Data Curation 101

ABCs of Data Curation & Scholarly Communication
Science & Engineering Academic Librarians (SEAL-S) Fall Program
15 October 2010

Jeanine Scaramozzino
College of Science & Mathematics Librarian
School of Education Librarian
Cal Poly State University
Overview

• Managing data & metadata
• Stakeholders
• Ethics, legality & copyright
• Data management plans
• Library roles
• Challenges
Data Curation Defined

“Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarly and educational activities.”

GSLIS, UIUC: http://criss.is.illinois.edu/CrillMeta/docp.html
Why manage data?

- Essential raw material of science
- Demand for transparency of science
- Funder requirements
- Publisher recommendations
Data Management

• Activities
  – Enable discovery & retrieval
  – Maintain data quality
  – Preserve & archive for re-use

• Tasks
  – Appraisal & selection
  – Data integrity
  – Interoperability
What is Data?

- Text (e.g. flat text files, Word, PDF)
- Numerical (e.g. SPSS, STATA, Excel, Acess, MySQL)
- Multimedia (e.g. jpeg, tiff, dicom, mpeg, quicktime)
- Models (e.g. 3D, statistical)
- Software (e.g. Java, C)
- Domain-specific (e.g. FITS in astronomy, CIF in chemistry)
- Instrument-specific (e.g. Olympus Confocal Microscope Data Format)
Size of Data Sets

Relevant References

"Big Science" was originally coined by Alvin M. Weinberg in his 1961 work Impact of Large-Scale Science on the United States, *Science* 134 (3473): 161-164.


Note: "Little Science" is now referred to as "Small Science".
Data Storage Options

- Local
- Campus/Funder/Discipline-based
- Cloud-based
Metadata

• Descriptive information
• Created for the purpose of retrieval & reuse
  – During creation of data
  – During archiving
• Standards
  – Data interoperability
  – Repository
Metadata Standards

• Example Standards
  – Darwin Core (for biology data)
  – Data Documentation Initiative (DDI)
  – Directory Interchange Format
  – ISO 19115:2003
  – PREMIS
  – Science Data Literacy (SDL)
  – Seeing Standards (http://www.dlib.indiana.edu/~jen/ile/metadatamap/seeingstandards_glossary_pamphlet.pdf)
Metadata Quality Issues In Repositories

Examples Include:

- Missing critical fields
- Insufficient information
- Inconsistent field values & defined standards
- Fields with: Not available, n/a, none, & nbsp (space), wrong year, etc.
- Broken links
- Outdated contact information
- Challenges with updating
- etc.
Digital Data Roles & Stakeholders

- Data Authors
- Data Scientists
- Data Managers
- Digital Curators
- Data Users
- Journal Publishers
Different User Communities

- Observational users
- Infrastructure users
- Modelers
- Students
- Scientists
- Policy makers
- General public
- etc.
Ethics, Legality & Copyright

- Data privacy, confidentiality, etc.
- Data disclosure
- Data ownership based on university & granting agency policies
- Data citation in development
  - MIT Libraries
  - Purdue’s Online Writing Lab
  - Dataverse Network Project
Funder Requirements & Data Management Plans

• Data Plans Include:
  – The types of data to be produced
  – The standards that would be applied for format, metadata content, etc.
  – Provisions for archiving & preservation
  – Access policies & provisions
  – Plans for eventual transition or termination of the data collection after funding period

• Resources:
  – SHERPA – JULIET
  – UMN Funding Agencies & Data Management Guidelines
  – Digital Curation Centre Data Management Tool
  – NSF
Data Plan Development

- What form and format is the data in?
- What is the expected lifespan of the dataset?
- How could the data be used, reused, and repurposed?
- How large is the dataset, and what is its rate of growth?
- Who are the potential audiences for the data?
- Who owns the data?
- Does the dataset include sensitive information?
- What publications or discoveries have resulted from the data?
- How should the data be made accessible?

From Purdue University Libraries, Conducting a Data Interview by Michael Witt and Jake Carlson
Why should libraries care about data curation?

- Information Formats Change
- Scientists Generate Vast Amounts of Information
- Library Collections & Services Must Adapt
What knowledge do librarians bring to the table?

- LIS & archival theory
- Collection development, discovery, etc.
- LIS & IT partnerships
- Data management & scholarly communications
- Data organization
Library Roles

• Self-education
• Outreach to scientists
• Provide new services –
  – Research,
  – Resources (repositories & databases)
  – Reference
• Collaborate with campus stakeholders
Data Management Education

• Why Manage Your Data?
• Managing Data
• Back-up Practices
• Ethics, Legality & Copyright
• Funder Requirements & Data Plans
• Science Data & Repositories

CAL POLY LIBRARY SERVICES

CAL POLY LIBRARY SERVICES

20
Managing Data Basics for Data Creators

1. Ensure data is accessible for the long-term, save a copy of data in a non-proprietary, commonly-accessible formats.
2. Keep a records of how data is produced & store it in a text file in the same directory as the data.
3. Use a folder/directory structure with a clear, documented naming scheme.
4. Use naming conventions for files
5. Lots of Copies Keeps Stuff Safe! Keep 3 copies of data in geographically distributed locations.
Data & Libraries

Data & Libraries Overview 2006-2010

Available online at: http://digitalcommons.calpoly.edu/lib_dean/27
Current Challenges in Curation

- Theory, policy, application, practice
- Relationships, data, practices, curation activities
- Emerging models, new divisions of labor and new roles
- Conceptualizing collections, observations, datasets, etc.
- Lifecycles, selection and appraisal
- Continuity of access to usable and useful data, sustainable service models
- Resource allocation, limited infrastructure
Questions?

Jeanine Scaramozzino
jscaramo@calpoly.edu

Data Curation Management Guide
http://libguides.calpoly.edu/data