformatted content.
- Digital preservation policies document an organization’s commitment to preserve digital content for future use; specify file formats to be preserved and the level of preservation to be provided; and ensure compliance with standards and best practices for responsible stewardship of digital information.
- Digital preservation strategies and actions address content creation, integrity and maintenance.
LIFECYCLE MODELS

- A visual way to depict the flow, relationships, and transitions of major components of large systems.
  - C. Humphrey (2006)

- Lifecycles are path dependent.

- The number of individuals and institutions involved at each stage of the lifecycle increases as the complexity of the data increases.
  - J Wallis (2008)
LIFECYCLE MODELS

- Time constants of data lifecycles are becoming shorter and the diversity of stakeholders and complexities of data are increasing.

- Iwata (2008)
DDI GENERIC LIFECYCLE

Thomas, Gregory, & Piazza (2005)
GENERIC RESEARCH LIFECYCLE

(Green & Gutmann, 2007)
LC PRESERVATION LIFECYCLE

(Library of Congress, n.d.)
DCC DIGITAL CURATION LIFECYCLE

http://www.dcc.ac.uk/resources/curation-lifecycle-model
DATA, any information in binary digital form, is at the centre of the Curation Lifecycle. This includes:

- **Digital Objects**: simple digital objects (discrete digital items such as text files, image files or sound files, along with their related identifiers and metadata) or complex digital objects (discrete digital objects made by combining a number of other digital objects, such as websites).

- **Databases**: structured collections of records or data stored in a computer system.

See more at: [http://www.dcc.ac.uk/resources/curation-lifecycle-model#sthash.xir54EJL.dpuf](http://www.dcc.ac.uk/resources/curation-lifecycle-model#sthash.xir54EJL.dpuf)
Description and Representation Information
Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand and render both the digital material and the associated metadata.

See more at: http://www.dcc.ac.uk/resources/curation-lifecycle-model#sthash.xir54EJL.dpuf
**DCC DIGITAL CURATION LIFECYCLE: FULL LIFECYCLE ACTIONS**

- **Preservation Planning**
  Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions.

- **Community Watch and Participation**
  Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software.

  - See more at: [http://www.dcc.ac.uk/resources/curation-lifecycle-model#sthash.xir54EJL.dpuf](http://www.dcc.ac.uk/resources/curation-lifecycle-model#sthash.xir54EJL.dpuf)
Curate and Preserve
Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle.

See more at: http://www.dcc.ac.uk/resources/curation-lifecycle-model#sthash.xir54EJL.dpuf
DCC DIGITAL CURATION LIFECYCLE: SEQUENTIAL ACTIONS

- **Conceptualise**
  Conceive and plan the creation of data, including capture method and storage options.

- **Create or Receive**
  Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation.
  - Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata.

- **Appraise and Select**
  - Evaluate digital objects and select those requiring long-term curation and preservation. Adhere to documented guidance, policies and legal requirements.
DCC DIGITAL CURATION LIFECYCLE: SEQUENTIAL ACTIONS

- **Ingest**
  Transfer digital objects to an archive, trusted digital repository, data centre or similar, again adhering to documented guidance, policies and legal requirements.

- **Preservation Action**
  Undertake actions to ensure long-term preservation and retention of the authoritative nature of data. Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity. Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.
DCC DIGITAL CURATION LIFECYCLE: SEQUENTIAL ACTIONS

- **Store**
  Keep the data in a secure manner as outlined by relevant standards.

- **Access, Use and Reuse**
  Ensure that data are accessible to designated users for first time use and reuse. Some material may be publicly available, whilst other data may be password protected.

- **Transform**
  Create new digital objects from the original, for example, by migration into a different form.
DCC DIGITAL CURATION LIFECYCLE: OCCASIONAL ACTIONS

- **Dispose**
  Rid systems of digital objects not selected for long-term curation and preservation. Documented guidance, policies and legal requirements may require the secure destruction of these objects.

- **Reappraise**
  Return digital objects that fail validation procedures for further appraisal and reselection.

- **Migrate**
  Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from hardware or software obsolescence.
WHY ARE DIGITAL/DATA CURATION IMPORTANT?
WHY DIGITAL CURATION?

- Ross Harvey *Digital Curation* book sums up well:
  - Improving access
  - Improving data quality
  - Encouraging data sharing and reuse
  - Protecting data
  - More data citation
  - Public good obligations
  - Compliance
STREAMS OF ACTIVITY RELATED TO DIGITAL CURATION

(FROM DR. CHRISTOPHER LEE)

- Care & properties of physical media
- Digital forensics & data recovery
- Hardware & software interoperability
- Institutional & manuscripts archives
- Social science data archives
- Physical science data archives
- Librarianship (esp. digital libraries)
- Art & museum curation
- Medical information (e.g. imaging, informatics, health records)
- Lawyers & auditors
- Computer-supported cooperative work (CSCW)
- Management of information systems (MIS)
- Research on documents & document-centric computing
- Standards development
DRIVERS FOR DIGITAL CURATION

- Overwhelming information growth
- Data driven science and commerce
- Need for data reuse
- Federal and funder mandates
- Data publishing and citation
- Regulation, compliance, and accountability across all sectors
- Data protection and freedom of information laws require digital data management
- Personal digital data management now an issue.
TYPES OF SCIENCE

- Observational science – 17th Century on
- Theoretical science – 17th Century on
- Simulation science – later 20th Century on
- Data-driven science – 21st Century
  - New types of computer clusters are emerging that are optimized for data movement and analysis rather than computing
  - Integrated data systems allow data analysis and storage on site instead of requiring download of large amounts of data.
DATA-DRIVEN SCIENCE

• Science as we may imagine it is changing
  • “Today, some areas of science are facing hundred- to thousandfold increases in data volumes from satellites, telescopes, high-throughput instruments, sensor networks, accelerators, and supercomputers, compared to the volumes generated only a decade ago.” –Bell, Hay, & Szalay, 2009
  • a data-intensive inductive approach to genomics (such as shotgun sequencing) is necessary to address large-scale ecosystem questions
  • Storage cannot keep up with the amount of data (and other digital objects)
  • Parallel and grid computing are becoming ubiquitous
DATA-DRIVEN SCIENCE - 2

• Collaborative, networked, and data-driven.
• E-science is “the synthesis of information technology and science that enables challenges on previously unimaginable scales to be tackled.” –Jim Gray, 2007
• Data-driven science slow to develop due to a general lack of understanding of databases, ontologies, schemas, and other data curation tools by the scientific community.
• Grid computing and cloud storage will play major roles in data-intensive science.
• Large role to play for data curators.
CHALLENGES TO DIGITAL & DATA CURATION
PLEASE WATCH

- http://www.youtube.com/watch?v=N2zK3sAtr-4
CHALLENGES TO DATA CURATION AND REUSE

- It takes work – many data creators don’t want to take on extra work after their research project is done
  - YouTube video you just watched captures the attitude toward reuse perfectly.
  - Reuse is based on high-quality metadata that help explain the data to the next user (who did not create the data or even the data creator if he/she uses the data in the future).
  - Best if metadata is created at time of data creation.
CHALLENGES TO DATA CURATION AND REUSE -2

- Data Curation requires effort around the data lifecycle.
  - Lifecycle approach is essential to data curation
  - Need to start curation as early as possible in the lifecycle for the greatest chance of successfully providing long-term access and data reuse.

- Data Curation relies on data standards and sound policies.
  - Like electronic records, preserved data should remain
    - Authentic, reliable, have integrity, and usable
CHALLENGES TO DATA CURATION AND REUSE

• It takes curation knowledge.
  o Most researchers (or any other data creators) have no idea how to curate their data for future use, whether it is their use or someone else’s.
  o There is a great need for digital/data curators as content creators don’t
    • Know how to curate their data well (or at all)
    • Have the time to curate their data (in their view at least)
    • Have strong enough incentives to curate their data

• It is not a matter of being lazy; the cost-benefit equation is just not right for most data curators to push them to learn to curate their data.
THREATS TO DIGITAL CONTINUITY

- Ross Harvey sums these points up well (p. 9).
  - Fragility of the media content
    - E.g., bit rot, accidental file overwriting, malicious changes and deletion
  - Short life of media formats
    - E.g., 8” disks; 5” disks; 3.5” disks; zip drives; jazz drives…
  - Short life of hardware and software/formats/etc.
THREATS TO DIGITAL CONTINUITY - 2

- There may not be money, facilities, or trained staff to curate data.
- Knowledgeable data curators (the folks whose articles we read) are just figuring data curation out (so how can we expect others to know how to do this???)
  - Data curation workflows are just being developed
  - Data curators are lobbying for funds, grants, and mandates
THREATS TO DIGITAL CONTINUITY-3

- Expense of preservation may not seem justifiable
  - People today need to pay for future use – hard sell
- **Data curation may not be anyone’s specific job**
  - If not in a job description (let’s say, of a data creator) it won’t get done
  - If it is in a job description but there is little reward (or punishment as it keeps you from what is viewed as more “valuable” work) it will not be done
- Great need for data curators
THREATS TO DIGITAL CONTINUITY - 4

- Technology is a moving target
  - It takes a while to figure out workflows and develop tools that work with specific technologies (esp. in the open-source world)
  - While curation technologies are being developed to work with specific data types and technologies, the scene continuously changes
    - Not just preservation of static objects
    - Web preservation; complex data preservation, etc.
THREATS TO DIGITAL CONTINUITY - 5

- Metadata, Metadata, Metadata
  - It is expensive to create – lots of person hours
  - Data creators do not generally want to create metadata and when they do, don’t usually use controlled vocabularies, etc.
  - Data curators may not understand the content well enough to create the metadata
  - Some automated metadata creation but a relatively small amount given the need of future re-users of data
- Data without metadata is useless.
THREATS TO DIGITAL CONTINUITY - 6

- Curators may not have legal access to data
- Long-term preservation involving migration may not capture significant properties of content that would be useful to future users.
SUMMARY

- Viewing and curating data over its lifecycle is essential
  - E.g., helping content creators to select appropriate file formats, intermediate storage media, variable names, and metadata
- It is the curator’s job to preserve the authenticity of data that will allow for reproducibility for future users.
- Curator’s add value through metadata, data cleaning, preservation of significant properties, and provision of access.
- There is a wide range of stakeholders.
- The importance of data curation is just being recognized outside of the data curation community.
WHAT LIES AHEAD?
ENGINEERING TOOLS & WORKFLOWS

- Building tools & workflows for content creators (e.g., scientists)
  - Must be easy, obvious, and apparent
  - Metadata capture is paramount
- Building tools & workflows for digital curators
  - We are starting to see these
EDUCATING ABOUT DIGITAL/DATA CURATION

- Everyone has content they want to keep and everyone needs help
- There is an enormous need for lifelong education in this area
- Instruction for
  - Individuals/the public
  - School children
  - College students
  - Graduate students/resesarchers
  - Professors/faculty
  - Workers in all sectors
EDUCATING GOVERNMENTS AND RESEARCH FUNDERS

- Governments and other funders must make data curation a priority
- They must provide funds to support data curation
- Data curation must be evaluated at end of grant projects
- Governments at all levels must employ data curation to manage their own content
PRODUCING AND HIRING MORE DIGITAL CURATORS

- Schools must produce more digital/data curators
- Governments, universities, research projects and labs, businesses, etc. must all hire more digital curators
- Digital and data curation are not going to get done by themselves!
DIGITAL AND DATA CURATION CONTINUING EDUCATION IN THE USA & PROGRAMS AT SILS
DIGITAL CURATION GAP

- Dramatic progress of research and development on digital curation and professional practices of archivists, librarians, and museum curators in the past decade.
- There are now many viable applications, models, strategies, and standards for long-term care of digital objects.
- However, many institutions are either not aware of the options or do not currently have the ability to evaluate and implement them.
LIFELONG EDUCATION

• Never before has there been such a need for continuing education for library, archiving, and IT staff.
• An amazing array of materials to monitor, read, and absorb.
• The web has not made this easier.
• Nice also to have face-to-face educational opportunities to make sense of it all and learn with people from similar backgrounds.
Library of Congress

The DPOE mission is to foster national outreach and education about digital preservation by building a collaborative network of instructors and partners to provide training to individuals and organizations seeking to preserve their digital content.

Train-the-trainer approach

I taught in the first week-long course, September 2011.

Each of the 24 students has gone home and conducted some level of training.
The DPOE Baseline Digital Preservation Curriculum consists of 6 easily understandable topics.

- **Identify** . . . the types of digital content you have.
- **Select** . . . what portion of your digital content will be preserved.
- **Store** . . . your selected content for the long term.
- **Protect** . . . your content from everyday threats and emergency contingencies.
- **Manage** . . . and implement requirements for long term management.
- **Provide** . . . access to digital content over time.
DPOE PYRAMID

- **Executive**
  - CEOs/CIOs/Administrators
  - Webinars, Corporate Briefings

- **Managerial**
  - Project Managers/Program Managers

- **Practical**
  - Practitioners/Support Staff/Hands-On Staff

- **Hands-On**
  - Workshops, Online Courses

- **Project Planning**
  - Workshops, Webinars
## DPOE AUDIENCES

<table>
<thead>
<tr>
<th>Audience</th>
<th>Outreach Methods</th>
<th>Educational Approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive Administrators</td>
<td>Webinars, Corporate briefings, Media</td>
<td>Provide information on trends in industry, the U.S., and worldwide, using tools such as webinars, corporate briefings, and media, to inform an organization-wide strategy/commitment in digital preservation.</td>
</tr>
<tr>
<td>CEOs, CIOs, Governance bodies, Funders</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Managerial</strong></td>
<td>Professional reading—e.g., journals, listservs, conferences—e.g., presentations, booths, panels, social media, traditional media</td>
<td>Provide a variety of educational options (e.g., half-day, daylong, onsite, Web-based) on setting up and running a successful digital preservation program.</td>
</tr>
<tr>
<td>Program managers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project managers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Line managers</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Practical</strong></td>
<td>Workshops, online tutorials, booklets and other publications, social media, traditional media</td>
<td>Provide a variety of instructional options (e.g., half-day, daylong, onsite, Web-based) on demonstrations of methods and workflows in digital preservation.</td>
</tr>
<tr>
<td>Practitioners</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hands-on staff</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Support staff</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## DPOE AUDIENCES

<table>
<thead>
<tr>
<th>Audience</th>
<th>Outreach Methods</th>
<th>Educational Approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive and Administrators</td>
<td>Webinars, Corporate briefings, Media</td>
<td>Provide information on trends in industry, the U.S., and worldwide, using tools such as webinars, corporate briefings, and media, to inform an organization-wide strategy/commitment in digital preservation.</td>
</tr>
<tr>
<td>CEOs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIOs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Governance bodies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Funders</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Managerial</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program managers</td>
<td>Professional reading—e.g., journals, listservs</td>
<td>Provide a variety of educational options (e.g., half-day, daylong, onsite, Web-based) on setting up and running a successful digital preservation program.</td>
</tr>
<tr>
<td>Project managers</td>
<td>Conferences—e.g., presentations, booths, Panels</td>
<td></td>
</tr>
<tr>
<td>Line managers</td>
<td>Social media, Traditional media</td>
<td></td>
</tr>
<tr>
<td><strong>Practical</strong></td>
<td>Workshops, Online tutorials</td>
<td>Provide a variety of instructional options (e.g., half-day, daylong, onsite, Web-based) on demonstrations of methods and workflows in digital preservation.</td>
</tr>
<tr>
<td>Practitioners</td>
<td>Booklets and other publications</td>
<td></td>
</tr>
<tr>
<td>Hands-on staff</td>
<td>Social media, Traditional media</td>
<td></td>
</tr>
<tr>
<td>Support staff</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SOCIETY OF AMERICAN ARCHIVISTS’ DIGITAL ARCHIVES SPECIALIST CERTIFICATE

• The DAS Curriculum is structured in four tiers of study
• Students can earn a certificate by completing required coursework and passing both course and comprehensive examinations.
• Three audiences similar to DPOE:
  • The Archivist Practitioner is a hands-on, front-line archivist who manages or will manage electronic records personally.
  • The Archivist Manager is an archivist who has oversight over the work of other professional archivists and who may or may not manage electronic records directly.
  • The Archivist Administrator is an archivist oversees archivist managers, who is responsible for organizational planning, and who does not manage electronic records directly but must ensure the organization’s capacity to do so.
DAS – FOUNDATIONAL COURSES

- Focus on the essential skills that archivists need to manage digital archives.
- They focus primarily, but not exclusively, on the needs of practitioners—archivists who are or will be working directly with electronic records.
- These courses present information that an archivist might implement in the next year.
DAS – TACTICAL & STRATEGIC COURSES

• Focus on the skills that archivists need to make significant changes in their organizations so that they can develop a digital archives and work seriously on managing electronic records.

• They focus primarily, but not exclusively, on the needs of managers—those archivists who manage other professionals and who oversee programmatic operations.

• These courses present information that an archivist might implement in the next five years.
DAS – TOOLS & SERVICES COURSES

• Focus on specific tools and services that archivists need to use for their work with digital archives.

• They are practical courses focused on specific software products and other tools and they focus primarily, but not exclusively, on the needs of practitioner archivists.

• These courses present information that an archivist could implement immediately.
DAS – TRANSFORMATIONAL COURSES

- Focus on the skills that archivists need to change their working lives dramatically and transform their institutions into full-fledged digital archives.
- They focus primarily, but not exclusively, on the needs of administrators—those archivists with oversight over the entire archival enterprise of an institution.
- These courses present information that an archivist might implement over the course of the next ten years.
DIGCURV

- Digital Curator Vocational Education Europe
- Funded by the European Commission’s Leonardo da Vinci programme to establish a curriculum framework for vocational training in digital curation.
- CURATE The Digital Curator Game is available for download to all Network Members.
Developed “lenses” to focus on digital curation skills and knowledge needed by:

- Practitioners
- Managers
- Administrators

Lenses built on DigCCurr Matrix, DPOE principles and curriculum, and SAA’s DAS curriculum
MANAGER’S LENS

- Understand digital preservation and repository audit and certification standards
- Be able to successfully prepare for an audit of curation functions and lead repository through certification process
- Understand risk management theory, standards, practice, and techniques in digital repositories
- Possess ability to assess, analyse, monitor and communicate risks
- Ability to create succession planning policy

- Create, manage and monitor project plans
- Possess firm knowledge of data management requirements
- Understand project management activities and innovative practices
- Possess ability to make sound decisions, based in information produced by project team
- Be knowledge of financial planning, cost analysis and economic sustainability
- Motive staff
- Create a team environment

- Possess extensive digital curation terminology, knowledge, & experience
- Understand fundamental digital curation principles, including lifecycles
- Be knowledgeable of designated community
- Be able to select appropriate technological solutions
- Develop a professional network for support

- Know and be able to articulate information and records management principles
- Understand and be able to articulate the benefits and long-term value of collections
- Contribute to institutional policies, including criteria, for selection/appraisal
- Apply selection/appraisal criteria to collections and train staff

- Prioritise digital objects/collections in terms of value to the institution & current risk level
- Continuously monitor and evaluate digital curation technologies
- Monitor & assess needs of designated community
- Know how to conduct user analysis

- Understand an array of data structures, file types, & systems
- Possess substantial knowledge of metadata standards
- Understand information seeking strategies, access technologies, & user data sharing behaviors

- Possess ability to identify malpractice (fraud, plagiarism)
- Understand what constitutes responsibility, accountability and good practice in digital curation
- Make decisions transparently
- Plan & implement sound staff training and development

- Communicate across disciplines and staff levels
- Plan and deliver dissemination activities
- Ability to articulate extent of knowledge and explain importance of digital curation to peers, other staff & wider public
- Ability to articulate value of collections to peers, other staff & wider public
- Make case for funding of curation activity
- Make case for practitioner training and development

- Be willing to incorporate new and emerging digital curation technologies and processes
- Translate knowledge of technology and processes into new services and tools for needs of designated community
- Assess, extend and generate digital curation models for cultural heritage and other domains

- Understand institutional policy frameworks in which digital repositories operate
- Apply appropriate actions to curation workflow

- Understand legal frameworks in which curation is taking place, e.g. data protection, freedom of information, moral rights
- Be familiar with domain policies and standards for management and preservation of digital objects
- Understand institutional policy frameworks in which digital repositories operate
- Apply appropriate actions to curation workflow

- Be knowledgeable of copyright and IPR law for digital materials
- Understand validation techniques to detect copyright infringement
- Incorporate rights management into curation workflows
- Understand information ownership principles

- Professional conduct
- Legal requirements
- Integrity
- Communication skills
- Responsiveness to change

- Ethics, principles and sustainability
- IPR and Copyright
- Managers Lens
STATE ELECTRONIC RECORDS INITIATIVE (SERI)

- Launched in July 2011 by Council of State Archivists (CoSA);
- Focused on “improving efforts to manage, preserve, and provide access to state government electronic records nationwide”

Phase 1

- Survey and phone interviews of state archivists and records managers about existing electronic records programs -> “composite nationwide profile including strategies used to create, fund, and maintain state electronic records programs”
- Generated a report: http://www.statearchivists.org/seri/phase_one_report.htm
Phase 2:
- Education and training
- Awareness raising about electronic records
- Addressing governance issues within states
- Best practices, tools, and implementation strategies

Supported through LSTA funds (IMLS) awarded by the Indiana State Library and the Kentucky Department for Libraries and Archives

In development: professional institutes for working professionals
Closing the Digital Curation Gap

IMLS Sponsored Project

Another THANK YOU to IMLS
CLOSING THE DIGITAL CURATION GAP

• IMLS-, JISC, and DCC, and SILS-funded project based at the School of Information and Library Services at the University of North Carolina at Chapel Hill.

• 2009-2013.

• Partners: Drs. Helen Tibbo and Christopher Lee and students Heather Bowden and Courtney Bailey at UNC and Neil Grindley (JISC) and Joy Davidson (DCC) in UK.

• Seeks to fill this gap between research and practice by providing guidance resources for professionals in small- to medium-sized repositories.
COLLABORATIVE

- The CDCG collaboration is serving as a locus of interaction between those doing leading edge digital curation research, development, teaching, and training in academic and practitioner communities; those with a professional interest in applying viable innovations within particular organizational contexts; and organizations charged with disseminating such innovation and best practices.
METHODOLOGIES

- User-centered design
  - Interviews
  - Focus Groups
  - Survey
- We grounded our advice and guidance in the real experiences of people working in cultural heritage institutions.
- Including video interviews.
- Online Digital Curation Guides, or “Getting Started Guides.”
RESEARCH APPROACH

• Drawn from previous studies, esp. those of the Northeast Document Conservation Center (NEDCC) in 2006 and Cornell University Library in 2005.

• Focus groups
  • 4 groups; 25 participants
  • Held at ALA, SAA, MCN

• Participant recruitment was directed toward professionals who were responsible for the care of digital collections and had sufficient background knowledge on digital curation to contribute to the discussion.
FOCUS GROUPS

- What kinds of digital curation activities do you currently practice?
- For these activities, what tools and resources have you used?
- Where and how did you find these tools and resources?
- What other types of tools and resources would you find to be helpful?
- We then presented a draft mockup of a “decision tree” tool and elicited the group’s feedback on it, including how they might use it, and what they would see as the main opportunities or benefits and challenges of using it.
COMMON SCENARIOS

• Common Scenarios form the basis of our Getting Started Guides.
  • Archiving web sites,
  • Building institution repositories,
  • Caring for digitized collections,
  • Managing data (as opposed to text-based materials),
  • Acquiring information off external storage media, and
  • Caring for digital audio visual materials.
GETTING STARTED GUIDES

• Built in an open source Drupal content management system to afford interactivity, adaptability, and sustainability.

• Guides are linked to the resources that we have already started collecting in the Drupal-based Digital Curation Exchange (DCE) website.
  • www.digitalcurationexchange.org
Welcome to the Digital Curation Exchange

The Digital Curation Exchange has been created to serve as a “town center” for YOU: the practitioners, researchers, educators, and students of digital curation. Create an account and add your digital curation events, jobs, resources, or questions and discuss what others have shared in the comments section. DON'T BE SHY! We welcome you and look forward to your contributions!
ORGANIZATION OF GUIDES

- Each guide is organized around seven main verbs, which we have borrowed from the work of the Library of Congress’s Digital Preservation Outreach and Education (DPOE) initiative:
  - prepare,
  - identify,
  - select,
  - store,
  - protect,
  - manage and
  - provide.
GUIDE CONTENT

• Under each verb, we provide one or more questions,
  • e.g. “How should I prepare to archive web sites?” and
  • “What do I need to identify in order to archive web sites?”
  • These questions serve as hyperlinks and titles to more detailed
    pages that summarize main considerations and pointers to
    existing resources that the user might find helpful.
  • Adding audio and video interviews with experts.
SOURCES OF GUIDE STRUCTURE & CONTENT

• Interviews with curation experts,
• Existing and Emerging Frameworks for Digital Curation Education,
• Digital Curation Lifecycle Models,
• Digital Curation Curriculum (DigCCurr)
  • Matrix of Digital Curation Knowledge and Competencies,
• Digital Preservation Management (DPM) Workshops,
• Digital Preservation Outreach and Education (DPOE) Program, and
• SAA’s Digital Archives Specialist (DAS) Certificate Program
DigCCurr Project
say: dij-seeker


- 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)

- Dr. Carolyn Hank (UTK) was project manager

- Ran July 1, 2006 – December 31, 2009
DIGCCURR I COMPONENTS

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

• Graduate-level (master’s) curricular framework
• Course modules
• Course development
• Experiential components
• International guest speakers

Two International Symposia:

 MATRIX OF DIGITAL CURATION KNOWLEDGE & COMPETENCIES

- Tool for thinking about, planning for, identifying & organizing material to cover in curriculum.
- Each unit of curriculum content can address one or more dimensions.
- Helping us to address a fundamental issue: All digital curation students should all get some aspects of the curriculum, but other aspects will only be necessary for students planning to work in particular types of places or jobs (i.e. balancing core vs. specialized knowledge).
SIX MATRIX DIMENSIONS

- Mandates, values & principles.
- Professional, disciplinary or institutional/organizational context.
- Transition point in information continuum/lifecycle.
- Type of resource.
- Function or skill.
- Prerequisite knowledge.
DIGCCCURR II

- This project is also funded with IMLS Laura Bush funds.
- A collaboration of the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill (UNC-CH) and the U.S. National Archives and Records Administration (NARA) and Univ. of Toronto through Seamus Ross.
- Project to run August 1, 2008 – March 31, 2013.
DIGCCURR II KEY ACTIVITIES

- PhD Fellowships - 6
- Digital Curation Exchange (DCE)
  - [http://digitalcurationexchange.org/](http://digitalcurationexchange.org/)
- Professional Institutes
  - Week-long summer workshop for practitioners.
  - Follow-up session held approximately six months later.
  - [http://www.ils.unc.edu/digccurr/institute.html](http://www.ils.unc.edu/digccurr/institute.html)
  - Fifth to be held May 12-17, 2013 (125 participants from US, Canada, & Europe so far)
  - Held session at Danish Royal Library in Copenhagen, June 2012.
- Public Symposia with DigCCurr & Bit Curator Advisory Boards
  - January 2010-2012
  - 2010-Engaging Communities for the Curation of Digital Products of Scholarly Endeavors; 2011-2012 Curate Gear
- Ph.D. Seminar Series, 2012-2013
Welcome to the Digital Curation Exchange

The Digital Curation Exchange has been created to serve as a "town center" for YOU: the practitioners, researchers, educators, and students of digital curation. Create an account and add your digital curation events, jobs, resources, or questions and discuss what others have shared in the comments section. DON'T BE SHY! We welcome you and look forward to your contributions.
DIGCCURR PROFESSIONAL INSTITUTE

- Conducted 2009-
- Supported by IMLS 2009-2011
- Held in Copenhagen for the Danish Royal Library, June 2012; will visit LAC in Feb. 2015
- Three stage institute
  - Before the institute: We send readings, use DCE
  - During institute: a week in Chapel Hill
  - After the institute: we follow participants, keep in touch with them, and they come back again in January to report on their progress on their projects.
• Detaching Bits from their Physical Media: Considerations, Tools and Methods
  LAB - Curation of Unidentified Files
• Returning to First Principles: Core Professional Principles to Drive Digital Curation
• Characterization of digital objects
• LAB - Assessing File Format Robustness
• Access and use considerations
• Access and user interface examples
• How and why to conduct research on digital collection needs
• LAB - Analyzing server logs and developing strategies based on what you find
DIGCCURR PROFESSIONAL INSTITUTE 2014-2015

- Institute Components: (may be subject to some revisions and reorganization)
  - Overview of digital curation definition, scope and main functions
  - Where you see yourself in the digital curation landscape
  - Digital curation program development
  - Engendering Trust: Processes, Procedures and Forms of Evidence
  - LAB - DRAMBORA in action
  - Strategies for engaging data communities
Overview and characterization of existing tools
LAB - Evaluating set of software options to support a given digital curation workflow
Formulating your six-month action plan - task for each individual, with instructors available to provide guidance
Summary of action plans
Clarifying roles and expectations for the next six months

January 5-6, 2015
Participants in the May event will return to Chapel Hill in Jan. 2015 to discuss their experiences in implementing what they have learned in their own work environments. Participants will compare experiences, lessons learned and strategies for continuing progress. Wednesday, January 6th will be a public symposium, free to the Institute participants.
DIGCCURR, JANUARY 2015

- January 5-6, 2015
  Participants in the May event will return to Chapel Hill in Jan. 2015 to discuss their experiences in implementing what they have learned in their own work environments. Participants will compare experiences, lessons learned and strategies for continuing progress. Wednesday, January 7th will be a public symposium, CurateGear, free to the Institute participants.
CurateGear: Enabling the Curation of Digital Collections

January 7, 2014 – Friday Center, Chapel Hill, North Carolina

An interactive day-long event focused on digital curation tools and methods. See demonstrations, hear about the latest developments, and discuss application in professional contexts.

Presenters:

- Jonathan Crabtree, Odum Institute for Research in Social Science
- Lisa Gregory, State Library of North Carolina
- Barbara Guttman, National Institute of Standards and Technology
- Carolyn Hank, McGill University
- Chien-Yi Hou, University of North Carolina
- Greg Jansen, UNC Libraries
- Leslie Johnston, Library of Congress
- Cal Lee, University of North Carolina
- Matt Kirschenbaum, Maryland Institute for Technology in the Humanities
- Nancy McGovern, MIT Libraries
- Richard Marciano, University of North Carolina
- Mark Matienzo, Yale University
- Trevor Owens, Library of Congress
- David Pearson, National Library of Australia
- Doug Reside, New York Public Library
- Seamus Ross, University of Toronto
- Ryan Scherle, Duke University
- Seth Shaw, University Archives, Duke University
- Katherine Skinner, Educopia Institute
- Helen Tibbo, University of North Carolina
- William Underwood, Georgia Tech
- Peter Van Garderen, Artefactual Systems
- Doug White, National Institute of Standards and Technology
- Kam Woods, University of North Carolina
Next DigCCurr Professional Institute
June 1-5, 2015
OBSERVATIONS & IMPLICATIONS FOR CONTINUING EDUCATION

• Commonalities across all these continuing education efforts but also many differences.
• All programs address the notion of Audience.
  • There is fundamental agreement that educational programs must be geared not only to experience, knowledge, and skill levels, but also to the specific roles and responsibilities that individuals assume.
  • For digital curation there is no effective training program that takes a one-size-fits-all-approach.
COMPLEXITY OF TRAINING LANDSCAPE

- Everyone wants a map – what to take; where to begin.
- Specialization results in extensive complexity and the need for many training resources and thus much cost and confusion.
- Because digital curation involves work across the entire information continuum and around the DCC Digital Curation Lifecycle Model, training cannot simply be aimed at novice, intermediate, or advanced audiences as one might with teaching algebra or a foreign language.
IMPLICATIONS FOR EDUCATION

• Overviews and awareness needed.
• Specialized courses are also needed - because the tasks are diverse, so too must be the training to support these functions.
• Digital curation practitioners need both knowledge and hands-on skills.
• Need to contextualize teaching within the participants’ own organizational contexts.
• Course delivery mechanisms must be appropriate and varied.
• Course length and intensity are question areas.
SILS DIGITAL CURATION PROGRAMS

• Master’s of Science in Library or Information Science.
  • Concentration in Archives and Records Management.
  • Certificate in Digital Curation.
  • Dual MSIS/MSLS – MPA program with UNC School of Government.

• 10 Ph.D. students in digital/data curation presently.
• Post Master’s Certificate in Data Curation.
• DigCCurr Professional Institute.
• Professional Science Master’s in Digital Curation
THANK YOU!

- Questions?
REFERENCES


Characterizing, analyzing and evaluating the producer information environment

Submission and transfer scenarios – push and pull (illustrative examples)

Defining submission agreements and policies

Strategies for writing policies that can be expressed as rules and rules that can automatically executed

LAB - Making requirements machine-actionable

Importance of infrastructure independence

Overview of digital preservation challenges and opportunities

Managing in response to technological change