The Data Life Cycle and Securing Access to Open Research Data

Rolando Garcia-Milian and Hannah Norton
UF Health Science Center Library
Open Access Week 2012

Image modified from: http://www.fotopedia.com/items/flickr-2194127541
Benefits of sharing data:

• Data can be used by other researchers with different objectives
• Accelerate the time of discovery by building upon previous research
• Results can be reproduced more easily and accurately
• Researchers receive the credit they’re due
• Data producers have a new channel by which to promote their work (increase impact of research)
Panton Principles for Open Data in Science

• Science is based on building on, reusing and openly criticising the published body of scientific knowledge.

• For science to effectively function, and for society to reap the full benefits from scientific endeavours, it is crucial that science data be made open.
Barriers to data sharing

• No time
• Lack of funding
• No place to put it
• Don’t have rights to make data public
• No standards
• Others don’t need my data
Data Lifecycle*

- Study Concept
- Data Collection
- Data Processing
- Data Analysis
- Data Distribution
- Data Archiving
- Data Discovery
- Data Analysis
- Repurposing

* Based on Data Documentation Initiative (DDI) version 3.0 Combined Life Cycle Model
Data Lifecycle*

Study Concept → Data Collection → Data Processing → Data Analysis

Data Analysis → Data Distribution → Data Discovery → Data Archiving

Repurposing

Based on Data Documentation Initiative (DDI) version 3.0 Combined Life Cycle Model
Data Lifecycle*

* Based on Data Documentation Initiative (DDI) version 3.0 Combined Life Cycle Model
Why do researchers need a DMP?

• To ensure that valuable data resources will be accessible in the future to members of the research team and the broader community.

• To make life easier – by planning ahead and documenting data throughout its life cycle, researchers can save time and focus on research.

• To increase the visibility of research.

• To satisfy funders’ requirements.
Funders’ requirements

• “…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’. This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results.”

• “The NIH expects and supports the timely release and sharing of final research data from NIH-supported studies for use by other researchers. Starting with the October 1, 2003 receipt date, investigators submitting an NIH application seeking $500,000 or more in direct costs in any single year are expected to include a plan for data sharing or state why data sharing is not possible.”
DMP Templates and Tools

Templates can give you a place to start, as long as you customize them for your project.

- UF HPC Center links:
  http://www.hpc.ufl.edu/proposals/

- https://dmp.cdlib.org/
National Institutes of Health: 1. Data sharing plan

Investigators seeking $500,000 or more in direct costs in any year should include a description of how final research data will be shared, or explain why data sharing is not possible.

Help

The precise content of the data-sharing plan will vary, depending on the data being collected and how the investigator is planning to share the data. Questions to consider as appropriate:

- When will you make the data available?
- Which file formats will you use for your data, and why?
- What transformations (to more shareable formats) will be necessary to prepare data for preservation/data sharing?
- What metadata/documentation will be submitted alongside the data or created on
Who are you willing to share your data with?

- Immediate collaborators: 95.8%
- Others in my department or institute: 35.4%
- Others in my field: 35.4%
- Others outside of my field: 16.7%
- Anyone: 6.3%

n=48
How are you sharing or planning to share your data?

- Depositing them in a discipline-specific data center or repository: 26.0%
- Submitting them to a journal to support a publication: 68.0%
- Depositing them in UF’s Institutional Repository (http://ufdc.ufl.edu/ir): 4.0%
- Making them available online via a project or institutional website: 22.0%
- Making them available informally to peers on request: 46.0%
- I do not share data: 10.0%

n=50

Percentage of Respondents
Securing Access to Open Research Data
Considerations When Submitting Research Data To An Open Repository

- Subject/research domain

- File format and data structure

- Metadata/ Standards

- The repository adds Digital Object Identifiers (DOIs) / permanent links

  [http://dx.doi.org/10.5061/dryad.585t4](http://dx.doi.org/10.5061/dryad.585t4)
  [http://ufdc.ufl.edu/IR00000452/00001](http://ufdc.ufl.edu/IR00000452/00001)

- Repository have the technical capacity to embargo or sequester access to data until the content has been approved for release to the public.
Excel file illustrating the structure of a GEO archive Affymetrix data submissions.
Data Repositories With Associated Visualization And Analysis Tools

**PhysioToolkit** is a library of software for physiologic signal analysis, and detection of physiologically significant events.

The **Web Image Browser** (Open Cell Centered Database) for viewing and annotating images. WIB reads and manages images similar to Google Maps.

**GEO2R** compares groups of samples in order to identify differentially expressed genes across experimental conditions.
What repositories are appropriate for a researcher to submit the data to?
“Research Data Management at UF” Guide

http://guides.uflib.ufl.edu/datamanagement
“Research Data Management at UF” Guide

DMP Tool
Guidance and Resources for your Data Management Plan

- Create ready-to-use data management plans for specific funding agencies.
- Use your gatorlink and password to login and use this tool.
- For training on the use of the DMP Tool, please contact us.

For individual or group consultation/training on how to prepare your Data Management Plan, please contact us.
"Research Data Management at UF" Guide

**UF HPC Center Storage**

High-Performance Computing (HPC) Center offers different options in terms of storage:

**Home storage:** Limited to 10 GB or less. Under no circumstances is home storage to be used for jobs running on the HPC Center systems (this will result in the suspension of your account).

**Scratch storage:** this is the primary source and destination for jobs running on the HPC Center cluster(s). Every user has access to a nominal amount of storage (500 GB) with no additional fees or charges. Data in scratch storage older than thirty (30) days and not associated with a running job may be automatically purged on a daily basis.

**Project storage:** is intended to be longer-term but less performant than scratch storage.

Please, contact the HPC Center for more information on how to have access to these storage options and costs.

Note that the HPC Center does not create or maintain backup copies of any user data.

**Comments (0)**

**Sharing Data at UF**

**IR@UF** The Institutional Repository at the University of Florida (IR@UF) "encourages university units to contribute their open access research, reports and other materials (e.g. journal articles, conference proceedings, white papers, audio, video, photographs, presentations, etc) to the IR@UF for archiving and dissemination free of commercial cost”

**Copyright at UF**

According to UF Intellectual Property Policy:

"University personnel are required to record all research data and information accurately and clearly and to keep all such data in a permanent and retrievable form [...]

"Personnel who leave the University may be permitted to copy their laboratory notebooks and take the copies with them" they are required to maintain the confidentiality of the data contained within the notebooks.

For copyright question, please contact **Office of Technology Licensing** and/or Christine Fruin (352) 273-2710 Library West Room 256.

**Secure File Sharing Service at UF**

**File-Express** allows any individuals with UF Gatorlink accounts to securely exchange files that may be too large, or otherwise blocked by traditional e-mail methods. Using a secure server and the Gatorlink Authentication system, users can easily share files with members of both the UF and Non-UF community.

**Maximum single file size:** 5GB

**Default expiration length:** 1 day(s)

**Maximum expiration length:** 5 day(s)

**Comments (0)**
Training on Best Practices in Data Management

<table>
<thead>
<tr>
<th>Workshop Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title:</strong> Best Practices in Data Management</td>
</tr>
<tr>
<td><strong>Provider:</strong> HSC Library</td>
</tr>
<tr>
<td><strong>Topic:</strong> Other Information Resources</td>
</tr>
<tr>
<td><strong>Duration:</strong> 1.0hrs</td>
</tr>
<tr>
<td><strong>Audience:</strong> Any UF</td>
</tr>
<tr>
<td><strong>Handout Link:</strong> View</td>
</tr>
<tr>
<td><strong>Online Training Link:</strong> View</td>
</tr>
<tr>
<td><strong>Prerequisites:</strong></td>
</tr>
<tr>
<td><strong>Detailed Description:</strong> Learn basic, practical strategies for best managing your research data. A number of U.S. funding agencies such as the National Science Foundation and the National Institutes of Health require researchers to supply plans for managing research data, called Data Management Plans (DMP), for all new grant proposals. This workshop will provide an overview of the questions to consider when creating a data management plan, with a focus on the DMPTool and tools for sharing your data (e.g. subject-specific repositories). Topics include metadata and annotation, file formats and organization, storage, backups and security, and data sharing. The workshop is geared toward graduate students, faculty, and researchers.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scheduled Dates:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
</tr>
<tr>
<td>11/6/2012 11:00 AM</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bibliographic Software</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Workshop</strong></td>
</tr>
<tr>
<td>NIH Public Access Policy</td>
</tr>
<tr>
<td>NIH Public Access Policy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Database Searching</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Workshop</strong></td>
</tr>
<tr>
<td>Web of Science (WOS)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Medical Terminology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Workshop</strong></td>
</tr>
<tr>
<td>Medical Terminology I</td>
</tr>
<tr>
<td>Medical Terminology II</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other Information Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Workshop</strong></td>
</tr>
<tr>
<td>Sharing Scholarship and Reaching Research: Open Access 101</td>
</tr>
<tr>
<td>Fair Use for Faculty and Other Instructors: Copyright issues in both the physical and online classroom</td>
</tr>
<tr>
<td>Best Practices in Data Management</td>
</tr>
<tr>
<td>Introduction to Systematic Reviews</td>
</tr>
</tbody>
</table>
A Great Number of Poster Presentations Are Never Published

<table>
<thead>
<tr>
<th>Publication rate of poster presentations</th>
<th>Discipline</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>47%</td>
<td>Orthopaedic</td>
<td><em>Donegan et al. (2010)</em> Publication Rates of Presentations at an Annual Meeting of the American Academy of Orthopaedic Surgeons</td>
</tr>
<tr>
<td>56%</td>
<td>Ophthalmology</td>
<td><em>Juzych et al. (1992)</em> Whatever Happened to Abstracts From Different Sections of the Association for Research in Vision and Ophthalmology?</td>
</tr>
<tr>
<td>42%</td>
<td>Nephrology</td>
<td><em>Harel, et al. (2011)</em> Frequency and factors influencing publication of abstracts presented at three major nephrology meetings</td>
</tr>
<tr>
<td>42%</td>
<td>Reproduction Embryology</td>
<td><em>Evers, 2000</em> Publication bias in reproductive research</td>
</tr>
<tr>
<td>19%</td>
<td>Dental Education</td>
<td><em>Galang et al. (2011)</em> Factors Influencing Publication Rates of Abstracts Presented at the ADEA Annual Session &amp; Exhibition</td>
</tr>
</tbody>
</table>
Research papers refer to poster presentation results.
The UF Institutional Repository

Provides permanent links for all or your submitted items (ideal for inclusion in your CV, tenure packet)

Share those items openly and ensure their permanent preservation.

Receive monthly reports on the usage of your items.

Indexed by online search engines (e.g. Google)
Acknowledgement

This project has been funded in part with federal funds from the National Library of Medicine, National Institutes of Health, under Contract # HHS-N-276-2011-00004-C.

This presentation is available for re-use under a creative commons attribution license.
References

• Panton Principles, Principles for open data in science. Murray-Rust, Peter; Neylon, Cameron; Pollock, Rufus; Wilbanks, John; (19 Feb 2010). Retrieved 10/18/2012 from http://pantonprinciples.org/


• Data Documentation Initiative (DDI) version 3.0 Combined Life Cycle Model: http://www.ddialliance.org/what