MEETING THE NSF DATA MANAGEMENT PLAN REQUIREMENT

Co-sponsored by IU Libraries, OVPR, and ORA
Stacy Konkiel, E-Science Librarian
skonkiel@indiana.edu
Learning Objectives

• Reasons behind the mandate
• Understand the basic requirements of the mandate
• Determine the many ways in which you can meet the NSF requirements
• Learn about IUB’s cyberinfrastructure and on-campus assistance in creating your data management plan
Overview

• Definitions
• NSF DMP Mandate: The Background
• Getting Prepared
• Five Requirements for your DMP & Exercises
• Examples
• How IU can help you meet the mandate
• Q&A/Feedback
Definitions

- **Cyberinfrastructure**: computing resources and networks, services, and people
- **Data management**: the technical processing and preparation of data for analysis
- **Data curation**: managing and promoting the use of data from its creation, to ensure it is fit for discovery and re-use
- **Data Sharing**: must take into account legal and ethical issues; a spectrum with many options
- **Data Citation**: mechanisms to enable easy reuse and verification, track the impact of data, and recognize and reward researchers
- **DMP** = Data Management Plan
  
  (Coates, 2012)
NSF DATA MANAGEMENT PLAN MANDATE: THE BACKGROUND
Historical Context

- NIH Public Access Policy (‘08)
- Office of Management & Budget (‘99)
- National Research Council (‘85)
- NSF DMP Requirement (‘11)

Historical Context

Similar policies at Wellcome Trust, Howard Hughes Medical Institute, NOAA, NEH:

http://www.cdlib.org/services/uc3/datamanagement/funding.html
Why do we have a data sharing mandate?
Why we have a data sharing mandate

“Such dissemination of data is necessary for the community to stimulate new advances as quickly as possible and to allow prompt evaluation of the results by the scientific community.” – NSF

- Accelerate scientific discovery
- Reproducible results
- ROI
Why we have a data sharing mandate

- **Organization = Easier Work**
- **Replicated Data = Safe(r) Data**
  - Digital data is more fragile than analog data
- **Open Data = More Citations** (Piwowar et al, 2010)

(Houston, 2011)
Why we have a data sharing mandate

• “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.”
Why we have a data sharing mandate

• “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.”
What is “a reasonable amount of time”? 

- Engineering Section: “no later than the acceptance for publication of the main findings of the final data”
- Earth Sciences: “No later than two (2) years after the data were collected.”
- Social and Economic Sciences: “within one year after the expiration of an award”
HOW TO PREPARE
DMP Basics

• No more than two pages
• Supplementary document: does not count towards page limit
• Even if no data produced, must submit a DMP
How to Prepare

- Data Inventory
- Audiences
- Obligations
  - Open Data? Intellectual Property? Confidentiality?
- Enduring value?
FIVE REQUIREMENTS FOR YOUR DMP

+ Exercises
Five Requirements for your DMP

- Types of data & Data Formats
- Metadata
- Access and Sharing
- Reuse and Distribution Policies
- Preservation
Requirements: Types of Data

“The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced”

• List any and all
  • Observational
  • Experimental
  • Simulation
  • Derived or compiled

• Be specific
Requirements: Data Formats

“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...)”

- Describe how your data will be recorded and stored
- Common formats above all else
- The more open/interoperable, the better
Exercise: Types of Data

• What type of Data or file formats do you produce in your work/research?

• Introduce yourself to your neighbor and describe the types of data you produce.

• Use the first section of the "Questions to consider" handout to give you ideas, like rate of growth to determine storage.

(Johnston et al, 2012)
Requirements: Metadata

“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…)”

• “Data about data”
• Metadata: basic information about data set(s)
• Preservation metadata: assure quality and provenance of data set(s)
• Guiding questions
## Requirements: Metadata

### Metadata

![Image of Christina’s World](image-url)

<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Christina’s World</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Creator</strong></td>
<td>Wyeth, Andrew</td>
</tr>
<tr>
<td><strong>Date</strong></td>
<td>1948</td>
</tr>
<tr>
<td><strong>Subject</strong></td>
<td>Painting, American Artists</td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td>Painting, Tempera on panel</td>
</tr>
<tr>
<td></td>
<td>32 1/4 x 47 3/4“</td>
</tr>
<tr>
<td><strong>Provenance</strong></td>
<td>&quot;Stolen in 1999; recovered by the Museum in 2003.&quot;</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th><strong>Ex. Dublin Core</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Creator</strong></td>
</tr>
<tr>
<td><strong>Date</strong></td>
</tr>
<tr>
<td><strong>Subject</strong></td>
</tr>
<tr>
<td><strong>Format</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Provenance</strong></td>
</tr>
</tbody>
</table>
Requirements: Metadata

Metadata

Ex. Darwin Core

ScientificNameID: Bolborhynchus aymara
DecimalLatitude: -23.8169444
DecimalLongitude: -65.4847222
Year: 2005
IdentificationID: 52356
Preparations: tissue, round skin, other

Image via http://aymary.wz.cz/members.htm
Requirements: Metadata

- Use existing standards and controlled vocabularies
- Where standards don’t exist, make note!
- Make metadata central to your study design
- Supply minimum information relevant to help others understand and access your data
- Consider supplying preservation metadata
  - Technical specifications
  - MD5 checksums
Exercise: Metadata

• What standards will you use for data documentation and metadata format?

• Imagine that I am about to enter your lab or department as a new grad student. What type of documentation about your files and data collection would I need to get up to speed? Are there metadata standards? If not, how would you make this process more standard so anyone in your field could understand? Discuss this possibility with your partner.

(Johnston et al, 2012)
Requirements: Access and Sharing

“policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements”
- With whom/how will you share?
- Will you “open it up” after time? When?
- Encrypt and store your ePHI and other data subject to IRB, HIPAA, FRPAA, etc regulations
Requirements: Reuse & Distribution Policies

Privacy & Confidentiality
• Interrelated with issue of access
• Subject to IRB regulations?

Reuse & Distribution Policies
• Also subject to IRB regulations
• How do you want others to…
  Use your data? (Non-commercial only?)
  Credit your work?
  Share your work with others?
Requirements: Reuse & Distribution Policies

“policies and provisions for re-use, re-distribution, and the production of derivatives”

• IU Legal Counsel is final word

• Recommended:
  • Open Data Commons Attribution License
  • Creative Commons Zero License

• Resources
  • Digital Curation Centre's "How to License Research Data"
  • Open Definition's list of recommended data licenses
Exercise: Access

• How will you release your data for access?

• Break into three groups.
• In each group, come up with 1-2 examples of steps needed to address the following data-related concerns:
  1. Privacy/Confidential Data
  2. Data Security (physical or digital)
  3. Intellectual Property rights

(Johnston et al, 2012)
Exercise: Sharing

• How will you share your data?

• In same groups, come up with one pro and one con for the following data sharing options:
  1. Posting data online (lab website or OA repository)
  2. Publish in a journal as supplemental material
  3. Make available on request

(Johnston et al, 2012)
Requirements: Preservation

“Plans for archiving data, samples, and other research products, and for preservation of access to them”

- Standard at IUB: “At least three years beyond the end of the project”
- Physical samples & Digital data
- Who assumes responsibility?
Exercise: Preservation

• How will you archive the data for preservation and long-term access?

• Thinking back to your (under)graduate degree, would you be able to access the data from your dissertation today? Why or why not? How might this be different for students going forward?

(Johnston et al, 2012)
Requirements: Other

Data Storage

• 3 copies in separate locations

• Yes
  Stable, short and long term storage for life of project+
  Attention to sensitive data issues
  Departmental/University tech support

• No
  Local storage (on lab computers, personal laptops, thumb drives)
FAQs

• If my data is freely available, how will I ensure that I am credited for my work?
• What if my research doesn’t produce data?
• What if it uses existing data?
• Do I have to make my data publicly available?
• How long do I need to keep my data?
• If data or samples are requested before I have completed all analyses on them, must I share them?

→ [http://1.usa.gov/MWv5ff](http://1.usa.gov/MWv5ff) <--
EXAMPLES
Example: Atmospheric Sciences

Atmospheric CO$_2$ Concentrations, Mauna Loa Observatory, Hawaii, 2011-2013

Example: Social Sciences


Image via http://go.iu.edu/6ll
Example: Ecology

The influence of plant functional types on ecosystem responses to altered rainfall

Image via http://commons.wikimedia.org/wiki/File:Rainfall_in_Amravati.jpg
Example: Microbiology

Biosignature Suites: Using Connections between Microbes & Minerals to understand Biogenic Carbonates

Image via http://www.geomar.de/uploads/pics/Mikrosonde_plag_map.jpg
HOW CAN IU HELP ME MEET THE MANDATE?
How IU can help you meet the mandate

IU Empowering People Strategic Plan for IT: Action 33:

“IU should provision a data utility service for research data that affords abundant near- and long-term storage, ease of use, and preservation capabilities. This data utility will need to offer a range of services for securing data, providing authorized access within and beyond IU; ensuring metadata description, annotation, and provenance; and providing backup/recovery services.”
How IU can help you meet the mandate

Staff Expertise
Developing your proposal (ORA/PDS)
Metadata, Checking your DMP (Libraries)
**Depositing** and **Preserving** Data (Libraries & UITS)

Cyberinfrastructure
Research File System (UITS)
Scholarly Data Archive (UITS)
IUScholarWorks Repository (Libraries)
IU Resources > Staff Expertise > IUB Data Management Service

- Preparing your data
- Storage (HIPAA-compliant)
- Preservation
- Access
- Data Management Plan consultations
  - Lab-based
  - Funder required
IU Resources > Staff Expertise

- Proposal Development Help, Grant Compliance
  - Proposal Development Services
  - Office of Research Administration
- Responsible Conduct of Research (RCR) & DM
- Poynter Center for Research Ethics
- RCR classes via ORA
Data Storage ("filesystem" storage)

- **Research File System** (UITS)
  - 100 GB
  - Individual and team accounts
  - HIPAA-aligned, not encrypted
  - Regularly backed up
  - Small files that are updated frequently
Data Storage ("filesystem" storage)

- **Data Capacitor** (UITS)
  
  Big Red, Quarry, Mason
  
  10 TB
  
  Not backed up, regularly cleared
  
  Large files
IU Resources > Tech > Storage

Data Storage ("filesystem" storage)

- **XSEDE** (multi-institution)
  - Visualizations
  - Software
  - “Science Gateways”
  - Big Data
  - For life of project (some persistent storage in SGs)
IU Resources > Tech > Storage

Data Storage ("filesystem" storage)

- **IU CTSI “Service Cores”** (IU, PU, ND)
  Subject-specific (social and physical sciences)
  Storage space & capabilities vary
  Computing and project planning

- **Advanced Information Technology Core** (UITS)
  Electronic personal health information (ePHI)
  HIPAA-protected information
  Biomedical research
  Paid service
IU Resources > Tech > Storage

Data Storage ("archival" storage)

- IU Box
  - 50 GB
  - Syncs with local storage
  - Permanent

- Scholarly Data Archive (SDA)
  - Where you back up RFS and Data Capacitor files
  - 50 PB
  - Permanent tape storage
IU Resources > Tech > Access

- **IU Box**
  - Role-based access
  - Share with anyone
  - No HIPAA, FERPA, FISMA, GLBA

- **IU Alfresco Share**
  - Role-based access
  - Multi-institution
  - No ePHI data allowed
IU Resources > Tech > Access

Access

- **IUScholarWorks Repository**
  - Open Access repository
  - Indexed by Google, et al
  - No sensitive data unless de-identified, etc

- **WinSCP**
  - SFTP
  - Access via username and password
  - Desktop client
Preservation

- **IUScholarWorks Repository**
  - Storage in perpetuity
  - Format migration
  - Interface for OA data stored on SDA
  - File integrity validation, embargoes
  - Managed by IU Libraries

- **Scholarly Data Archive** (SDA)
  - Permanent, perpetual tape storage
  - Mirroring & redundancy
LEADII Vortex2 Archive Dataset

Title: LEADII Vortex2 Archive Dataset

Author: Plale, Beth

Date: 2011-02-25

Date(s) Covered: May 01, 2010 - June 15, 2010

Geographic / Spatial Information: Oklahoma

Methodology: Executing 214 workflows, using 109,568 CPU hours and generating 215 GB of data and over 9,100 2D products, LEAD II produced short-term, highly accurate weather forecasts each morning and made the results instantly available to researchers in the field using mobile phones and a field viewer.

File Information: Each 28GB zipped tarball contains at the top level a directory "out" which consists of a number of directories, each containing the results of a single forecast. The directory, out/forecast_20100501090000EDT_run001, for instance is for 9:00 May 1st, 2010 (Eastern Daylight Time). Within a forecast directory are approximately 9 data products (sometimes less than 9 data products because visualization generation for some variables failed). Each data product is in its own file, and described by a metadata file of the same name but ending in .metadata.xml.

Location: http://hdl.handle.net/2022/12987

Type: Dataset

Embargo release date: 2212-04-11

External Files

http://purl.dlib.indiana.edu/iusw/data/2022/12987/LEADII-Vortex2-dataset.tar.gzaa
http://purl.dlib.indiana.edu/iusw/data/2022/12987/LEADII-Vortex2-dataset.tar.gzab
http://purl.dlib.indiana.edu/iusw/data/2022/12987/LEADII-Vortex2-dataset.tar.gzac
http://purl.dlib.indiana.edu/iusw/data/2022/12987/LEADII-Vortex2-dataset.tar.gzad
http://purl.dlib.indiana.edu/iusw/data/2022/12987/LEADII-Vortex2-dataset.tar.gzae
http://purl.dlib.indiana.edu/iusw/data/2022/12987/LEADII-Vortex2-dataset.tar.gzaf
Q&A / Feedback / References

Download this presentation at:

> http://hdl.handle.net/2022/14722 <

