Priorities for Data Curation Education:

Data Center Partnerships & Long-Tail Science

Carole L. Palmer
Center for Informatics Research in Science & Scholarship
Graduate School of Library & Information Science
University of Illinois at Urbana-Champaign

Microsoft eScience Workshop
9 October 2012
Overview

➤ Background: co-evolution of education and research informed by LIS fundamentals

➤ Lessons from student field experience program built with research partners

➤ Building knowledge base on long-tail science
  - emphasis on re-use value

➤ Trends in student placement
Education programs informed by research

- **Information and Discovery in Neuroscience (NSF)** 2006-09
- **Data Curation in the Sciences 2006-11**
- **Curation Profiles Project (IMLS)** 2008-12
- **Data Conservancy in Research Centers 2010-**
- **Socio-Technical Data Analytics 2012-Cathy Blake, PI**
- **Data Curation in the Humanities 2008-12**
- **Summer Institutes in Data Curation 2008-11**
- **Data Conservancy Education Initiatives 2009-12**

**Participants:**
- Biological Information Specialists
- Cathy Blake, PI

**Partners:**
- NSF
- IMLS
- Data Conservancy (NSF)
- Institute of Museum and Library Services
Underpinned by LIS principles and core expertise

The “true essence” of the profession...is
“the maximization of the effective use of graphic records.”
(Shera, 1971, p. 57)

- **add value** to information to improve current use and potential for future use  
  (Taylor, 1986)

- alignment with complex **social structures and practices**  
  (Shera, 1972)

LIS core – collect, preserve, and provide **access for user communities**

- information behavior
- representation and retrieval of content
- collection and service development and management

(Palmer, Renear, Cragin, 2008)
Emphasis on LIS “metascience” responsibilities

(Bates 1999)

- Provide access within and across disciplines
  in the tradition of research libraries, union catalogs, bibliographies of bibliographies, national libraries

- Promote sharing and interoperability
  across institutions and fields of research

But, supporting data intensive and Interdisciplinary research requires

a larger “ecology” of collaborating institutions and professionals.

(Smith, 2010; Parsons & Fox, 2012)
<table>
<thead>
<tr>
<th>Internship Sites</th>
<th>Practicum Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>* National Center for Atmospheric Research</td>
<td>Oxford Internet Institute</td>
</tr>
<tr>
<td>* National Snow and Ice Data Center</td>
<td>SLAC National Accelerator Laboratory Archives and History Office</td>
</tr>
<tr>
<td>Woods Hole Oceanographic</td>
<td>** Field Museum of Natural History</td>
</tr>
<tr>
<td>National Library of Medicine</td>
<td>Institute for Advanced Technology in the Humanities</td>
</tr>
<tr>
<td>National Agriculture Library</td>
<td>* MD Institute for Tech in the Humanities</td>
</tr>
<tr>
<td>** Smithsonian Institution Digital Services</td>
<td>Northwestern University Library</td>
</tr>
<tr>
<td>Smithsonian Institution Archives</td>
<td>Center for Multimedia Excellence, Illinois</td>
</tr>
<tr>
<td>* Johns Hopkins Library</td>
<td>* University of Illinois Library</td>
</tr>
<tr>
<td>* Purdue Distributed Data Curation Center</td>
<td>IDEALS, Institutional Repository, Illinois</td>
</tr>
<tr>
<td>* Brown University Women Writers Project</td>
<td></td>
</tr>
<tr>
<td>State Historical Society of North Dakota</td>
<td></td>
</tr>
</tbody>
</table>

* = Research partners
** = Project advisors
Data Curation Education in Research Centers (DCERCC)

Model for graduate education:

• Shared core masters curriculum & intensive workshop

• Field experiences in science data centers

masters students – 7 week internship
doctoral students – 2 semesters

Data Mentors and Science Mentors
Evaluations of core course, workshop, and internships:

- Strongly positive feedback from students and mentors
- Evidence of reciprocity

Areas for development:

Increase student preparation for data-intensive environment.

- earlier internship project planning
- more hands-on experience working with data
- additional experience in academic scientific settings

Build long-term partnerships, also integrating academic atmospheric science
Qualitative studies informing curriculum

- long tail - complex, heterogeneous data
- re-use value across disciplines
- implications for curation of research data

Illinois Data Practices team

Doctoral students:

Nic Weber
Tiffany Chao
Karen Baker
Andrea Thomer

Collaborator:
Melissa Cragin

Promoting data preservation and re-use across disciplines.
Emphasis on the long / “big” tail

12,025 NSF grants awarded in 2007 = $2,865,388,605

<table>
<thead>
<tr>
<th></th>
<th>Range</th>
<th>Number of Grants</th>
<th>Total dollars</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$300,000 - $38,131,952</td>
<td>2405</td>
<td>$1,747,957,451</td>
</tr>
<tr>
<td></td>
<td>$579 - $300,000</td>
<td>9621</td>
<td>$1,117,431,154</td>
</tr>
</tbody>
</table>

(Heidorn, 2009)
Earth & life science case studies

Oceanography
Climate science - modern
Climate science - paleo
Soil ecology
Volcanology
Stratigraphy
Mineralogy
Microbiology
Sensor network science
Environmental engineering
Photonics

Curation Profiles Project
2007-2009

Anthropology
Plant sciences
Kinesiology
Speech and Hearing
Earth and Atmospheric

earth and life science intersection
## Utility for producers – compound units

<table>
<thead>
<tr>
<th>Data unit</th>
<th>Geobiology</th>
<th>Volcanology</th>
<th>Soil ecology</th>
<th>Sensor science</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Site-specific time series:</strong></td>
<td><strong>Rock profile:</strong></td>
<td><strong>Database:</strong></td>
<td><strong>Database:</strong></td>
<td></td>
</tr>
<tr>
<td>- spreadsheets averaged rock, water chemistry measures</td>
<td>- physical rock</td>
<td>- multiple abiomatic soil measures</td>
<td>- soil data</td>
<td></td>
</tr>
<tr>
<td>- microscopy images</td>
<td>- thin section</td>
<td>- associated metadata</td>
<td>- sensor data</td>
<td></td>
</tr>
<tr>
<td>- annotated field photo</td>
<td>- chemical analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- microbial genomic data</td>
<td>- photographs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- field notes</td>
<td>- field notes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Rock profile:
- physical rock
- thin section
- chemical analysis
- photographs
- field notes

### Database:
- multiple abiotic soil measures
- associated metadata

### Sharing conventions
- by request
- no repository
- by request
- no repository
- public resource collection
- Reference data
- Limits – “vertical” dev.
Utility for reuse – components of compound units

...somebody more knowledgeable about isotopes can take the data that I produced and do a whole different series of investigations.

...there are people who might work on little iron and titanium oxides which I don’t really care about.

...there’s a lot of geochemical work that’s done that relies less on field context.
### User communities

<table>
<thead>
<tr>
<th>Designated community</th>
<th>Geobiology</th>
<th>Volcanology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Microbiology</td>
<td>Igneous petrology</td>
</tr>
<tr>
<td></td>
<td>Geobiology</td>
<td>Geophysics</td>
</tr>
<tr>
<td></td>
<td>Geology</td>
<td>Geochemistry</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Potential communities</th>
<th>Geobiology</th>
<th>Volcanology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chemistry, Evolutionary biology</td>
<td>Glaciology</td>
</tr>
<tr>
<td></td>
<td>Bioprospecting</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U.S. Park Service</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Public Health</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reuse applications</th>
<th>Geobiology</th>
<th>Volcanology</th>
</tr>
</thead>
<tbody>
<tr>
<td>(parts of unit)</td>
<td>Microbial data - assess presence and extent of disease</td>
<td>Field photos – assess spacio-temporal glacier change over time</td>
</tr>
</tbody>
</table>
Value and use

“A classic example is the NSIDC glacier photo collection, which 10 years ago no one had heard of, and no one thought was worth digitization. It is now NSIDC's 2nd most popular data set.”

(Ruth Duerr, National Snow & Ice Data Center)

How do we predict what data will become highly valuable?

“The value of data increases with their use.” (Uhlir, 2010)

How do data gain in value through use?
Value indicators

Climate / Ocean modeling
Soil Ecology
Volcanology
Stratigraphy
Sensor and Network Engineering

- Reputation of data collector
- Spatial coverage
- Longitudinal coverage *
- Site factors:
  - unique conditions*, rarely studied,
  - politically volatile*, permitting requirements*
- Multiple sources for triangulation and context *
- Documentation of workflows and provenance

* associated with Systems Geobiology at YNP
Yellowstone National Park
Mecca for data collection in systems geobiology.
Research questions from origin of life on Earth to life on other planets.

Collaborators:
- Bruce Fouke, U of I, Geology, Microbiology, Genomic Biology
- Ann Rodman, National Park Service
- Sayeed Choudhury, Data Conservancy

Research on policy and curation processes feeding into education:

LIS – site-based curation, complement to work of repositories
Geobiology – curation principles for undergrad and graduate curriculum
YNP – build awareness among YNP scientists
Specialization in Data Curation placements

49/55 students, 2008 to date

- 33% - Research libraries & museums – LC, Newberry, Chicago Art Institute
- 20% - Research / data centers - USGS, ISGS, WHOI, NSIDC, MITH
- 20% - Industry – Adobe, Industrial Data Associates, Am. Health Information Management, Computer Science Corp, Byte Managers

Sample position titles:

- Data Curator
- Data Management Consultant
- Research Data Librarian
- Data Analyst
- GIS Specialist
- Digital Asset Manager
- Digital Curation Librarian
- Digital Preservation Librarian
- Science Librarian
- Information Architect
clpalmer@illinois.edu

Center for Informatics Research in Science and Scholarship