Microsoft®
Research

Each year Microsoft Research hosts hundreds of influential speakers from around the world including leading scientists, renowned experts in technology, book authors, and leading academics, and makes videos of these lectures freely available.

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

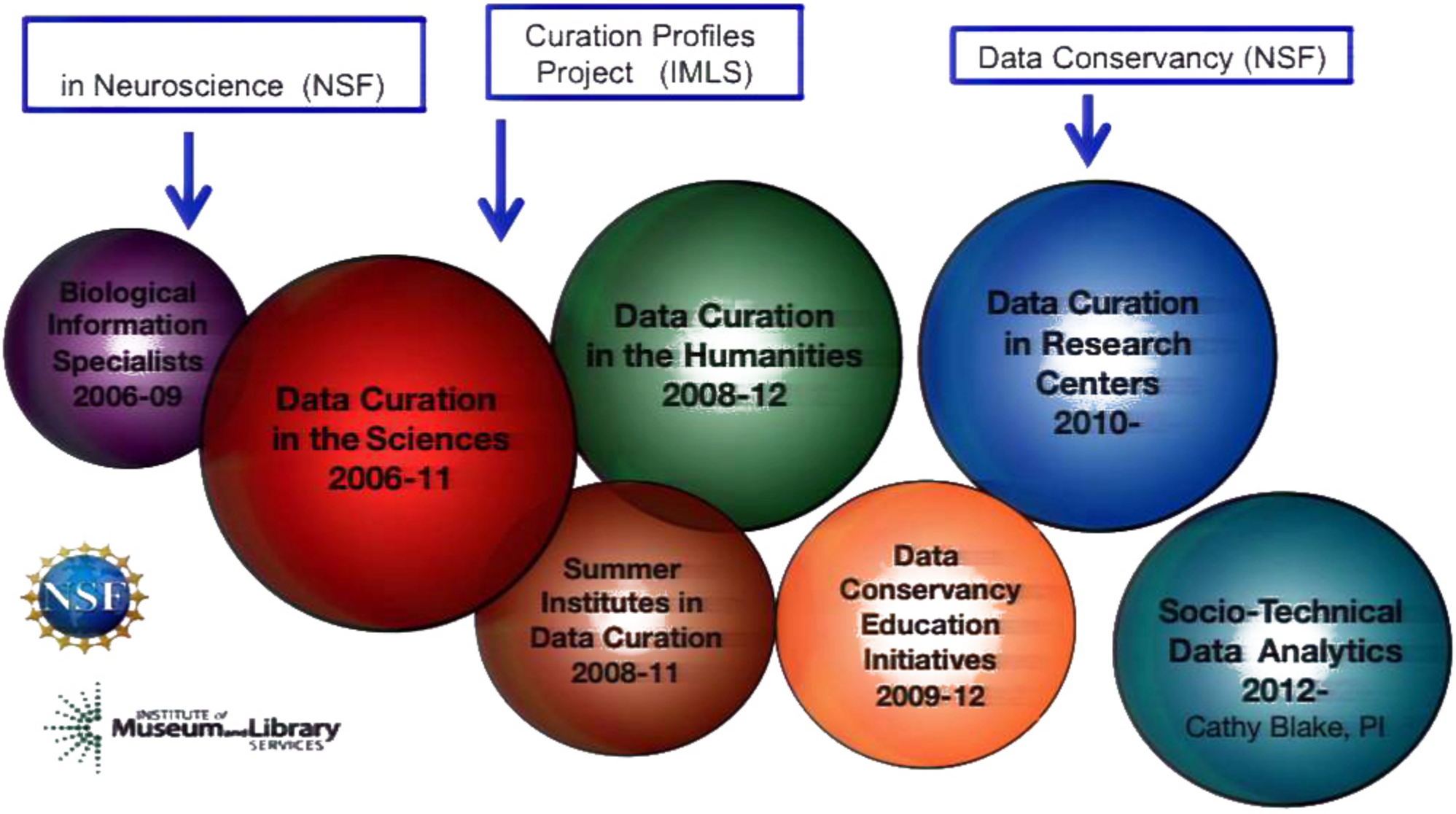
Overview

- 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



Underpinned by LIS principles and core expertise

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)

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)

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)

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)

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

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

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)

Leveraging institutional partners

Internship Sites	Practicum Sites
* National Center for Atmospheric Research	Oxford Internet Institute
* National Snow and Ice Data Center	SLAC National Accelerator Laboratory Archives and History Office
Woods Hole Oceanographic	** Field Museum of Natural History
National Library of Medicine	Institute for Advanced Technology in the Humanities
National Agriculture Library	* MD Institute for Tech in the Humanities
** Smithsonian Institution Digital Services	Northwestern University Library
Smithsonian Institution Archives	Center for Multimedia Excellence, Illinois
* Johns Hopkins Library	* University of Illinois Library
* Purdue Distributed Data Curation Center	IDEALS, Institutional Repository, Illinois
* Brown University Women Writers Project	* = Research partners
State Historical Society of North Dakota	** = Project advisors

Data Curation Education in Research Centers (DCERC)

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

NCAR

THE UNIVERSITY of TENNESSEE KNOXVILLE

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**

INSTITUTE of Museum and Library SERVICES

DCERC assessment after 2 years

DCERC assessment after 2 years

Evaluations of core course, workshop, and internships:

Strongly positive feedback from students and mentors
Evidence of reciprocity

DCERC assessment after 2 years

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

DCERC assessment after 2 years

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

DataConservancy

Preserve * Share * Discover

PI – Sayeed Choudhury

JOHNS HOPKINS UNIVERSITY

Promoting data preservation and re-use across disciplines.

Illinois
Data Practices team

Doctoral students:

Nic Weber
Tiffany Chao
Karen Baker
Andrea Thomer

Collaborator:

Melissa Cragin

Cornell University NCAR MBL Biological Discovery in Woods Hole DURASPACE UCLA Tessella PORTICO NSIDC

DataConservancy Preserve * Share * Discover

PI – Sayeed Choudhury

JOHNS HOPKINS UNIVERSITY

Promoting data preservation and re-use across disciplines.

Illinois Data Practices team

Doctoral students:

Nic Weber
Tiffany Chao
Karen Baker
Andrea Thomer

Collaborator:

Melissa Cragin

Qualitative studies <u>informing curriculum</u>

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

Cornell University NCAR MBL Biological Discovery in Woods Hole DURASPACE UCLA Tessella PORTICO NSIDC

Emphasis on the long / "big" tail

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

Range	$300,000 - $38,131,952	$579 - $300,000
	20%	80%
Number of Grants	2405	9621
Total dollars	$1,747,957,451	$1,117,431,154

(Heidorn, 2009)

Emphasis on the long / "big" tail

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

Range	$300,000 - $38,131,952	$579 - $300,000
	20%	80%
Number of Grants	2405	9621
Total dollars	$1,747,957,451	$1,117,431,154

(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

	Geobiology	Volcanology	Soil ecology	Sensor science
Data unit	**Site-specific time series:** - spreadsheets averaged rock, water chemistry measures - microscopy images - annotated field photo - microbial genomic data	**Rock profile:** • physical rock • *thin section* • chemical analysis • photographs • field notes	**Database:** • multiple abiotic soil measures • associated metadata	**Database:** • soil data • sensor data
Sharing conventions	• by request • no repository	• by request • no repository	• public resource collection	• Reference data • Limits – customization "vertical" dev.

Utility for producers – compound units

	Geobiology	Volcanology	Soil ecology	Sensor science
Data unit	**Site-specific time series:** - spreadsheets averaged rock, water chemistry measures - microscopy images - annotated field photo - microbial genomic data	**Rock profile:** • physical rock • *thin section* • chemical analysis • photographs • field notes	**Database:** • multiple abiotic soil measures • associated metadata	**Database:** • soil data • sensor data
Sharing conventions	• by request • no repository	• by request • no repository	• public resource collection	• Reference data • Limits – customization "vertical" dev.

Utility for producers – compound units

	Geobiology	Volcanology	Soil ecology	Sensor science
Data unit	**Site-specific time series:** - spreadsheets averaged rock, water chemistry measures - microscopy images - annotated field photo - microbial genomic data	**Rock profile:** • physical rock • *thin section* • chemical analysis • photographs • field notes	**Database:** • multiple abiotic soil measures • associated metadata	**Database:** • soil data • sensor data
Sharing conventions	• by request • no repository	• by request • no repository	• public resource collection	• Reference data • Limits – customization "vertical" dev.

Utility for producers – compound units

	Geobiology	Volcanology	Soil ecology	Sensor science
Data unit	**Site-specific time series:** - spreadsheets averaged rock, water chemistry measures - microscopy images - annotated field photo - microbial genomic data	**Rock profile:** • physical rock • *thin section* • chemical analysis • photographs • field notes	**Database:** • multiple abiotic soil measures • associated metadata	**Database:** • soil data • sensor data
Sharing conventions	• by request • no repository	• by request • no repository	• public resource collection	• Reference data • Limits – customization "vertical" dev.

Utility for reuse – components of compound units

Data Type

Field notebooks

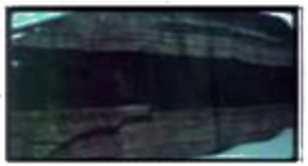
Rock samples

Thin Section slides

Polarized images of Thin Section

Chemical analyses

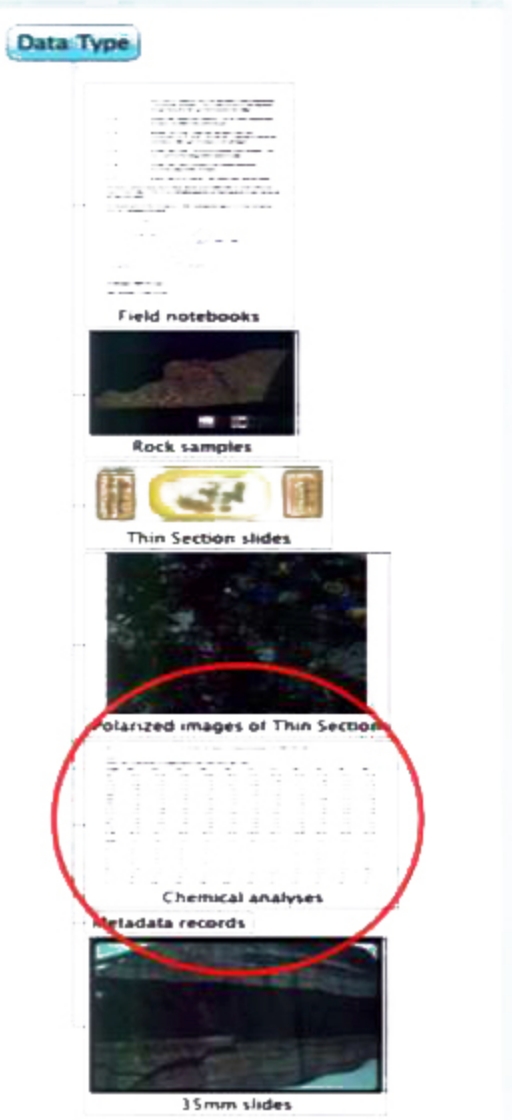
Metadata records



35mm slides

Utility for reuse – components of compound units

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

	Geobiology	Volcanology
	Time series	Rock profile
Designated community	Microbiology Geobiology Geology	Igneous petrology Geophysics Geochemistry
Potential communities	Chemistry, Evolutionary biology Bioprospecting U.S. Park Service Public Health	Glaciology
Reuse applications (parts of unit)	**Microbial data -** **assess presence and extent of disease**	**Field photos –** **assess spacio-temporal glacier change over time**

Value and use

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)

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)

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 triangluation and context*
- Documentation of workflows and provenance

Site-Based Data Curation @ 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

Used with permission from B. Fouke

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