Illinois Data Practices Research

Carole L. Palmer
Center for informatics Research in Science and Scholarship
CIRSS

Berliner Bibliothekswissenschaftliches Kolloquium
8 June 2010

Data Conservancy

JOHNS HOPKINS UNIVERSITY

PI, Sayeed Choudhury, Sheridan Libraries

Network of domain and data scientists, information and computer scientists, enterprise experts, librarians, and engineers.

Co-PIs and Partners

Carl Lagoze	Cornell University
Mary Marlino	National Center for Atmospheric Research (NCAR)
Carole Palmer	CIRSS, GSLIS, University of Illinois at U-C
Paddy Patterson	Marine Biological Laboratory
Chris Borgman	University of California Los Angeles
Ruth Duerr	National Snow and Ice Data Center
Mark Evans	Tessella, Inc.
Eileen Fenton	Portico
Sandy Payette	DuraSpace / Fedora Commons

Other partners

- Australian National Data Service
- Australian National University
- British Library
- Digital Curation Centre
- Microsoft Research
- Monash University
- Nature Publishing Group
- Optical Society of America
- Sakai Foundation
- Space Telescope Science Institute
- SPARC
- Sun Microsystems (Data Curation Center of Excellence)
- University of Queensland
- Zoom Intelligence

Primary assertion and approach

Research libraries will be a core part of the emerging, distributed network of data collections and services.

Integrated and comprehensive data curation strategy

- to collect, organize, validate, and preserve data
- to address grand research challenges that face society

Infrastructure builds on and connects existing exemplar projects and communities

- deep engagement with scientists
- extensive experience with large-scale, distributed system development.

Astronomy as an exemplar community

Success in data standards, practices, documentation, and associated services

Ingest astronomy data into preservation archive,
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.

Scope to include: life sciences
earth sciences
social sciences

Domain science and library based hubs

Marine Biological Laboratory

- Encyclopedia of Life - taxonomic organization, ontology indexing
- species identification queries for climate change analyses

National Snow & Ice Data Center

- extensive sensor network, fieldwork, aircraft and satellite data
- access node on the DC network, test bed for distributed services

National Center for Atmospheric Research

- civic decision making and climate science in megacities

Cornell University Library

- DataStar - promotes archiving to disciplinary data centers
- arXiv eprints service capability for linking research data with publications

Data Conservancy objectives

Infrastructure	⇒	Technical requirements
Information science Computer science	⇒	Scientific requirements
Sustainability	⇒	Business requirements
Broader impacts	⇒	Educational requirements

- Balance among research and implementation
- Sustainability planning immediate and integral
- Commitment to training and advancing professional workforce

Data framework

Carl Lagoze - Cornell

- Start with a common conceptualization that applies across domains
  -- scientific observation
- Examine, adapt, and adopt existing models
  - National Virtual Observatory
  - Scientific Observations Network (Sonet)

Allen Renear – Illinois

- Define fundamental concepts and identity conditions for collections, data sets, version; preservation targets

Research on data producers and users

	Astronomy	Life Sciences	Earth Sciences	Social Sciences	
NCAR	Task-based design and usability testing ⇒ User cases, data requirements, system recommendations				
UCLA	Ethnography, oral histories ⇒ Use cases, Data reqs.	Interviews, surveys, worksheets, Comparative analysis ⇒ Framework of data characteristics, data practices, & curation requirements			**ILLINOIS**

Illinois data practices research aims

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.

Research questions

- What disciplinary differences in data practices need to be accommodated by DC curation principles and policies?
- How should data be represented, in terms of units, attributes, context, etc. for different functions, such as registry, discovery, interpretation, sharing, and reuse?
- What curation practices and data services can promote better interoperability and federation among research, reference, and resource collections?
- What are the indicators of reuse potential for data sets?

  Conditions or support necessary for reuse?

Progressive data collection

How to get to the point where you're talking shop with scientists about their data, and the right scientists about the right things.

Scientists leading research

- Pre-interview worksheets
- Semi-structured interviews
- follow up sessions with selected participants

Scientists managing data

(post docs, others from labs and research groups)

- Data deposit & sharing worksheet
- Data samples, related documents, vitas, publications

Curation profiling

Supports the ongoing development of system and policy for

Data types:

facilitates workflows and services

pertinent to use within the data community or sub-discipline

- Data communities

guidance on policies for *selection, appraisal, retention, attribution, embargoing . . .*

datacurationprofiles.org

Profile outline

Data kinds and stages - sharing targets, workflow/ provenance, context

Intellectual property - owner(s), stakeholders, terms of use, attribution

Ingest org /description – formal / local standards, documentation

Access - embargo, access control, mirror site

Preservation – duration, migration

Tools - analytical, visualization, integration

Interoperability - needs, APIs, 3rd party data, etc.

Storage, integrity, security - audits, version control

Discovery – browse, search, external

Profiling data complexities & differences

Data Characteristics	Crystallography	Geobiology
Type	1. "Raw data" • Most information rich, long-term value for re-use … 4. "CIF file" – crystallography exchange • Most commonly shared data type	1. "Reduced spreadsheet" – table with average values for multiple observations • Most often requested by others
Intellectual Property/Data Owners	• Service model provide a service to chemists by solving crystal structures • Ownership of the data is ambiguous, and require negotiation before data "hand-off	• Depends on source of funding governmental and private grants, gov. institutions, industry • Ownership of and right to the data range from full to very limited, some long-term "embargoes"
Accessibility	• Field-wide repositories • Many journals require deposit of CIF files • OAI-PMH tools becoming available for CIF files	• Difficult and ad hoc • Well-known researchers receive direct requests for data, often based on publications

Data communities

What are the meaningful social units for organization and use of data over the long term?

Sub-disciplines focused on particular kinds of data that support specific measurements or analysis

Research area where scientists are focused on a research problem, often interdisciplinary in nature

Researchers working to develop and use a shared, community-level data collection (i.e., “Resource Collection”, NSB 2005)

Private vs. public sharing

- Supplying data on request is common
  - current collaborators or close colleagues
  - control and exchange
- Limited experience exposing to wider “publics”
- Attribution problems and misuse incidents highly influential
  - decreased willingness to share
  - increased cynicism about data sharing initiatives
- Most easily or willingly shared is not always the most valuable for re-use

Data Conservancy objectives

Infrastructure	⇒	Technical requirements
Information science Computer science	⇒	Scientific requirements
Sustainability	⇒	Business requirements
Broader impacts	⇒	Educational requirements

- Balance among research and implementation
- Sustainability planning immediate and integral
- Commitment to training and advancing professional workforce

6th International Digital Curation Conference

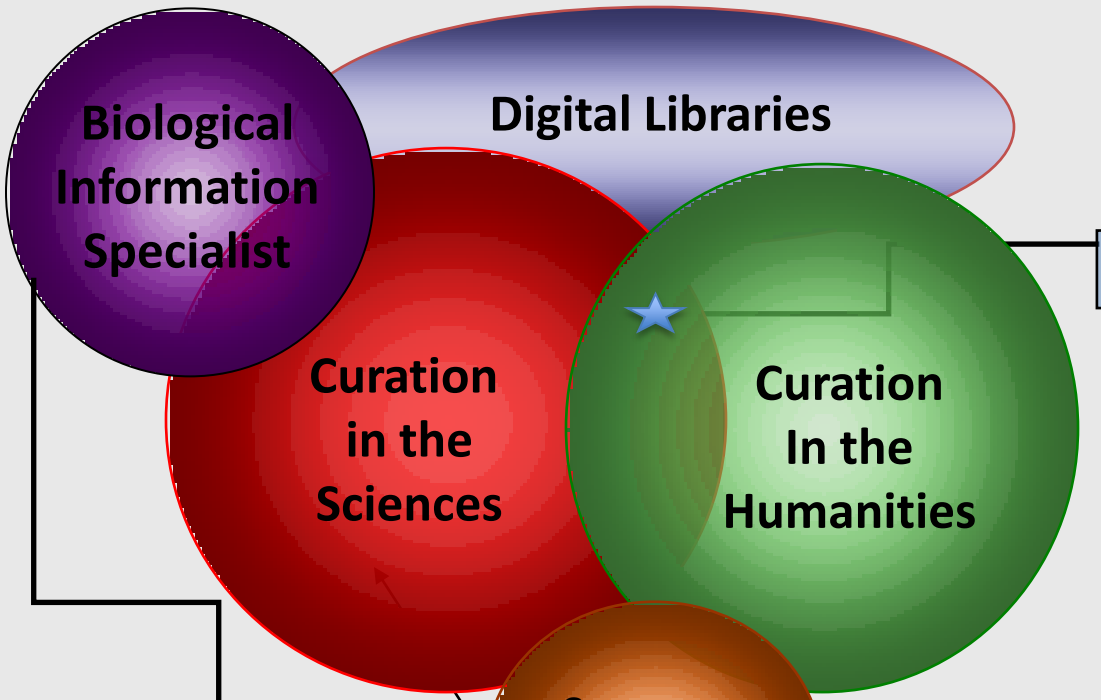
Chicago, IL
Dec. 6-8, 2010

hosted by
CIRSS / GSLIS

in partnership with
Digital Curation Centre, UK

- pre-conference DataNet Education Summit
- post-conference LIS Research Summit

Strategy for building the new workforce



Shared GSLIS Core

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

Campus level bioinformatics masters. Core requirements in biology, CS, & bioinformatics.

Data Conservancy
cases, expertise, best practices

Questions & comments, please

clpalmer@illinois.edu

http://DataConservancy.org

http://cirss.lis.uiuc.edu/

Center for Informatics Research in Science and Scholarship