Data Curation in LIS Education and Libraries

ACRL-STS Panel
Big Science, Little Science,
E-Science: The Science Librarian's Role in the Conversation

Melissa Cragin
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

July 13, 2009

Overview

- Data Curation defined
- Problems and implications
- Education initiatives
- Related research

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

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

Tasks

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

Current problems in curation

Theory, policy, application, practice

- conceptualizing collections
- lifecycles
- selection and appraisal
- continuity of access to usable and useful data
- sustainable service models
- new divisions of labor and new roles
- limited infrastructure
  - technical and human
- resource allocation

Data curation: What's new for libraries?

- engaging with scientists during research production cycles
  - technical infrastructure, research groups
- new service examples
  - supporting data handling and management
  - facilitating data deposition
  - data literacy training and support
- new collaborations with various offices
  - campus IT, Research officers, archives

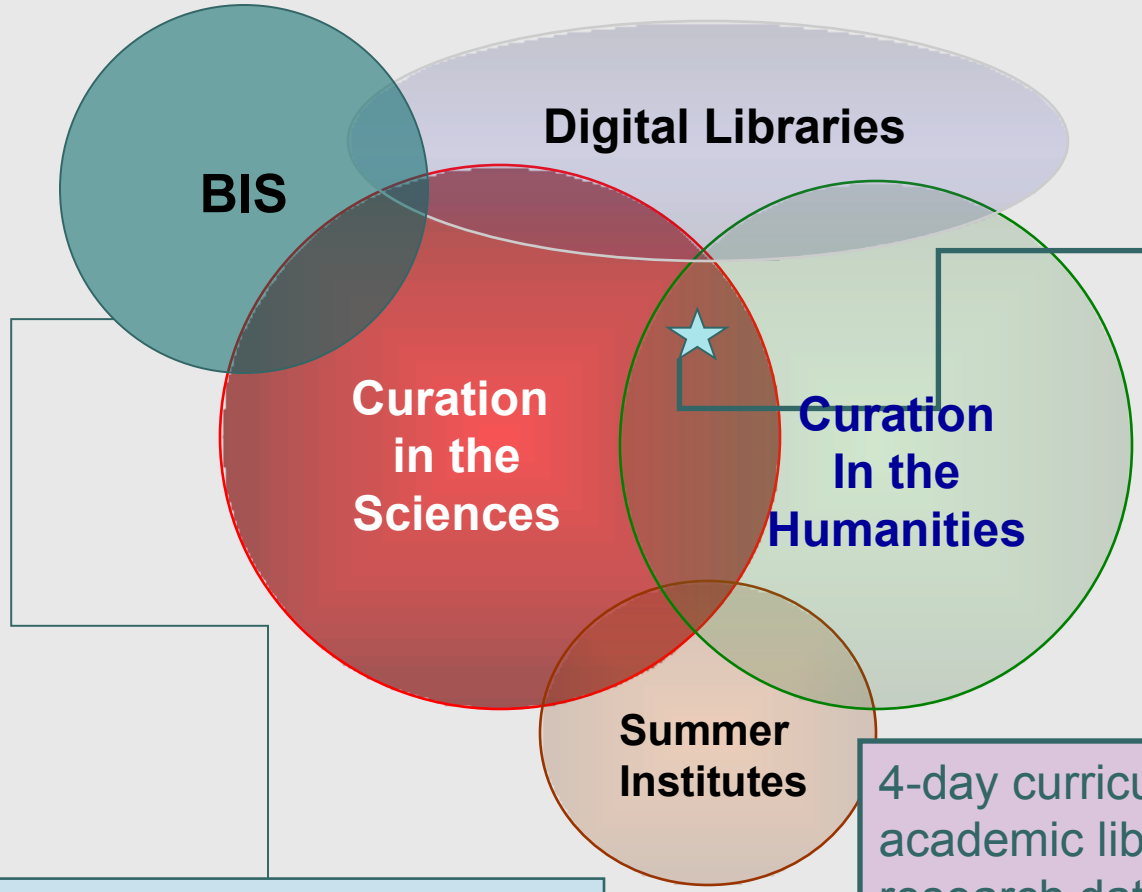
Data literacy

- search and retrieval
- selection
  - assessing quality
  - determining appropriate “fit”
- manipulation
- citation and attribution
- consultation and training

Additional complexities with implications for services

- data complexity
  - variation for what is transformed, when, and how
    - representation of evidentiary value
    - what is shared when…
- engagement with intersecting data communities
- collection forms can highly complex
  - web-based content may need re-construction for ingest into IRs
  - databased collections can have additional, non-trivial requirements
- management of value-added services to meet needs of distinct primary user groups *and* re-use groups

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

Biological Information Specialist

Began in 2006 - part of campus-wide bioinformatics program –
others in CS, crop sciences, animal sciences, chemical & biomolecular eng.

GSLIS only department not focused on computational molecular biology, but biological informatics broadly construed:

> "Tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data. (NIH, BISTI)

Knowledge base: user communities, interoperability, digital preservation, data modeling, ontology development, digital aggregation, information architecture

- Emphasis on data curation and supporting integrative science.

Requires expertise in LIS and research domain.

(NSF/IIS/CISE – 0534567, 2006, Palmer, PI)

GSLIS

MASTER OF SCIENCE IN BIOLOGICAL INFORMATICS

BIS curriculum

Campus core requirement in biology, CS, & bioinformatics.

GSLIS distribution requirement of one course in 3 of 4 areas:

- Information Organization and Knowledge Representation
- Information Resources, Uses and Users
- Information Systems and Access
- Disciplinary Focus

New course offerings:

- Literature-based discovery
- Data mining

Thesis strongly recommended

Skills for Biological Information Specialists

ISs will develop functional applications that are integrated with current science practice, training professionals to support science by building expertise in three areas:

1) Evaluation and implementation of information systems
   - user based assessment and continual quality improvement for the development of tools that work and are used.
2) Information acquisition, management, and dissemination.
   - development of digital libraries, data archives, institutional repositories, and related tools.
3) Information organization and integration
   - structuring information for optimal use and sharing, and standards development.

Data Curation Education Program

1. Data Curation Education Program (DCEP) - IMLS/LB, 2006 - Heidorn, PI
2. Extending Data Curation to the Humanities (DCEP+) - IMLS/LB - 2008, Renear, PI

- Masters concentration in MSLIS, distance option
- Foundation in digital data collection & management, representation, preservation, archiving, standards, policy.
- Emphasis on enabling data discovery and retrieval, maintaining quality, adding value, and providing for re-use over time.

Skills for data curation

- knowledge of scholarly communication processes and how research works
- domain knowledge (or access to it!)
- ability to talk to domain experts, programmers, and technologists
- systems analysis
- ability to track and assess emerging technology
- metadata (incl. disciplinary standards)
- understanding of how databases work
- technical and programming expertise

Core curation content

Foundations of Data Curation

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

Assignments:

20 cases developed this semester

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

Partnerships with research & data centers

Science

- BIRN (Biomedical Informatics Research Network) Maryann Martone
- Smithsonian Libraries, Biodiversity Heritage Library T. Garnett & M. Kalfatovic
- U.S. Geological Survey David Soller
- Marine Biological Laboratory Indra Neil Sarkar
- Missouri Botanical Garden Chris Freeland & Chuck Miller
- Field Museum of Natural History Joanna McCaffrey
- US Army ERDC-CERL General William D. Goran
- Snow and Ice Data Center Ruth Duerr
- Johns Hopkins Libraries Sayeed Choudhury

Humanities

- Perseus Project Greg Crane
- OCLC Lorcan Dempsey
- Women Writers Project, Brown University Julia Flanders
- Unit for Digital Documentation, University of Oslo Christian-Emil Ore
- IATH, University of Virginia Daniel Pitti
- Center for Computing in the Humanities, Kings College Harold Short

Summer Institute on Data Curation:

Extending the DC Curriculum to Practicing LIS Professionals

1st Summer Institute on Data Curation

focus on scientific data topics:

- Digital data
- Data integrity & authenticity
- Appraisal and selection
- Preparation for ingest
- Digital preservation standards
- Day-to-day preservation work
- Repository architectures

2nd Summer Institute: Humanities Data Curation

focus on managing textual data topics:

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

SIDC feedback:
needed skills and requested content

- Policy, management and legal issues
- The Data Interview
  - and then what..?
- Examples of functioning projects and current collaborations
- Meeting reluctance – moving the DC agenda forward
- Metadata
  - hands-on – applications and working with forms

There is so much that is new…

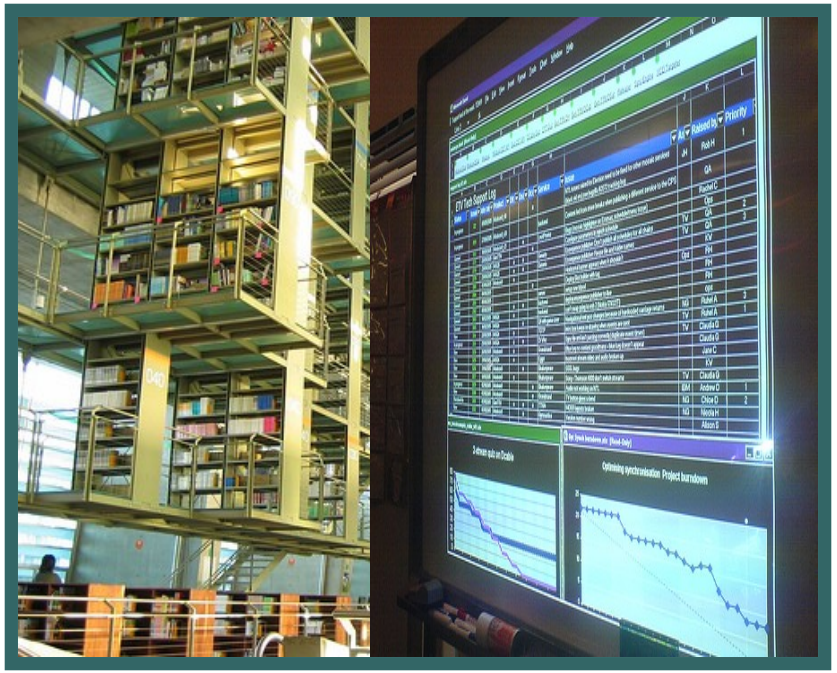
it is critical for research to inform what we teach.

- BECHAMEL Markup Semantics Project
- Digital Collections and Content
- ECHO DEPository
- Quality and Reliability Dynamics
- BIS & DCEP
  - Needs Assessments
- The Information Environments of Humanities Scholars
- BioGeomancer (BG) Project
- Information and Discovery in Neuroscience
- Creation and use of Gene Ontology annotations in model organism databases

Studies within and across domains

research practices & needs → e-research libraries & repositories

Vasconcelos Library Flickr user: rageforst creative commons



Flickr users: stancia, rh creative commons

- How should research data communities be defined for curation purposes?
- What domain differences make a difference for curation requirements?
- How do we aggregate and represent data collections to add value and aid access and use for researchers?

GSLIS Curation Research

Data Curation Profiles Project

(Purdue University Libraries, D. Scott Brandt, PI, IMLS NLG 2007-2009)

In collaboration with librarians, working closely with scientists to study

- research data management / metadata workflow
- policies for archiving and access
- system requirements for managing data in a repository
- librarians roles and skill sets to support archiving and sharing

Data Conservancy (an NSF DataNet award)

(pending, Led by Johns Hopkins Univ. Libraries, Sayeed Choudhury, PI)

Establish a new library-based data cyberinfrastructure paradigm

- partnering with Illinois, UCLA, Cornell, NCAR, MBL, Snow & Ice Data Center

6th International Digital Curation Conference

Chicago, IL
Dec. 7-9, 2010

Digital Curation Centre, UK
and co-hosted by
Graduate School of Library and Information Science

Thank you

cragin@illinois.edu

This work is funded in part by IMLS, grant award # RE-05-06-0036-06.