MEETING THE NSF DATA MANAGEMENT PLAN REQUIREMENT

Co-sponsored by IU Libraries, ORA, UITS

Stacy Konkiel
Science Data Management Librarian
skonkiel@indiana.edu

Overview

- Definitions
- NSF DMP Mandate: The Background
- Getting Prepared
- Five Requirements for DMPs
  - 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
- **DMP** = Data Management Plan

(Coates, 2012)

NSF DATA MANAGEMENT PLAN MANDATE: THE BACKGROUND

Historical Context

- Before “data management” there was “data sharing”

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.”

With the responsibility to share data, proper data management becomes essential.

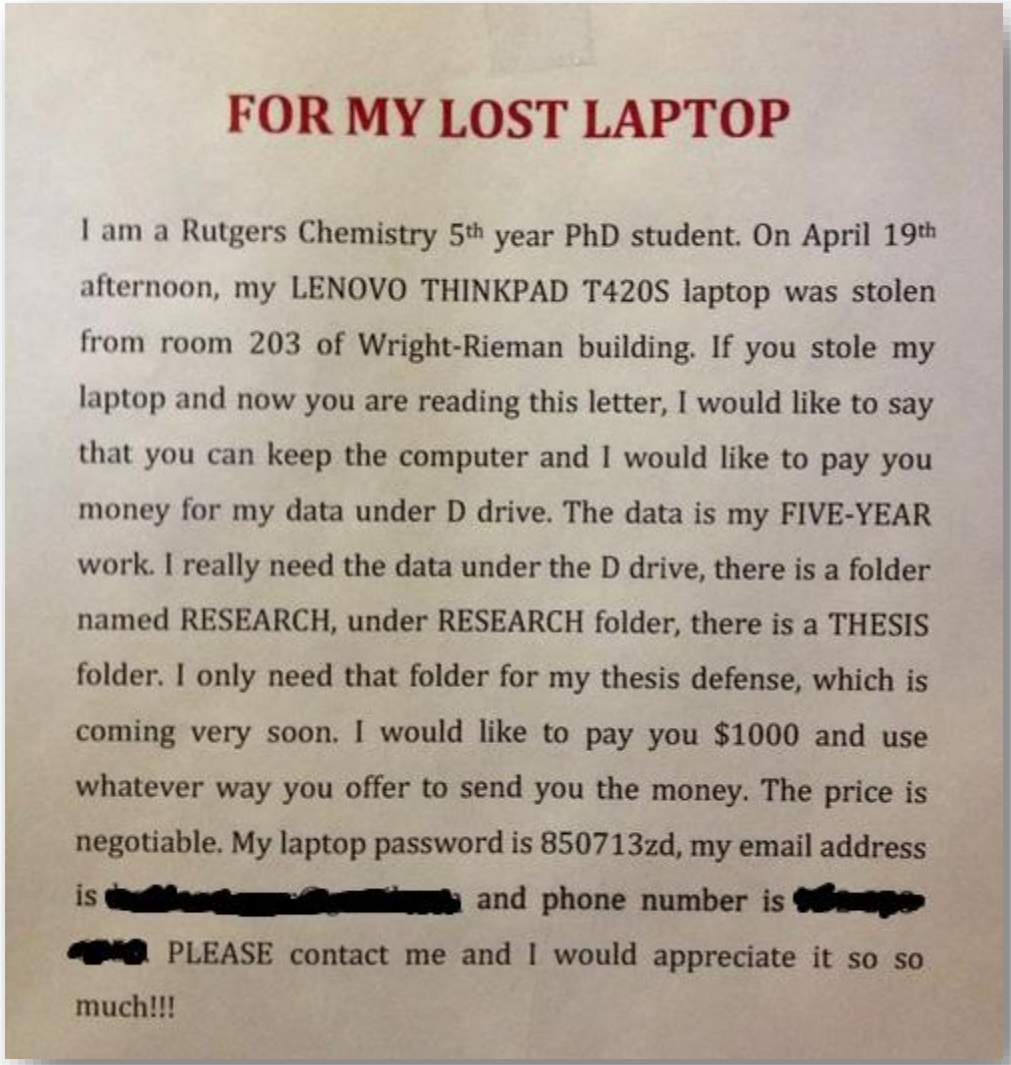
Why worry about managing data?

The consequences are **stark**:

- Loss of cultural heritage
- Inability to move cancer research insights from lab to clinic
- Retracted papers
- Faked data
- Cherry-picked results
- Inefficiencies, wasted time, wasted tax dollars

Source: Dorothea Salo's "Data Horror Stories" Pinboard

Why worry about managing data?



FOR MY LOST LAPTOP

I am a Rutgers Chemistry 5th year PhD student. On April 19th afternoon, my LENOVO THINKPAD T420S laptop was stolen from room 203 of Wright-Rieman building. If you stole my laptop and now you are reading this letter, I would like to say that you can keep the computer and I would like to pay you money for my data under D drive. The data is my FIVE-YEAR work. I really need the data under the D drive, there is a folder named RESEARCH, under RESEARCH folder, there is a THESIS folder. I only need that folder for my thesis defense, which is coming very soon. I would like to pay you $1000 and use whatever way you offer to send you the money. The price is negotiable. My laptop password is 850713zd, my email address is [redacted] and phone number is [redacted] PLEASE contact me and I would appreciate it so so much!!!

Why worry about managing data?

“Trying to understand my old spreadsheets.”

(WhatShouldWeCallGradSchool)

Why worry about managing data?

Changing research landscape and increased expectations of reusability and shareability of your data from:

Funding agencies

Others in your discipline

Tax payers

University research administration

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

Take a step back and make note of the following:

- 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


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

Ex. Dublin Core

Title: Christina's World

Creator: Wyeth, Andrew

Date: 1948

Subject: Painting,
American Artists

Format: Painting

Tempera on panel
32 1/4 x 47 3/4"

Provenance: "Stolen in 1999; recovered by the Museum in 2003."

Requirements: Metadata

Metadata

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

Ex. Darwin Core

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

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

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

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”

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?

CC libraryman

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

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?

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**

  Unencrypted local storage (on lab computers, personal laptops, thumb drives)

EXERCISE

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 ➔

EXAMPLES

Example: Atmospheric Sciences

Atmospheric CO_2 Concentrations, Mauna Loa Observatory, Hawaii, 2011-2013

Image via http://www.nytimes.com/2010/12/22/science/earth/22carbon.html

Example: Social Sciences

Social Pricing: Image Management, Social Preferences and Pay-What-You-Want

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

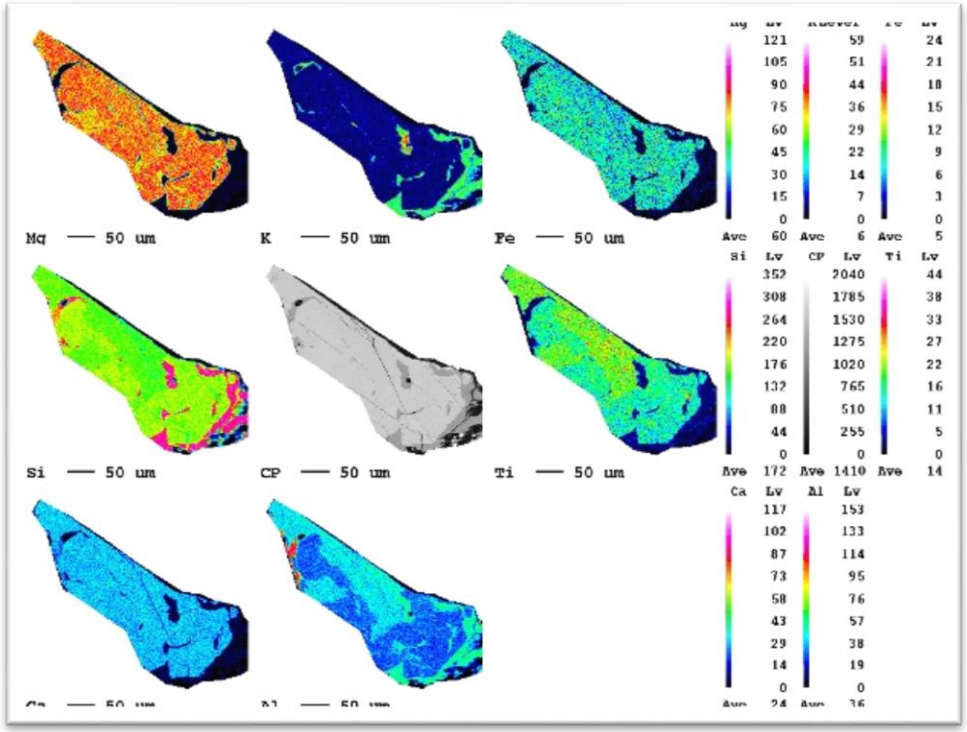


Image via http://www.geomar.de/uploads/pics/Mikrosonde_plag_map.jpg

HOW CAN IU HELP ME MEET THE MANDATE?

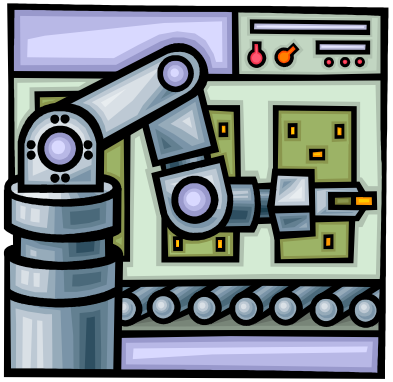
How IU can help you meet the mandate

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)

Research Technologies – Storage Options

- **Data Capacitor 2 (DC)***
- **Research File System (RFS old and new)***
- **Scholarly Data Archive (SDA)***
- **Alfresco Share**
  - for capturing, sharing, and retrieval of information across virtual teams.
- **RedCap**
  - create and design online surveys and databases (or a mix of both)
  - primarily intended for biomedical researchers
- **Research Database Complex**
  - research-related databases and data-intensive applications that require databases.
  - Oracle and MySQL databases, and provides an environment for database-driven web applications focusing on research

* Today's Focus

What is Data Capacitor 2?

- High-speed, large capacity storage system
  - Located in Bloomington Data Center
  - New replacement for original Data Capacitor (DC)
  - 3.5 Petabytes of space
  - Over 40GigaByte/second aggregate bandwidth
  - Available on all HPS systems (Big Red 2, Quarry, Mason)
- DC2 data is not backed up
  - It is intended for short-term usage
  - Use SDA to backup and archive critical data

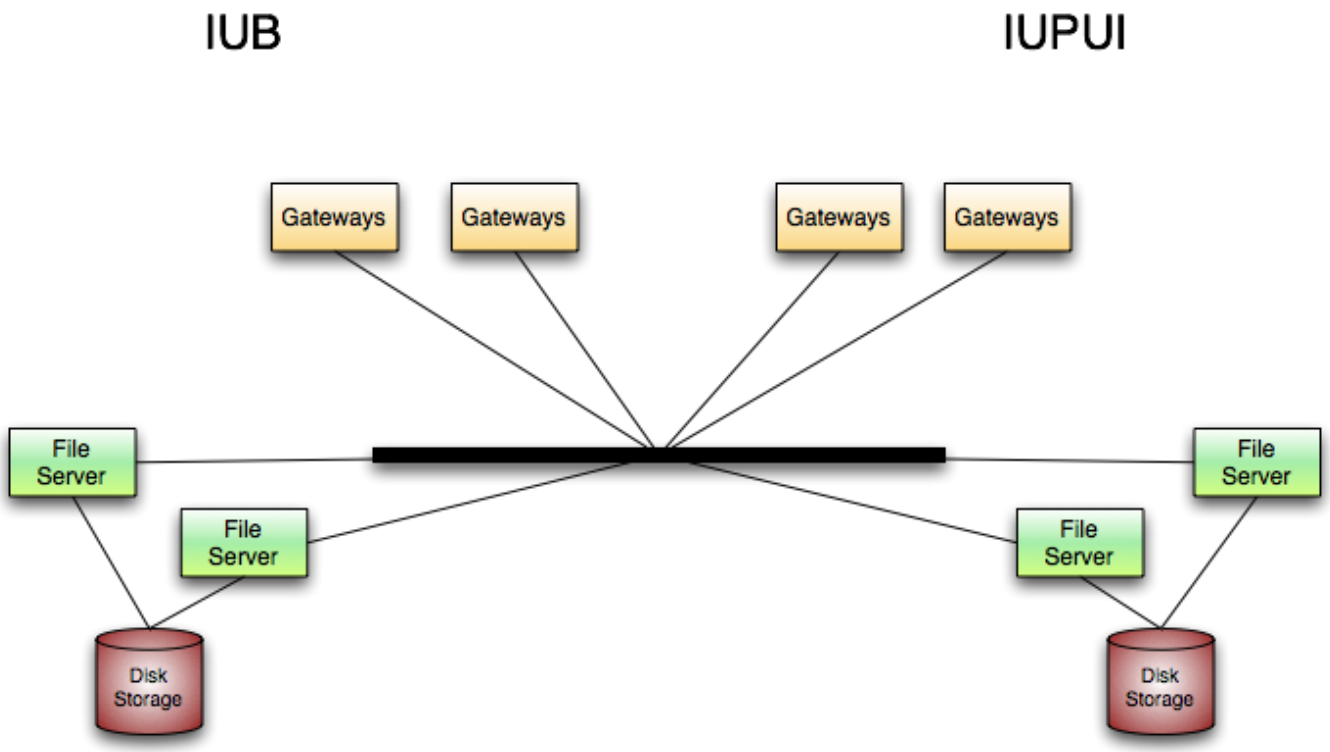
Data Capacitor Policies

- Two kinds of storage space
  - Data in ***scratch space*** is purged if it hasn't been used in **60** days
    - /N/dc2/scratch/
  - Data in ***project space*** is purged if it hasn't been used in **180** days
    - /N/dc2/projects/
- HIPAA-aligned

Data Capacitor Usage

Best Use Cases	Problematic Use Cases
Large files (up to Terabytes in size)	Lots of tiny files <1MB • Slow performance on both reads and writes
Files used frequently for computation	Long-term storage
Sharing data with collaborators • Project Space	

What is the Research File System (RFS)?



- Distributed File System intended for Research Data
- Based on OpenAFS
- 4 gateways (providing access via Samba/CIFS, sftp/scp, web)

Who can use RFS?

- Available to faculty, staff and graduate students
  - undergraduates can be sponsored by faculty or staff
- Request personal account via https://itaccounts.iu.edu/
- Individuals: 100 GB default quota, extensions generally granted
- Project spaces: 100 GB default quota, extensions generally granted
  - Can be shared with other RFS users
  - No group account required
  - Granular permission options (e.g. Grad Students read-only, Faculty read-write)
  - Request via email to store-admin@iu.edu

RFS Uses

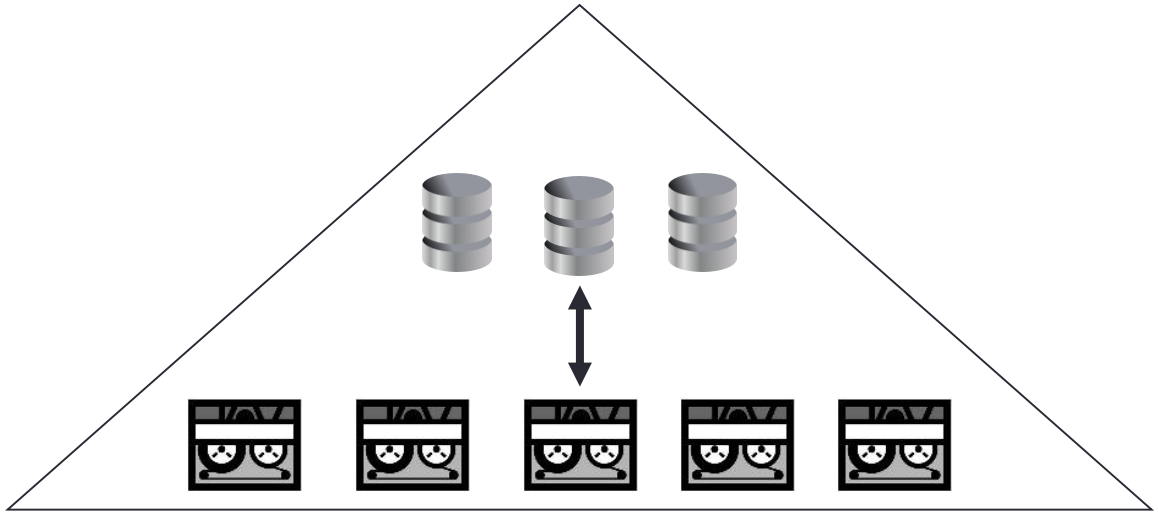
Best Use Cases	Problematic Use Cases
Relatively small files (up to GBs in size)	Not intended for backups or archiving ▪Use Scholarly Data Archive (SDA) for archiving
Files that are updated/accessed frequently	Files updated by multiple users at once ▪E.g. Access and other databases
Editing directly in RFS, using Samba or the AFS client	
Files that need to be shared, i.e. group project work	

Why should I use RFS?

- Data only stored in IU data centers
- Data is backed up nightly; restores available for up to 30 days
- You can access previous day's changes yourself in 1day-backup
- HIPAA-aligned
- Available on your desktop and also IU supercomputers (Quarry, BigRed)
- Project spaces for collaboration
- Support (electronically and we have staff at IUPUI and IUB)

What is the Scholarly Data Archive (SDA)?

- Massive data archive for Indiana University
- Default stores 2 copies of data, IUB and IUPUI
- Operating since 1999
- Primarily tape storage
- HIPAA-aligned



Who can use the SDA?

- Available to faculty, staff and graduate students
  - undergraduates can be sponsored by faculty or staff
- Request personal account via https://itaccounts.iu.edu/
- default quota 5TB
  - 2nd copy of data is not counted
  - additional storage is readily available

SDA Uses

Best Use Cases	Problematic Use Cases
Files of at least 1MB - Single file can be up to 10TB	Small files - Small files should be aggregated with a tool like WinZip or tar
Archive files - Files rarely updated - Files need to be kept long time	Files that will frequently change
Files are read often - Frequently accessed files tend to stay on disk cache	Do not edit files in place

Why should I use the SDA?

- Data only stored in IU data centers
- Data integrity
  - By default two copies of data, IUB and IUPUI each get a copy
  - checksum (data “fingerprint”) storing and validation available
- HIPAA-aligned
- Available on your desktop
- Support (electronically and we have staff at IUPUI and IUB)

Working with Research Technologies

- RT Staff at IUPUI and IUB, but support for all 8 campuses
- Opportunities to Interact in person http://pti.iu.edu/calendar
  - Research Tech Expo http://researchtech.iu.edu/
    - October 8 @ IUB
    - October 10 @ IUPUI
  - Visiting departments and labs by request
  - Workshops/Training
- Electronically
  - researchtech@iu.edu

IUB Data Management Service

- Preparing your data
- Basic Storage (HIPAA-compliant)
- Preservation
- Access
- Data Management Plan consultations
- DMPTool.org
  - General NSF & some directorate/division templates, NIH, NEH
- Libraries Data Management Guide
  http://libraries.iub.edu/data

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

Q&A / Feedback / References

Download this presentation at:

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

Coates, H. (2012). Data management plans & planning: Meeting the NSF Requirement [presentation]. Presented June 13, 2012. Retrieved from http://www.slideshare.net/goldenphizzwizards/ovcr-workshop-meeting-the-nsf-dmp-requirement-20120613-final.

Houston, B. (2011). Data Management Plans (DMP) for NSF Proposals [presentation]. Presented December 2, 2011. Retrieved from http://www.slideshare.net/herodotusjr/data-management-plans-dmp-for-nsf-10435609

Johnston, Lisa; Lafferty, Meghan; and Petsan, Beth. (2012). "Training Researchers on Data Management: A Scalable, Cross-Disciplinary Approach." *Journal of eScience Librarianship* 1(2): Article 2. http://dx.doi.org/10.7191/jeslib.2012.1012