Growing the Curation Community in LIS

Data Curation Education Program & the Data Conservancy

Cultural and organization initiatives that meet the challenges of e-Science

31st Annual IATUL Conference

Purdue University

June 21, 2010

Overview

- Defining data curation
- DataNet initiative: The Data Conservancy
- Data Curation Education Program

- Conversations through the week:
- Implications for growing the LIS curation community
  - New ways of engaging with scientists
  - Managing uncertainties

Data curation is...

The active and on-going management of (research) data through its lifecycle of interest and usefulness to scholarship, science, and education.

Activities

- developing collections
- enable data discovery and retrieval
- maintain data quality
- add value
- provide for re-use over time
- preservation
- archiving

Tasks

- appraisal and selection
- representation
- data integrity
- authentication
- creating and maintaining links
- format conversions
- migration

Data Conservancy

- One of two current awards through the National Science Foundation DataNet program
- Other award is DataONE led by William Michener at University of New Mexico
- Each award is $20 million, 5 year award with multiple partners

NSF DataNet Partners Initiative: Three Primary Goals

- Achieve long-term preservation and access capability in an environment of rapid technology advances.
- Create systems and services that are economically and technologically sustainable.
- Empower science-driven information integration capability on the foundation of a reliable data preservation network.

DC model

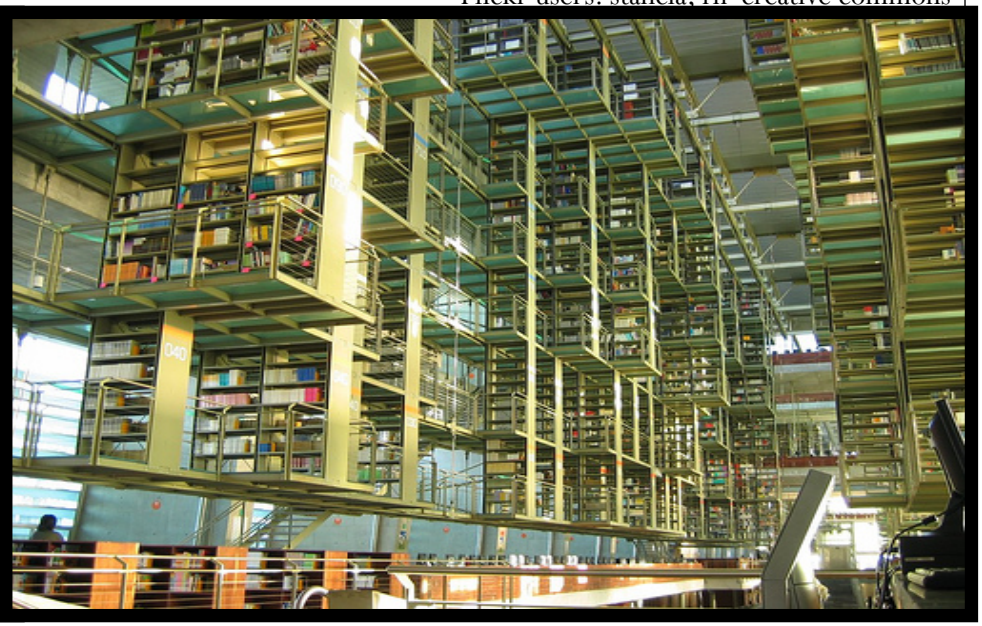
Asserts research libraries as a core component of the emerging distributed network of data collections and services

- Data will be like other collections that have support through research libraries' base budgets.

A shared vision: data curation is a means to collect, organize, validate and preserve data so that scientists can find new ways to address the grand research challenges that face society.

flickr.com/photos/001fi/2907653323/

Flickr users: stancia, rh creative commons



Partner institutions

- Johns Hopkins University (Lead institution)
- Cornell University
- DuraSpace
- Marine Biological Laboratory
- National Center for Atmospheric Research
- National Snow and Ice Data Center
- Portico
- Tessella, Inc.
- University of California Los Angeles
- University of Illinois at Urbana-Champaign

Astronomy as an exemplar scientific community

Achieved notable success in community data standards, practices, documentation, and associated services for research and learning.

Initial goal - ingest astronomy data, connect data to existing services used by astronomers.

** SDSS 140 TB, 3 times that currently held on JHU campus

Demonstrate utility of hosting data in environment that supports existing scientific capabilities in a sustainable manner.

Extend to: **life sciences**
earth sciences
social sciences

Dry Valleys as an exemplar data collection

- 19 years of exploration and data collection - exposed volcanic plumbing!
- Compound data sets organized around vertical "profiles" in rock face; a "profile" consists of:
  - field notes
  - rock sample
  - thin sections (on slides)
  - chemical analysis
  - photographs
  - maps
- Initial data sets in process of ingest – 1 TB

Domain coverage/methods

- Multi-site user research methods are a blend of:
  - Case study & domain comparisons
  - Depth & breadth
  - Local & global

	Astronomy	**Earth Sciences**	**Life Sciences**	**Social Sciences**	
UCAR	Task-based design and usability testing ⇒ Use cases, data requirements, system recommendations				**UCAR**
UCLA	Ethnography, virtual ethnography, oral histories ⇒ Use cases, data requirements	Interviews, Surveys, Worksheets, Content analysis ⇒ Curation requirements, taxonomy, metadata/ provenance framework			**Illinois**

Illinois Data Practices Team – research goals

Comparative analysis of disciplinary differences in data practices to determine varying expectations and needs:

deposition, sharing, and quality control

for participating research communities.

focus on complex, heterogeneous data produced in small science research.

Data Conservancy Broader Impacts

Outreach and education (examples):

- Data curation "boot camps" – one through the UCAR SOARS program undergraduate-to-graduate bridge
- Baltimore area high school students (Sun Microsystems funding, as part of the Data Curation Center of Excellence agreement with DC)
- UCAR-DC, with UCAR staff in U.S. Climate Change Science Program plan and develop outreach/communications strategy on DC outcomes to university, scientific, and citizen stakeholders.
- NSIDC embedding data and earth scientists in DC's IS/CS teams
- providing data curation mentors for students

Data Curation Education Program

Graduate School of Library and Information Science

1. Data Curation Education Program (DCEP) - IMLS/LB, 2006 – Heidorn / Cragin, PI

2. Extending Data Curation to the Humanities (DCEP+) - IMLS/LB - 2008, Renear, PI

- data curation specialization in MSLIS
- continuing education for practicing LIS and IT professionals
- program:
  - curriculum building (for distance option)
  - field work opportunities
  - curation needs assessment

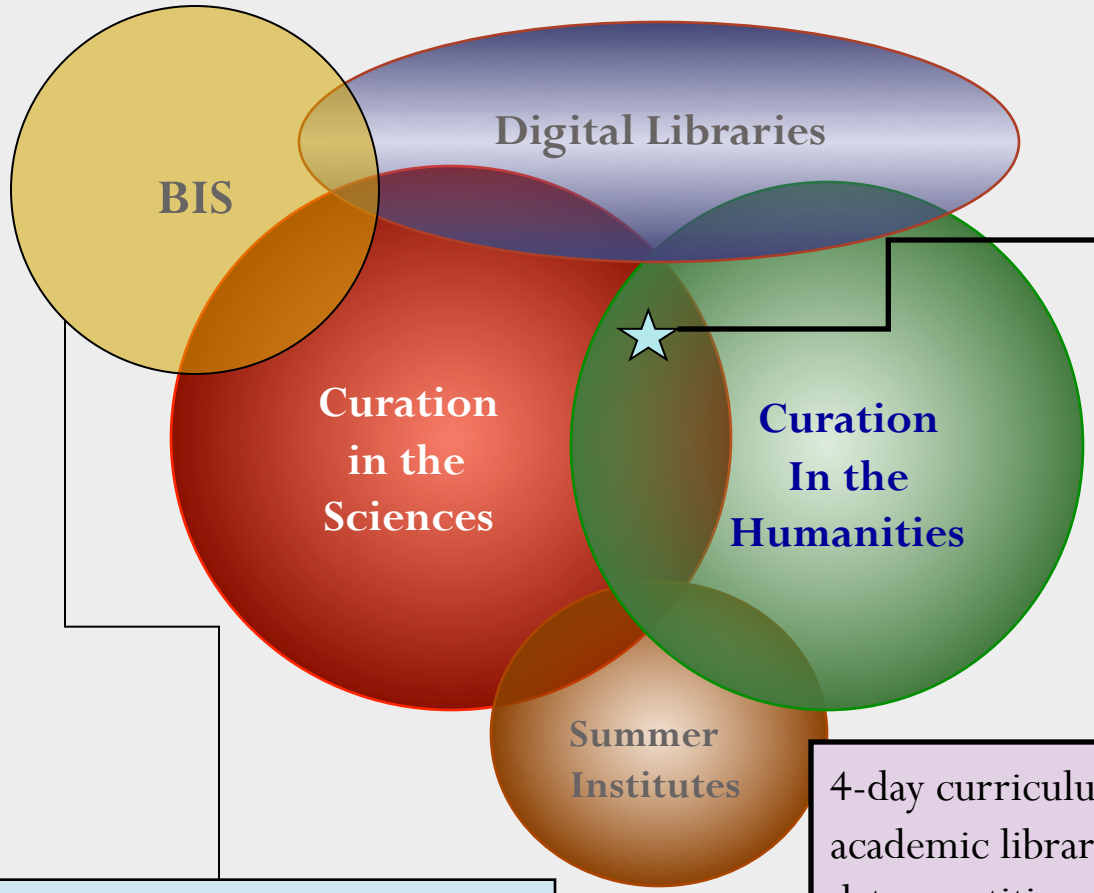
Same foundational principles

The true essence of librarianship…is the maximization of the effective use of graphic records… . (Shera, 1971, p. 57).

- adding value to information to improve current use and potential for future use (Taylor, 1986)
- coordinating and integrating information in alignment with complex social structures and practices (Shera, 1972)

Next-generation information professionals will assist domains in building and maintaining specialized data and information systems

Building a workforce through professional education



Required Core Courses

- Foundations of Data Curation
- Digital Preservation
- Systems Analysis & Mngt

4-day curriculum for practicing academic librarians and other research data practitioners

- Ontology Development
- Intro: Biological Informatics Problems and Resources
- Information Transfer and Collaboration in Science

Core curation content

Foundations of Data Curation

- Digital Data
- Scholarly Communication
- Lifecycles
- Collections
- Selection and Appraisal
- Metadata, Standards & Protocols
- Infrastructures & Repositories
- Archiving & Preservation
- Intellectual Property & Legal Issues
- Workflows; Data Re-use & Value
- Policy & Cooperative Alignments
- Scientific Information Work

Assignments:

Case studies of data infrastructures

Critiques of data management plans

Digital Preservation

- Archival Theory & Diplomatics
- OAIS Reference Model
- Data Formats
- Digital Archival Objects
- Preservation Strategies:
- Emulation vs. Migration
- Authenticity, Integrity & Trust
- Evaluation & Value
- Digital Preservation & The Law

Assignments:

Planning Grant Application

Trusted Repository Assessment

Learning in the field

Internships

- 6 week, funded placements
- project-oriented

  - Digital Research and Curation Center, Sheridan Libraries, Johns Hopkins University (2008)
  - National Agriculture Library, USDA (2009)
  - Smithsonian (2009)
  - Distributed Data Curation Center, Purdue University Libraries (2010)
  - National Library of Medicine (2010)
  - National Snow and Ice Data Center (2010)

Practica

100 hours, course credit

- organizational orientation; shadowing

  Sampling:

  - Nat'l Snow and Ice Data Center (2009)
  - IDEALS (Illinois IR) (2009; 2010)
  - UIC Library / DataOne (2010)

Snapshot of DCEP students

- 42 out of 85 students (signed-up in GSLIS community) report:
  - Background Education: 35 BA ; 7 BS
    - 11 Social Science
    - 20 Humanities
    - 9 Science
- Of the 85 students, previous degrees include: 5 MA / 5 MS / 4 PHD
- Graduates
  - Archivist & Digital Librarian at Center for Black Music Research, Columbia College
  - Enrolled in the Illinois Bioinformatics Master's program
  - GIS Specialist, Illinois State Geological Survey
  - Manager, Public library

Summer Institute on Data Curation:

Extending the DC Curriculum to Practicing LIS Professionals

1st Summer Institute on Data Curation (**focus on scientific data**):

scoping digital data; data integrity & authenticity; appraisal and selection; preparation for ingest; digital preservation standards and day-to-day preservation work; repository architectures

2nd Summer Institute (**focus on humanities and textual data**):

metadata; XML/TEI text encoding; format and encoding management; institutional repository systems; digital preservation; management of versions and provenance

3rd Summer Institute (**focus on Earth and environmental sciences**):

current curation problems in these fields; data standards; GIS and LIDAR data; technical aspects of data systems; repository development and institutional planning; emerging roles for library professionals

4th – 2011 will **focus on life sciences**

Data Curation Education in Research Centers – DCERC

- $1,256,569 funded by IMLS, Carole Palmer, PI
  - Supporting doctoral students in data curation research at GSLIS

- Program partners
  - UT Knoxville School of Information Sciences (Tenopir and Allard) - supporting Master's students
  - National Center for Atmospheric Research (Marlino)
    - providing year-long field placement for doctoral students and summer internships for master's students

Converging research and service roles in the profession

- Information science researchers, studying how best to provide information to researchers, *and*
  - increasingly partners, collaborators, consultants with scientists and organizations involved in data initiatives.

- Librarians and information professionals, providing access and liaison services to researchers, *and*
  - increasingly involved in studying practices and needs of researchers

THANK YOU

INSTITUTE of Museum and Library SERVICES

RE-05-06-0036-06

OCI-0830976