Data Curation and Data Curation Services in Academic Libraries - an Overview

CLIR Postdoctoral Fellowship Program Summer Seminar
Bryn Mawr College | 30 July 2014

Patricia Hswe | Pennsylvania State University Libraries
Digital Content Strategist
Head, ScholarSphere User Services
My Context

How I Got Here

“Cubicle wall” - by sidewalk_flying CC BY 2.0
First CLIR postdoc cohort

- U. Illinois at Urbana-Champaign: Slavic & East European Library, 2004-2006
- Slavic Digital Humanities Fellow
- Digital projects, reference services, workshops

Summer 2004
Library and information science school - GSLIS

- Digital Library track
  - Document Modeling; Use and Users; Interfaces to Information Systems; Digital Libraries; Information Modeling; Metadata; Information Transfer & Collaboration in Science; Ontology Development

- Graduate and research assistantships - Digital preservation projects, Engineering and Mathematics libraries
National Digital Information & Infrastructure Preservation Program

{ Since 2008 }

Publishing and Curation Services
Penn State University Libraries

<table>
<thead>
<tr>
<th>Data Curation</th>
<th>Scholarly Publishing</th>
<th>Digital Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>About ScholarSphere</td>
<td>Journals and Proceedings</td>
<td>Digital Collections</td>
</tr>
<tr>
<td>Data Management</td>
<td>Open Access</td>
<td>Learning as Play Project</td>
</tr>
<tr>
<td>DMP Tutorial</td>
<td>Copyright</td>
<td>People's Contest Project</td>
</tr>
</tbody>
</table>
Template for the Talk

- What is data curation?
- Curation is *not* a solo undertaking.
- Developing a repository service - one piece of the data curation puzzle.
- Suggested essentials for getting started.
What is data curation?

What is data?
What isn’t data curation?
Purposes.

“20080528-002corrupted” - by nvshn CC BY SA 2.0
Research data may be -

- Observational (e.g., sensor data, data from surveys)
- Experimental (e.g., gene sequencing data)
- Simulation (e.g., climate modeling data)
- Derived or compiled (e.g., text mining, 3D models)
- Reference or canonical (e.g., static, peer-reviewed data sets, likely published and curated)

(from U. of Edinburgh, Information Services)
Examples of data formats -

- Text
- Numerical
- Multimedia
- Models

- Data specific to a discipline (e.g., Chrystallographic Information File, or CIF)
- Data specific to an instrument (e.g., images from a digital microscope)

from U. Edinburgh, Information Services
Leaves of Grass

Brooklyn, New York: 1855.
This XML file does not appear to have any style information associated with it. The document tree is shown below.

```xml
http://www.loc.gov/standards/mets/profile_docs/mets.profile.v1-2.xsd">
    <LOCTYPE="URL">http://www.loc.gov/mets/profiles/00000023.xml</LOCTYPE>
</METS_Profile>
```

**The Walt Whitman Archive METS Profile Version ARCHIVE**

The Walt Whitman Archive uses this profile to describe the organization and structure of the thematic research collection. Digital objects that comprise the archive described by this profile include but are not limited to: images, html, xml, and tei-based structured text. The archive hierarchy describes objects as one of two types: Whitman Authored Materials and Non-Whitman Authored Materials. Digital objects can consist of one or both types. This profile references METS instance files for items that comprise the Archive. Any Archive items not represented by METS instance files are referenced by this profile through the respective html pages directly.

**<date>2007-01-30T00:00:00</date>**

**<contact>**
- **<name>Stacey Berry</name>**
- **<address>Center for Digital Research in the Humanities, University of Nebraska–Lincoln, 319 Love Library/P.O. Box 884100, Lincoln NE, 68588-4100</address>**
- **<phone>402.472.4547</phone>**
- **<email>cdrh@unl.edu</email>**

**<related_profile RELATIONSHIP="parents" URI=""/>**

**<extension_schema ID="ext1">**
- **<name>MODS</name>**
- **<URL>http://www.loc.gov/standards/mods/v3/mods-3-1.xsd</URL>**
- **<context>mets/dmdSec/mdWrap/xmlData</context>**

**<note>**
While the schema referenced above represents v.3.1 of MODS, all v.3.x MODS schemas are largely backwards compatible with previous incremental releases of version 3. METS documents using any version 3 MODS schema are accommodated by this profile.

</note>

</extension_schema>

**<extension_schema ID="ext2">**
- **<name>NISO MIX</name>**
- **<URL>http://www.loc.gov/standards/mix/mix.xsd</URL>**
- **<context>mets/adcSec/techMD/mdWrap/xmlData</context>**

**<note>**

http://www.loc.gov/standards/mets/profiles/00000023.xml
What isn’t data curation?

- Only backing up data
- Not equal to data management, though overlaps
- Digital preservation
- Vs. digital curation

“Data” - by UWW ResNET CC BY-NC-SA 2.0
Collection of data (selection and appraisal)

Documentation / description (metadata)

Ingest / preservation / storage

Dissemination / sharing access / use / reuse

Transformation (migration, derivatives)

Standards

INFRASTRUCTURE

Policies, rights, conditions

"Gulliver's collections" by Yohan Creemers CC BY-SA 2.0
**Examples of data curation activities** -

- Organizing data
- Reformatting, compressing (e.g., .tar, .zip), normalizing
- Planning for long-term security
- Protecting, redacting, anonymizing data
- Applying licenses to data
- Providing data citation guidance

“Clean up, or you’re out! :Brooklyn Street Sign” - by [emilydickinsonridesabmx](https://www.flickr.com/photos/emilydickinsonridesabmx) CC BY 2.0
Purposes: proper curation of research data should enable -

- Verifiability
- Understanding of data use
- Identification of responsibility
- “Findability”
- New research
- Improved guidelines, policies
Keep in mind: use cases

- Researchers’ goals
- Current research workflows and desired workflows
- Role / context of “data keeper”
- Intended uses
- Primary users of the data
- Types of access
- Frequency of access / use
- Potential outcomes / impacts
Curation is not a solo undertaking

Thinking beyond models

“Soloist” - by Chris JL CC BY-NC-ND 2.0
Plan research/design methodology

Collect/Generate data

Document/Describe data

Process/Analyze data

 Preserve (Prepare for ingest)

Create derivatives/reuse

Share/Access data

Store data

MANAGE DATA
Plan research/design methodology

Collect/Generate data

Document/Describe data

Process/Analyze data

Preserve (Prepare for ingest)

Store data

Share/Access data

Create derivatives/reuse

Data curators/librarians, IT

Researchers
Researchers
Data curators / librarians, IT

Plan research/design methodology

Collect/Generate data

Document/Describe

Process/Analyze data

Preserve (Prepare for ingest)

Store data

Share/Access data

Create derivatives / reuse

Manage data

COLLABORATION

COLLABORATION
Researchers

Data curators / librarians, IT

Publishers

Plan research/design methodology

Collect/Generate data

Document/Describe

Process/Analyze data

Preserve (Prepare for ingest)

Create derivatives/reuse

Share/Access data

Store data

Users of the data!

COLLABORATION

COLLABORATION

COLLABORATION

COLLABORATION
Curation involves building and sustaining -

- Community
- Partnerships
- Integration toward solutions

“Relationships” - by beatnic CC BY 2.0
Examples of libraries leading in data curation services

- Purdue University Libraries
- The Library - UC San Diego
- NYPL Labs
- University of Minnesota Libraries
Developing a Repository Service

One piece of the data curation puzzle
ScholarSphere didn’t happen overnight.
Building a Community of Practice: The Curation Architecture Prototype Services (CAPS) Project at Penn State University

**BACKGROUND:** Review of Legacy Platforms
- Myriad, unconnected workflows
- Two platforms effectively moribund
- Another platform without upgrades since 2007
- Impossible to search content across platforms
- No support for e-records, little for research data sets
- Lack of unifying architecture

**MORE BACKGROUND:** Curation Microservices Pilot
- Mix and match, swap in/swapped out services as needed
- Explored curatorial use cases to drive development of microservices
- **KEY:** Laid foundation for prototyping a service

**How Prototyping a Curation Service Led to Building a Community of Practice – The Story of CAPS**
- **Aggressive time line** – January to March 2011
- **Key project roles:** project manager, digital library architect, programmer analyst, metadata librarian, access archivist, and digital collections curator
- **Use cases** kick-started prototyping
- **Daily meetings** allowed team to check in briefly for updates & discussion
- **Weekly meetings** with stakeholders (archivists, subject experts, digitization staff)
- Stakeholders were consulted early on for metadata needs, with concerted attention to creating a framework allowing for annotation of objects throughout their lifecycle
- **Followed agile methods** to drive development, collaboration, feedback, iterations
- **Listened to & learned from stakeholders,** being as responsive as possible
- **User requirements were prioritized,** but all of them were documented; no need was dismissed – both to inform future development but also to assure stakeholders their perspectives counted and were crucial to future development and iterations
- **Survey of stakeholders** at project’s end revealed they felt listened to by the project team, and the resulting prototype reflected their requirements and deliverables
- **Next steps:** chart a production path and prioritize pilot projects for further testing

**CAPS architecture – our point of departure**
- **Dashboard interface giving view of objects in system**
- **Ingest tool interface to upload and describe digital object**
- **Wiki where stakeholders documented testing of CAPS**
- **Prototype of public interface for curated collections**

Co-authors: Michelle Belden, Kevin Clair, Daniel Coughlin, Michael Giarlo, Patricia Hswe, and Linda Klimczyk

CAPS Poster - [https://scholarsphere.psu.edu/files/r781wj67c](https://scholarsphere.psu.edu/files/r781wj67c)

Platform review, pilot project, prototype

2010 through 2011
In 2012: We did 9 months of development before beta release . . .
9 months during which we prioritized user engagement

- Enlisted liaison librarians
- Held bi-weekly “stakeholder” meetings to discuss use cases
- Met with researchers to gauge their needs, learn their use cases
- Demoed the evolving tool, got feedback
- Frequent communications about progress
- Usability testing w/ librarians, faculty & students; developers present
Hydra is a Repository Solution

- IMPORTANT change of Hydra email address
- Hydra Connect #2 announced – booking open
- CALL for Expressions of Interest in Hosting the Annual Open Repositories Conference, 2016
- Hydra-head 7.1.0 released
- Registration Open for Fall Hydra Camp – August 26-29 – Princeton, NJ

See All News & Events

Get Involved
- News & Events
- Hydra events diary
- Hydra wiki
What is ScholarSphere?

ScholarSphere is a secure repository service enabling the Penn State community to share its research and scholarly work with a worldwide audience. Faculty, staff, and students can use ScholarSphere to collect their work in one location and create a durable and citable record of their papers, presentations, publications, data sets, or other scholarly creations. Through this service, Penn State researchers can also comply with grant-funding-agency requirements for sharing and managing research data.

Recently Uploaded

- **Drunkenness, Gender, and...**
  - Danielle Janine Driscoll
  - ACURA Poster 2014 .pdf
  - crime, London, 1885, Gender, drunkenness, Thomas Saunders, Franklin Lushington, Thames Police Court Records

- **Rethinking Medical...**
  - Esther Dell
  - Rethinking Medical Edu...
  - CEO Perspectives, Penn State Hershey, Penn State Medicine

- **Personalized Medicine.pdf**
  - Esther Dell
  - Personalized Medicine.pdf
  - CEO Perspectives, Penn State Hershey, Penn State Medicine

**Contribute**

- Share Your Work
- Terms of Use

**My Latest**

Login to see your recently added documents

- Low-barrier, self-deposit
- Four required metadata fields
- Versioning
- Provenance tracking
- Citable, persistent URLs
- Flexible permissions/access/visibility
- No storage maximums
Uses of ScholarSphere include -

- Supplemental files for electronic theses and dissertations (ETDs)
- Compliance with NSF, NIH, and other funding agency requirements
- “Legacy” data sets - i.e., “I need to get these off my hard drive” data files
- Data sets linked to journal articles (publisher requirement)
- Student research (“scholarship as pedagogy”)
- Service / platform to help diversify what Libraries collect
Cleveland and Ammon, 2013: Relocated earthquakes map

Abstract or Summary

Precise relative earthquake location using surface waves, Journal of Geophysical Research-Solid Earth, 118(6), 2893-2904. Earthquake locations provide a fundamental tool for seismological investigations. While dense seismic networks can provide robust locations, accuracy and precision of these locations suffer outside dense networks. This is particularly true in offshore areas, where location analysis relies heavily on distant seismic observations. We present a method for estimating precise relative seismic source epicentral locations using surface waves. Several reasons, including lower velocities and strength of the signal at distance, make use of surface waves for event location appealing. We focus on the Panama Fracture Zone region and relocate 81 strike-slip earthquakes to produce tectonically consistent epicentral locations. The resulting pattern of earthquakes more clearly delineates recently active regional structures than original body-wave locations. The mean shift between the US Geological Survey National Earthquake Information Center epicenter and our epicentroids is about 14km (the median is about 11km), and typical origin time changes are generally less than 2s. We find that north of 6.5 degrees N, the plate boundary motion is split across two roughly north-south striking structures, the Panama and Balboa Fracture zones. For the last 36 years, slip along these two structures roughly matches slip along the Panama Fracture Zone to the south (from 4.5 degrees N to 6.25 degrees N), but the Balboa Fracture zone has roughly three times the moment than the northern Panama Fracture Zone. Our analyses show that observed Rayleigh-wave signal-to-noise ratios for moderate-to-large shallow earthquakes are suitable for applying the procedure and that Rayleigh-wave observations form a self-consistent set of constraints on the relative location of earthquake centroids.
Penn State Hershey Medical Center Nursing

Publications, Poster and Podium Presentations, Other Scholarly Activities, Invited Presentations, Symposia, Panels, and Funded Research

### Actions

### Descriptions

<table>
<thead>
<tr>
<th>Title</th>
<th>Penn State Hershey Medical Center Nursing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>Sonja Heisey</td>
</tr>
<tr>
<td></td>
<td>Penn State Hershey Nursing Department</td>
</tr>
<tr>
<td>Abstract or Summary</td>
<td>Publications, Poster and Podium Presentations, Other Scholarly Activities, Invited Presentations, Symposia, Panels, and Funded Research</td>
</tr>
<tr>
<td>Date modified</td>
<td>2014-06-23</td>
</tr>
<tr>
<td>Date uploaded</td>
<td>2013-11-12</td>
</tr>
</tbody>
</table>

Sort By: relevance ▼  Show 10 ▼ per page  ▼ Refresh

#### Items in this Collection

1. **Because I Understand What You're Going Through: Patient/Family Mentor Program in Adult Patient Population**

   ![Image](image_url)  
   **Title:** Because I Understand What You’re Going Through: Patient/Family Mentor Program in Adult Patient Population  
   **Depositor:** Sonja M Heisey  
   **Creator:** Chris Ewing  
   **Description:** The Patient/Family Mentor Program (P/FMP), adopted by the Patient-and Family-Centered Care Advisory Council (PFCCAC) at Penn State Hershey Medical Center, is based on actual patient s’ experiences of
We aimed for breadth, rather than depth, initially.

Because we knew we’d be evolving and improving the service continually. And we are.
- 10 releases since 2012
- Next release will be ScholarSphere 2.0 in fall 2014: new UI/UX
- Expectation management is a thing!
How engagement is paying off

- Inquiries from researchers looking for help with managing their data >> new partners.
- Subject specialist inquiries about the service >> more faculty and student users.
- Inquiries from departments to use service for student scholarship >> more content.
- Feature requests from users >> improved service.
- Librarians + researchers understand better the importance of data & data curation >> opportunities for infrastructure, collections, publishing, partnerships, outreach, new roles.
Current work, next steps

- **Current:** ScholarSphere Users Group and incremental feedback on UI/UX work
  - Usage statistics
  - More integration with cloud services
  - Marketing and promotion / outreach (ongoing)

- **Next:** mediated deposit service, “work” data model, different metadata templates, citation support (DOIs), notification framework, Zotero
Suggested essentials- 1

- Find out who your collaborators will be.
  - Who do you want to collaborate with?
- Learn the technology infrastructure at your library and institution.
- Be your own use case.
- Learn the basics of project management.
- Aim to work in partnership with researchers.
Discover new mentors.

Definitely seek out training that skills you up - problem-based.

- Also: MANTRA, online data mgmt course.

Looking for examples of digital curation in practice? There’s an active Google group for this.

Cultivate a sense of experimentation & of play by trying out tools and applications new to you.
Thank you!

Keep in touch: phswe@psu.edu