Data Curation in a Research Library: A Start-Up Story

Patricia Hswe | Penn State University Libraries

Presentation for Philadelphia Chapters of Medical Library Association | Special Library Association
March 16, 2012
Plot points of the story

• Prologue
  o (Penn State facts, Libraries' context)
• Data curation - defined, modeled, mapped
• NIH, NSF, NEH - the plot thickeners
• Data curation at Penn State: a start-up story
  o Technical infrastructure development
  o Collaborations: within & beyond Libraries
    ▪ Research Data Management Services Team
    ▪ Data Curation Working Group
  o Inreach / Outreach (workshops, DMP Tool)
• Where we are . . . where we hope to be
Prologue

The Penn State Libraries Context
Penn State University

- 24 campuses ~ 86,200 students
- World Campus ~ 10,300 students
- 96,500 total*
  * total includes full-time and part-time students
- 6,010 full-time faculty
- 2009-2010: $780 million in research funding, 60% of it federal $$
  - $102 million - industry-sponsored
- 8 interdisciplinary research institutes
- 2 institutes focused on defense research
University Libraries' Strategic Plan 2008-2013

• Accelerating the Transition to a Digital Collection
  o improve access, improve collections in new, interdisciplinary areas

• The Knowledge Commons at University Park
  o integrate student research support with technology services

• Cyberinfrastructure, e-Content, and Data Stewardship
  o cohesive suite of access, security, discovery, preservation, curation, archival, and storage services
We rock at digitization

• We have a process
  o Digital Collections Review Team
  o Digitization and Preservation Department
  o Digital Operations Team
• Almost 80 digitized collections online
• Known for our digitized historical newspapers
• Long-running ETD service
• Office of Digital Scholarly Publishing
  o E-journals and conference proceedings
BUT . . .

• No unifying infrastructure (e.g., no institutional repository) - *as yet*
  o Four different platforms delivering digital content
• Little attention to use and users of our digitized content - *slowly improving on this*
• Digitized collections are not a formal part of overall collection development process
• No support yet for data in digital format - i.e., data generated by faculty and students - *this is coming!*
Data Curation

defined, modeled, mapped
Data curation is the “active and ongoing management of data through its life cycle of interest and usefulness to scholarly and educational activities.”

~ Data Curation Education Program, Graduate School of Library & Information Science ~
How is data curation different from digital curation?

Maintains, preserves, adds value to research data through its lifecycle
~ Digital Curation Centre ~
The terminology may differ, but the goals align, which means overall stewardship of data.
Modeling the data lifecycle

pick one, any one
Lots of models out there - which one makes sense? How do you know?
and then there's just the simple workflow you want

Mapping as an approach

~ adapted from Mike Furlough, Associate Dean for Research and Scholarly Communications, Penn State Libraries
<table>
<thead>
<tr>
<th>Action</th>
<th>What/How of Data?</th>
<th>Who of Data?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conceptualize</td>
<td>Design study or experiment</td>
<td>Researcher</td>
</tr>
<tr>
<td>Create</td>
<td>Collect observations</td>
<td>Researcher</td>
</tr>
<tr>
<td>Analyze</td>
<td>Regression analysis</td>
<td>Researchers</td>
</tr>
<tr>
<td>Receive</td>
<td>My grant is over. Can you take this for the next 40 years?</td>
<td>Researcher/Data caretaker</td>
</tr>
</tbody>
</table>
| Prepare (preservation actions) | Create documentation, metadata, or convert to standard format                                             | Caretaker? Researcher?  
Who's the expert? Where does caretaker reside? |
| Ingest                  | Final hand-off & validation, register in repository, assign identifiers, assign access restrictions & rights. | Caretakers, local or international    |
| Store                   | Hardware acquisition & migration, backup center, data center, maintenance                                  | Caretakers                            |
| Access                  | Find citation in publication, search data repository, search Google, download                              | Students, researchers, public         |
The learning outcomes of mapping

• In deconstructing the model, we become familiar with it
  ○ Actions (column 1)

• In becoming familiar with the model, we contextualize it
  ○ What/How of Data (column 2)
  ○ Who of Data? (column 3)

• In contextualizing, we start seeing paths, processes, services
Mapping helps kick off "the putting into practice" part informs consultations and roles and responsibilities, and can shape how we define uses, users, and services
NIH, NSF, NEH

the plot thickeners
The trend toward data sharing & data management

- National Institutes of Health - helped set the standard with data sharing policy (2003)
- National Science Foundation - data management planning (DMP) requirement (2011)
  - Driven by community of interest
  - General guidance - non-prescriptive, even vague
- National Endowment for the Humanities Office of Digital Humanities
  - Sustainability and data management plan

DMPs are growing in significance . . .
heard for the 1st time from an NSF program officer that meritorious proposals have not funded because of an inadequate data plan #CASCSM
Data curation at Penn State

a start-up story
Technical infrastructure development

• Platform review in spring 2010
• Explored curation microservices approach
  o Gathered use cases in curation as part of proof-of-concept (summer 2010)
  o Prototyped a curator dashboard tool for digital object management (fall 2010/spring 2011)
• Developing repository services platform (fall 2011-fall 2012)
  o To support suite of services for publishing and curation of scholarship and data produced by Penn State Research Community
Diagram showing technical architecture for Hydra-based repository services platform.

By Mike Giarlo, Digital Library Architect, Penn State University Libraries.
Collaborations within & beyond the Libraries

• Research Data Management Services Team
  ○ Diverse team make-up (e.g., liaison librarians, metadata librarian, digital library architect, digital collections curator, data archivist from Population Research Institute, specialist in storage services)
Toolkit site:
http://www.libraries.psu.edu/psul/scholar/datamanagement.html
Data Management

Documenting your data is an initial step in managing them

Managing data is an integral part of the research process. It can be challenging particularly when studies involve several researchers and/or when studies are conducted from multiple locations. How data is managed depends on the types of data, how data is collected and stored, and how it is used through the length of the study.

The outcome of your research depends in part on how well you manage your data. Managing data helps you as a researcher organize research files and data for easier access and analysis. It helps ensure the quality of your research. It supports the published results of your work and, in the long term, helps ensure accountability in data analysis. Good data management starts with comprehensive and consistent data documentation and should be maintained through the life cycle of the data.

- Designate the responsibilities of every individual involved in the study.
- Designate how data will be stored and backed up.
- Designate how data will be dealt with through each modification of the study.
- And make sure the rules are followed through!

Data documentation encompasses the following:

- names, labels and descriptions for variables, records and their values
- explanation of codes and classification schemes used
Collaborations beyond the Libraries - 1

• **Data Curation Services Working Group**
  - Assoc. VP for Research
  - Heads of Research Institutes
  - Assoc. Dean for Research & Scholarly Communications - Libraries
  - Senior Director - Information Technology Services
  - Director, Strategic Interdisciplinary Office
  - Director, Harrell Health Sciences Library (Penn State Hershey)
  - Digital Library Architect
  - Digital Collections Curator
Collaboration beyond the Libraries - 2

- Strategic Interdisciplinary Research Office, located in the Office of VP for Research
  - Responsible for helping researchers with proposals on much larger scale (multimillion $$ grants)
  - Libraries are occasionally pulled in
    - Collections in Support of Biological Research (CSBR) - an NSF grant program
    - Sustainability Research Networks Competition - also NSF
  - Major collaborative effort - local guidance for Penn State Researchers
Inreach / Outreach Efforts