Data Management Planning Consultation Workshop

Lisa Zilinski, Purdue University
LISA ZILINSKI
Data Services Specialist

Special Thanks to:
The Data Education Working Group

May 3, 2013
1. Introduction to Data Management
2. Types of Data
3. Data Formats and Metadata Standards
4. Issues in Data Access and Sharing
5. Data Re-Use and Distribution
6. Preservation and Archiving
7. Additional Resources
8. DMP Exercise
9. Wrap-Up and Questions
Data Sharing and Management Snafu in 3 Short Acts
By Karen Hanson, Alisa Surkis & Karen Yacobucci
NYU Health Sciences Libraries
August 3, 2012 (Last Update: December 12, 2012)
http://www.youtube.com/watch?v=N2zK3sAtr-4
DATA MANAGEMENT

STAKEHOLDERS

- External sources
  - Funding agency mandates
  - Governmental policy
  - Open access agenda
  - Publishers

- Internal Sources
  - Faculty
  - Institutional policy

- Library
  - Data collections
  - Open access
• Libraries Strategic Plan: Liaisons will assist researchers with data management planning by:
  1. Liaisons will be able to engage with faculty and graduate students in data management planning.
  2. Liaisons will be able to relate data lifecycle to restricted or sensitive data.
  3. Liaisons will be able to discuss issues related to data “ownership”.

• Relationship Building

• Potential Benefits
  – Discovery
  – Context
  – Security
  – Preservation
  – Impact through citation
  – Documentation
  – Funding mandates

National Science Foundation (NSF) Data Management Plan (DMP) Requirements

• Plans for data management and sharing of the products of research. Proposals must include a supplementary document of no more than two pages labeled “Data Management Plan”. This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (see AAG Chapter VI.D.4), and may include:

1. the **types of data**, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
2. the **standards** to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
3. policies for **access and sharing** including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
4. policies and provisions for **re-use**, re-distribution, and the production of derivatives; and
5. plans for **archiving** data, samples, and other research products, and for preservation of access to them.

• Data management requirements and plans specific to the Directorate, Office, Division, Program, or other NSF unit, relevant to a proposal are available at: [http://www.nsf.gov/bfa/dias/policy/dmp.jsp](http://www.nsf.gov/bfa/dias/policy/dmp.jsp). If guidance specific to the program is not available, then the requirements established in this section apply.

It is 2:45 pm on Tuesday when this email hits your inbox:

Dear Super Librarian,

I attended the data management workshop last week and would like to work with you on developing my data management plan and learning about how to incorporate PURR into my proposal. I am preparing a proposal for the NSF Early Career Award for Physics. I need to have the first draft ready early May for an internal review. I would certainly like to get some help, especially for the DMP section. The full proposal deadline is July 24, 2013. Thank you for your help.

Jonathan Osterman, PhD
Assistant Professor, Physics
drmanhattan@university.edu

What do you do?
SCENARIO

A. Immediately conduct an in-depth literature review on intrinsic field subtraction in physics, unbury yourself from books and articles three weeks later, finally composing an email to Dr. Osterman, emphasizing your sincere desire to help by applying your new found love of particle physics.

B. Take a moment to review the NSF Early Career Award requirements on the NSF website. Reply to the email and set up a time to meet in person.

http://4-plates.com/
SCENARIO

A. Immediately conduct an in-depth literature review on intrinsic field subtraction in physics, unbury yourself from books and articles three weeks later, finally composing an email to Dr. Osterman, emphasizing your sincere desire to help by applying your new found love of particle physics.

Dr. Osterman was distracted while working on his proposal. He stepped into the intrinsic field subtractor, got locked in while the machine was warming up, and was disintegrated.

B. Take a moment to review the NSF Early Career Award requirements on the NSF website. Reply to the email and set up a time to meet in person.

Dr. Osterman was impressed by your invitation to meet and invited you out for coffee, his treat.
You meet Dr. Osterman at Starbucks. Dr. Osterman has brought a rough draft of his proposal, including a data management plan. What do you do?

A. You tell Dr. Osterman to scrap everything he has. You have worked on this for the past three weeks and you have everything ready to go!

B. Have a conversation with Dr. Osterman. Ask questions about his research and the data to be produced.
A. You tell Dr. Osterman to scrap everything he has. You have worked on this for the past three weeks and you have everything ready to go!

In your excitement to share all of your knowledge, you spill your coffee all over Dr. Osterman and his proposal.

Luc Reid Photography, http://www.lucreid.com/?tag=real-life-examples
B. Have a conversation with Dr. Osterman. Ask questions about his research and the data to be produced.

You work through Dr. Osterman’s DMP, asking questions and providing recommendations along the way.

TYPES OF DATA

NSF DMP SECTION 1
Example questions to ask:

- What are the products of your research?
- How is the data created, acquired and recorded?
- What instruments are used?
- What hardware or software is required (if any) to access or analyze the data?
- What format(s) will your data be in?
- What size will the data or files be?
- What other types of files will be created (project files, documentation, etc)?
Examples of data types:
- Samples
- Specimen
- Numeric data
- Physical collections
- Software
- Text
- Curricula
- Images
- Video
- Sound recordings

PURR Example
- https://purr-dev.lib.purdue.edu/publications/203
- https://purr-dev.lib.purdue.edu/publications/193
Exercise (10 minutes)

Examine the candy and wrappers on the tables and list the following:

– Describe the data.
– Describe any standards or formats of the data.
– List any metadata you find.
– Ingest data.

August Storck KG http://en.wikipedia.org/wiki/File:Merci_Chocolates_KF.JPG
DATA FORMATS AND METADATA STANDARDS

NSF DMP SECTION 2
Example questions to ask:

- Are there standard formats for data in your field or for your data type (word processing, plain text, tabular data)?
  - See the File Format Recommendations handout for specifics.
- If not, how will you format your data so others can use it?
- What metadata or standards will be applied to your data?
- How will you communicate procedures that explain the creation or management of your data files?
- Who will be responsible for ensuring data is properly formatted and that standards are applied?
Glossary of Metadata Standards
Content: Jenn Riley, Design: Devin Becker
Indiana University

SGML
Standard Generalized Markup Language
http://www.nec/is/iso/catalogue_detail.htm?cnum=16357

SGML is the precursor and current parent meta-language to XML. It is less strict in its structure than XML, including the ability to not require closing tags. Several metadata standards of interest to the cultural heritage community began as SGML languages and later migrated to XML, including EAD and TEI. HTML versions through HTML 4 are SGML languages, whereas XHTML is an XML language. Currently, XML is favored over SGML for the development of new markup languages, largely due to XML’s stricter structure.

SKOS
Simple Knowledge Organization System
http://www.w3.org/2004/02/skos/

SKOS is a Semantic Web-driven method of encoding structured vocabularies in RDF. The RDF SKOS vocabulary focuses on describing concepts, which are represented by terms, and documenting relationships between concepts. SKOS-encoded data is a key building block in the Semantic Web’s Linked Data movement. While SKOS can be used for encoding thesauri like those commonly used in the cultural heritage community, it fits less well for other types of controlled vocabularies common in this community such as name authorities. A high-profile use of SKOS in the cultural heritage community is the http://id.loc.gov service.

RSS
Really Simple Syndication
http://cyber.law.harvard.edu/rss/rss.html

RSS is a syndication format for Web content, allowing frequently updated information such as news feeds to be pushed to subscribed users. The most frequent use of RSS is to embed an RSS-encoded news feed into an otherwise human-readable web page such as a news service or a blog. An RSS feed is divided into “channels” for individual items, each of which have some required data such as title and description and some optional data such as publication date and category. RSS 2.0 allows enclosures, which support embedding of content and allow applications such as podcasting.

The main alternative to RSS for syndicated content is Atom. The RSS 2.0 specification calls for representation in XML, whereas the 1.0 specification represents information in RDF. RSS has also been known to stand for Rich Site Summary.

SCORM
Sharable Content Object Reference Model
http://www.adlnet.gov/Technologies/scorm/default.aspx

SCORM was created as an effort of the Advanced Distributed Learning initiative of the US Department of...
Disciplinary Metadata Directory

- Information about disciplinary metadata standards, including profiles, tools to implement the standards, and use cases of data repositories currently implementing them.
  - [http://www.dcc.ac.uk/resources/metadata-standards](http://www.dcc.ac.uk/resources/metadata-standards)
METADATA TOOLS AND RESOURCES

DCMI SCIENCE & RESEARCH DATA ALLIANCE

- Dublin Core Metadata Initiative—Science and Metadata Community
  - Focuses on metadata challenges specific to scientific data curation, and solutions that will benefit from the architecture and global reach of the Dublin Core Metadata Initiative.

- Research Data Alliance
  - Aims to accelerate and facilitate research data sharing and exchange.

- Prototype directory of metadata schemes for managing scientific data
Example questions to ask:

- Are there privacy or confidentiality issues on the data you are using or generating?
- Who will have access to the data during the project and after the project?
- What data will be shared?
- At what point will data be shared (raw, derived, published)?
- How will others gain access to your data?
DATA ACCESS AND SHARING
PRIVACY AND SECURITY

http://guides.lib.purdue.edu/content.php?pid=219836
1. Sharing & Access

2. Reuse & Distribution

3. How does PURR do it?
DATA ACCESS AND SHARING

4 OBJECTIVES

To collect, **publish** and preserve the digital data sets and associated documentation generated by researchers affiliated with Purdue or associated with Purdue's research projects.

To enable researchers at Purdue to satisfy the requirements of funding agencies in managing, **sharing** and preserving research data.

To provide the means for researchers, policy makers, and others to **discover** and **access** data sets generated through research done at or in conjunction with Purdue for the long term.

To provide a sustainable preservation environment where deposited research data are available to support the historical record of research, and **accessible** for use for contemporary scholarship.

PURR’s Digital Preservation Policy: [https://purr.purdue.edu/legal/digitalpreservation](https://purr.purdue.edu/legal/digitalpreservation)
Example questions to ask:

- Who owns the data?
- Under what conditions will data be shared (CC licenses, embargos)?
- Under what conditions will you allow others to create and publish derivatives from the data?
- How will others cite your data sets?
CC0 enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

Creative Commons 0 license: http://creativecommons.org/about/cc0
Terms of Deposit: https://purr.purdue.edu/legal/termsofdeposit
PURR provides a **free** online research data collaboration platform and service solution for Purdue faculty, graduates students, and staff.

**Research data** - spreadsheets, images, output from sensors and instruments, transcripts, surveys, software source code and tools, video, and observation logs.

**PURR provides:**

- Data management plan (DMP) [resources and consultation](http://datacite.org/whatisdoi)
- Collaborative research [data project space](http://datacite.org/whatisdoi)
- Dataset publication with Digital Object Identifier (DOI) *
- Long-term preservation in a Trusted Digital Repository

* PURR uses [Datacite](http://datacite.org/whatisdoi) DOIs. To learn more go to [http://datacite.org/whatisdoi](http://datacite.org/whatisdoi)
Published Datasets

Open Access Scholarship

Publisher’s Version of Scholarship
Example questions to ask:

- What backup methods and procedures are in place for your data?
- How long will you need to store your data?
- Which of your data sets will have long-term value to others and why?
- What supplemental information (reports, software, etc.) will need to preserved along with the data?
- Will funding or other institutional commitments be necessary for archiving and preserving?
- What repository or archive will you deposit your data?
- What requirements and restrictions does the repository or archive have?
PRESERVING AND ARCHIVING

LONG-TERM STORAGE AND MANAGEMENT OF DATA

PURRR Digital Preservation Policy
https://purr.purdue.edu/legal/digitalpreservation

MetaArchive
http://www.metaarchive.org/
DATA MANAGEMENT

ADDITIONAL RESOURCES
DATA MANAGEMENT RESOURCES

DMPTOOL AND DATABIB

DMPTool
https://dmp.cdlib.org/

DataBib
http://databib.org/
Supporting Information for Data Services LibGuide
http://guides.lib.purdue.edu/content.php?pid=214801&sid=1787106

DMP Self-Assessment Questionnaire
https://purr.purdue.edu/resources/7

OVPR Sponsored Programs Awards Listing
http://www.purdue.edu/research/vpr/publications/archivedawards.php
DATA MANAGEMENT RESOURCES

PEOPLE

Lisa Zilinski  
Data Services Specialist  
lzilins@purdue.edu

Jake Carlson  
Data Services Specialist  
jakecarlson@purdue.edu

Michael Witt  
Interdisciplinary Research Librarian  
mwitt@purdue.edu

Amy Hatfield  
Metadata Specialist  
hatfiea@purdue.edu

Amy Van Epps  
Engineering Librarian  
vanepa@purdue.edu

Courtney Matthews  
Digital Data Repository Specialist  
matthew6@purdue.edu

Dave Scherer  
Scholarly Repository Specialist  
dscherer@purdue.edu

Carly Dearborn  
Digital Preservation & Electronic Records Archivist  
cdearbor@purdue.edu

Sammie Morris  
University Archivist  
morris18@purdue.edu
Exercise (15 minutes)

Review the sample DMP at your table:

- Provide feedback for the sample “before” DMP.
- Compare your feedback to the “after” DMP.
- Present your recommendations.
You have completed your DMP consultation with Dr. Osterman. He submitted his proposal and won funding! Dr. Osterman sent you an email thanking you for your time and assistance in creating a winning DMP! What do you do?

A. Email Dr. Osterman back and tell him how you will be monitoring his progress to ensure his strict compliance to the DMP that you developed. You also want a citation in his paper.

B. Contact Dr. Osterman and congratulate him on his award. Remind him that the library liaisons are available for consultation throughout the data lifecycle if he needs assistance and you would love to be updated on any of his research progress.
A. Email Dr. Osterman back and tell him how you will be monitoring his progress to ensure his strict compliance to the DMP that you developed. You also want a citation in his paper.

Dr. Osterman is shocked by your response. He rushes into his research, gets locked in the intrinsic field subtractor, is disintegrated, and reforms himself, atom by atom. This unfortunate event has changed his molecular structure and he can now experience time in a non-linear, quantum fashion. Feeling the new gap between him and all of humanity, Dr. Osterman goes into self-induced exile on Mars.
B. Contact Dr. Osterman and congratulate him on his award. Remind him that the library liaisons are available for consultation throughout the data lifecycle if he needs assistance and you would love to be updated on any of his research progress.

Dr. Osterman is very happy with the result. He invites you to the next department meeting to meet the rest of his colleagues. He sends a “kudos” email to you and Dean Mullins. You win a Collaboration Award, both in the Libraries and in the Department. You go on to be a Super Librarian and protector of data management!
Thank you.