Research Data in Libraries

Rebecca C. Reznik-Zellen
University of Massachusetts Amherst

Follow this and additional works at: https://escholarship.umassmed.edu/lib_articles

Part of the Library and Information Science Commons

Repository Citation

This material is brought to you by eScholarship@UMassChan. It has been accepted for inclusion in Library Publications by an authorized administrator of eScholarship@UMassChan. For more information, please contact Lisa.Palmer@umassmed.edu.
Outline

1. What is Research Data?

2. Why Manage Research Data?

3. Approaches to Managing Research Data
   1. UMass Amherst & Institutional and Consortial Activities
   2. UMass Medical & Skills and Competencies
Disambiguation Page

- **e-Science, e-Research**
  - A methodology: “shorthand for the set of tools and technologies required to support collaborative, networked science.” (Hey, 2006)

- **Cyberinfrastructure**
  - Techno-social environment that supports data intensive, information intensive research

- **Data management**
  - Data management is the systematic organization and planning for data throughout the research cycle

- **Data curation**
  - Description and selection of data sets for long-term preservation and access

*Research data* is the critical component of each of these concepts
What is Research Data?

“Research data, unlike other types of information, is collected, observed, or created, for purposes of analysis to produce original research results.”

University of Edinburgh Information Services
What is Research Data?

<table>
<thead>
<tr>
<th>Research publication</th>
<th>Research data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information transformed into results</td>
<td>Information not transformed into results</td>
</tr>
<tr>
<td>Use requires basic software and instruments and their command</td>
<td>Use often requires special software and instruments and their command</td>
</tr>
<tr>
<td>Self-explanatory</td>
<td>Requires additional information and documentation if not archived</td>
</tr>
<tr>
<td>Should not include sensitive information</td>
<td>May include sensitive and confidential information</td>
</tr>
<tr>
<td>Use does not require permission</td>
<td>Use often requires permission</td>
</tr>
<tr>
<td>Ownership and copyright often clear</td>
<td>Ownership and copyright often unclear</td>
</tr>
<tr>
<td>Openly accessed by the scientific community for a fee or for free</td>
<td>Several degrees of openness (from completely open to closed)</td>
</tr>
<tr>
<td>Understood as scientific output</td>
<td>At the moment not understood as scientific output/merit</td>
</tr>
<tr>
<td>Ready to be used by others</td>
<td>Use requires processing</td>
</tr>
</tbody>
</table>

From Kuula (2008) Open access to and reuse of research data: the state of the art in Finland, as excerpted in Feijen (2011) What Researchers Want
Research Data Lifecycle

Creating data
- design research
- plan data management (formats, storage etc)
- plan consent for sharing
- locate existing data
- collect data (experiment, observe, measure, simulate)
- capture and create metadata

UK Data Archive:
http://www.data-archive.ac.uk/create-manage/life-cycle
Types of Research Data...

- Observational
- Experimental
- Simulation
- Derived or compiled
- Reference or canonical
Various Formats of Research Data:

• Text
• Numerical
• Multimedia
• Software
• Models

• Can be specific to a discipline: crystallographic information files in chemistry
• Can be specific to instrumentation: Olympus Confocal Microscope Data Format
Research Data May Include:

- Documents
- Lab notebooks
- Questionnaires, transcripts, codebooks
- Audiotapes, videotapes
- Photographs, film
- Test responses
- Slides, artifacts, specimens, samples
- Models, algorithms, scripts
- Methodologies and workflows
- Data files
- Database content
- Standard operating procedures and protocols
Why Manage Research Data?

“Stolen laptop contains cancer cure data”
http://news.cnet.com/8301-17938_105-20028475-1.html
Why Manage Research Data?
Why Manage Research Data?

- Protect federal investment in research and development
- Expedite the scientific process

- Use or reuse
  - the value of the data
  - the uniqueness of the data
  - the importance of the data

- Validate
- Heritage

- Obligation
Funder Requirements

• National Institute of Health (NIH)
  • Data sharing plan for grants >500k
  • Public Access Mandate for published research

• National Science Foundation (NSF)
  • Data management plan for all grant proposals

• Others: USDA, NIST, NASA...
  • See University of Minnesota Data Management Pages
    (http://www.lib.umn.edu/datamanagement/funding)
NSF Mandate

Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled “Data Management Plan (DMP).” This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.

b) “Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.”
NSF Mandate

Key pieces of information to include are:

- Types of data
- Metadata standards to be used
- Policies for access and sharing (including provisions for privacy/intellectual property)
- Provisions for re-use, re-distribution, and production of derivatives
- Plans for archiving and preservation of access

Data Management throughout the Research Cycle

Note to self: Data management is not limited to data sharing!
Facilitating Compliance

Libraries are considering research data as material that falls within their scope of responsibility.

• Why
  • Faculty do not want to do this
  • Office of Research may not want to do this
  • Neither have the expertise/infrastructure to do this
    
    But the Libraries do: it is the libraries’ natural area of expertise

• How
  • By engaging faculty and crafting meaningful services to the extent that they are able
The e-Science Call to Action

e-Science presents unique opportunities to exploit and develop the capabilities of libraries/librarians.

ARL 2007, 2008

- Engaging with researchers
- Being conversant in science subjects
- Understanding nature of research methodologies and how scholarly exchange is communicated
- Understanding archival and life-cycle aspects of scientific information
- Developing standards and systems for digital content
- Data curation and preservation
UMass System Libraries’ Response

• UMASS 5 Science Librarians Ad Hoc Committee for e-Science (2008)

• Activities designed to facilitate e-Science on our campuses
  • Exploring shared electronic resources
  • Drafting Principles fundamental to e-Science
  • Creating Professional Development and Continuing Education opportunities

• A series of events (currently in third consecutive year)
  • Professional development day
  • E-Science symposium
  • Science Boot Camp
UMass System Libraries’ Response

• Some principles fundamental to e-Science
  • Collaboration
  • Curation of Primary Scientific Data
  • Digital Stewardship and Preservation
  • Metadata Standards and Creation
  • Virtual Communities
  • Communication
  • Open Access
  • Professional Development and Investment
UMass Amherst Libraries Approach

- 21,000 undergraduate students
- 6,000 graduate students
- 1,174 full-time instructional faculty
- 51 doctoral and 73 master’s degree programs
- $170 million of sponsored research in FY 2010
- 1,134 awards from 1,294 proposals submitted in FY 2010
- 498 federal awards (42% NSF)
UMass Amherst Libraries Approach

• Established a Data Working Group (DWG) to make explicit recommendations to the Libraries’ regarding research data (2010)

Determine if the University Libraries should accept broad responsibility for curating research data and, if so, how that should be done, what would be expected, and who would be involved.

• Education on the issues involved with data curation
• Understanding of the University’s current research environment and data outputs
• Evaluation of current Library practice for supporting active and archival data
• Exploration of partnerships for data curation both in and outside of the University
• Propose interim steps to assist the University in meeting its own policies
• Create a vision of data curation for the Libraries
Understanding the Current Environment

Faculty Interviews and Graduate Student Focus Group

Heterogeneous array of strategies and infrastructure scenarios, with common themes across disciplines.

| Data storage & infrastructure | • Size  
|                                 | • Backup/storage  
|                                 | • Computational power  
|                                 | • Preservation of proprietary programs/file formats  
|                                 | • Various infrastructure ownership scenarios |
| Procedures & training          | • Workflow routinization  
|                                 | • Knowledge transfer |
| Documentation & metadata       | • Project and discipline-specific practices  
|                                 | • Impact of external requirement on practices |
| Data Reuse & sharing           | • Versioning  
|                                 | • Making data public vs. making data useful  
|                                 | • Collaboration for publication vs. collaborative projects  
|                                 | • Occasional need to reproduce research |
| IP & data sensitivity          | • Privacy/IRB constraints  
|                                 | • USPTO rules for confidential exposure  
|                                 | • Emulation and Post-publication sharing |
Partnerships and First Steps

• Joint letter on NSF mandate with the Office of Research
  • Outlines Services
  • Consultation on Data Management Plans

• Data Management Web Page

• Data Management Plan
  Template
  • Project Overview
  • Data Description
  • Data Storage
  • Access and Dissemination
  • Preservation

http://www.library.umass.edu/data-management/
Institutional Approaches

• Informational/Educational
  • University of Minnesota
  • University of Nebraska

• Consultative
  • MIT
  • University of Wisconsin-Madison

• Technical
  • Purdue University
  • Rutgers
Large-scale Consortial Approaches

• Linking Publications and Data
  • Dryad

• Data Management Planning
  • California Digital Libraries and Partners

• Persistent Identifiers for Data Sets
  • DataCite

• Metadata Schema for Data Sets
  • Data Documentation Initiative
UMass Medical School Approach

- 3 graduate schools: Medicine, Nursing, Biomedical Sciences
- Private and federally funded research grants > $200 million for fiscal year 2009
- Awarded CTSA Summer 2010
- Lamar Soutter Library oversees the National Network of Libraries of Medicine New England Region (RML)
- RML outreach programs: continuing education
Evaluating local needs and competencies

• Conducted a learning needs assessment of New England science and medical librarians for planning continuing education programs and portal in 2009

• Developed an e-Science portal for New England science and medical librarians ([http://esciencelibrary.umassmed.edu](http://esciencelibrary.umassmed.edu))

• Spring 2011 assessment of New England science and medical librarians’ data management competencies
UMMS/WPI Data Management Project

IMLS National Leadership Planning Grant

Objectives

1. Data management curriculum for science and medical/health science students
2. Identify requirements for a data repository
3. Communications plan
What New England libraries are doing now:

• Initiating projects to understand the scope of campus data management needs—education and environmental scanning
• Working with other campus departments to determine best practices for data management
• Developing formal policies for data management support
• Reorganizing library to have department for specialized content
• Conducting data interviews with researchers (this may be done as part of a team with subject librarians)
• Evaluating requirements for an institutional data repository
• Consulting with researchers on data management plans
Skilling Up To Do Data

Pryor and Donnelley. 2009. Skilling Up To Do Data. JIDC 2(4)
Data Management skills/competencies

• Understanding research methods, data lifecycle, data security
• Build, populate, and maintain digital databases
• Use a variety of programming languages (e.g. XML, SQL)
• Knowledge of metadata standards (interoperability standards, Dublin Core, MODS, OAI_PMH, etc)
• Work with metadata manipulation, crosswalk, validation, and portals
• Provide data mining, interpretation, representation, and visualization services
• Work with and develop digital lab notebook applications
• Promote digital data sharing, open access, and/or participation in IR
• Ability to work collaboratively with librarian colleagues, IT, IRB, and faculty
References


Data Curation and Management Competencies of New England Region Health Sciences and Science and Technology Librarians

e-Science Portal for New England Librarians
[http://esciencelibrary.umassmed.edu](http://esciencelibrary.umassmed.edu)


Katz L. *Stolen Laptop Contains Cancer Cure Data*. CNet News 13 January 2011.

Kuula A and Borg S. *Open access to and reuse of research data - The state of the art in Finland*. Finish Social Science Data Archive, 2008.


*Nature* Special Issue: Big Data (2008)
*Nature* Special Issue: Data Sharing (2009)
The Economist Special Issue: The Data Deluge (2010)
*Science* Special Issue: Dealing With Data (2011)


UK Data Archive
[http://www.data-archive.ac.uk/create-manage/life-cycle](http://www.data-archive.ac.uk/create-manage/life-cycle)