Writing an Effective Data Management Plan

Lisa Spiro, Melissa Wentz & Erik Engquist
Rice University

March 11, 2016

Outline

1. Discuss challenges in developing data management plans (DMPs)
2. Review examples of agency guidelines
3. Highlight best practices for data management
4. Evaluate a sample plan
5. Experiment with DMP Tool
6. Explore resources for writing DMPs

1. What challenges do you face in dealing with data?

2. Examples of agency guidelines

Nearly All Federal Funding Agencies (& Some Nonprofits) Require or Will Soon Require DMPs

- NSF (specific guidelines by directorate)
- NIH
- CDC
- NEH Office of Digital Humanities
- DOE
- DOT
- FDA
- NOAA
- USAID
- USGS
- Moore Foundation
- Alfred P. Sloan Foundation
- ...

Why do funding agencies require DMPs?

- Facilitate replication of results
- Allow alternative hypotheses to be tested
- Enable comparative studies
- Promote new research
- Foster education
- Maximize investment of research money

Some Principles Underlying Data Management/ Sharing Requirement

- Data: “the recorded factual material commonly accepted in the scientific community as necessary to validate research findings”
- Values openness for fostering scientific progress & integrity.
- Respects norms of disciplinary communities.
- Recognizes constraints such as confidentiality & intellectual property.
- Promotes “timely access” while respecting rights of researchers to analyze data & publish results.

Rice University’s Research Data Management Policy

- PI is the primary steward of data & is responsible for:
    - Educating research team on “obligations regarding research data”
    - Ensuring accuracy, security & management of data
    - Complying with sponsor requirements
- Researcher has right to choose research directions, publish work & share findings.
- Rice holds legal title to data.
- Normal retention period for data = 5 years after grant expiration.

Information to Include in NSF DMPs

Guidelines vary by directorate, but generally require:

- Types of data
- Standards to be used for data & metadata
- Policies for access and sharing (including IP)
- Policies and provisions for re-use & re-distribution
- Plans for archiving data and for preserving access

Read the Guidelines.

- Pay attention to the specific requirements of your funding agency.
- Typically DMPs are 2 pages long.

DMPs and Compliance

- Proposals without DMPs will not be reviewed.
- Some agencies/directorates (e.g. NSF Bio) require reporting on DMP implementation in annual & final reports.
- Some directorates will consider DMP implementation in evaluating future proposals.
- Pay attention to policies governing how data should be handled, e.g. HIPAA.

3. Some Best Practices for Managing Research Data

1. Understand your data.

- What kind of data will you produce/ use?
  - What computing resources are needed?
  - What will be the workflow for managing data?
  - How much data will you be generating?
- What costs will be associated with managing data?
  These can often be written into grants.
- Are there restrictions on the data (e.g. HIPAA)?

2. Draw upon data management norms for your discipline.

- Ecology: British Ecological Society and ESA
- Environmental science: DataONE
- Social science: ICPSR, Dataverse & The American Economic Review: Data Availability Policy

>> Know up front what is required to share data through your discipline’s repository (e.g. ICPSR).

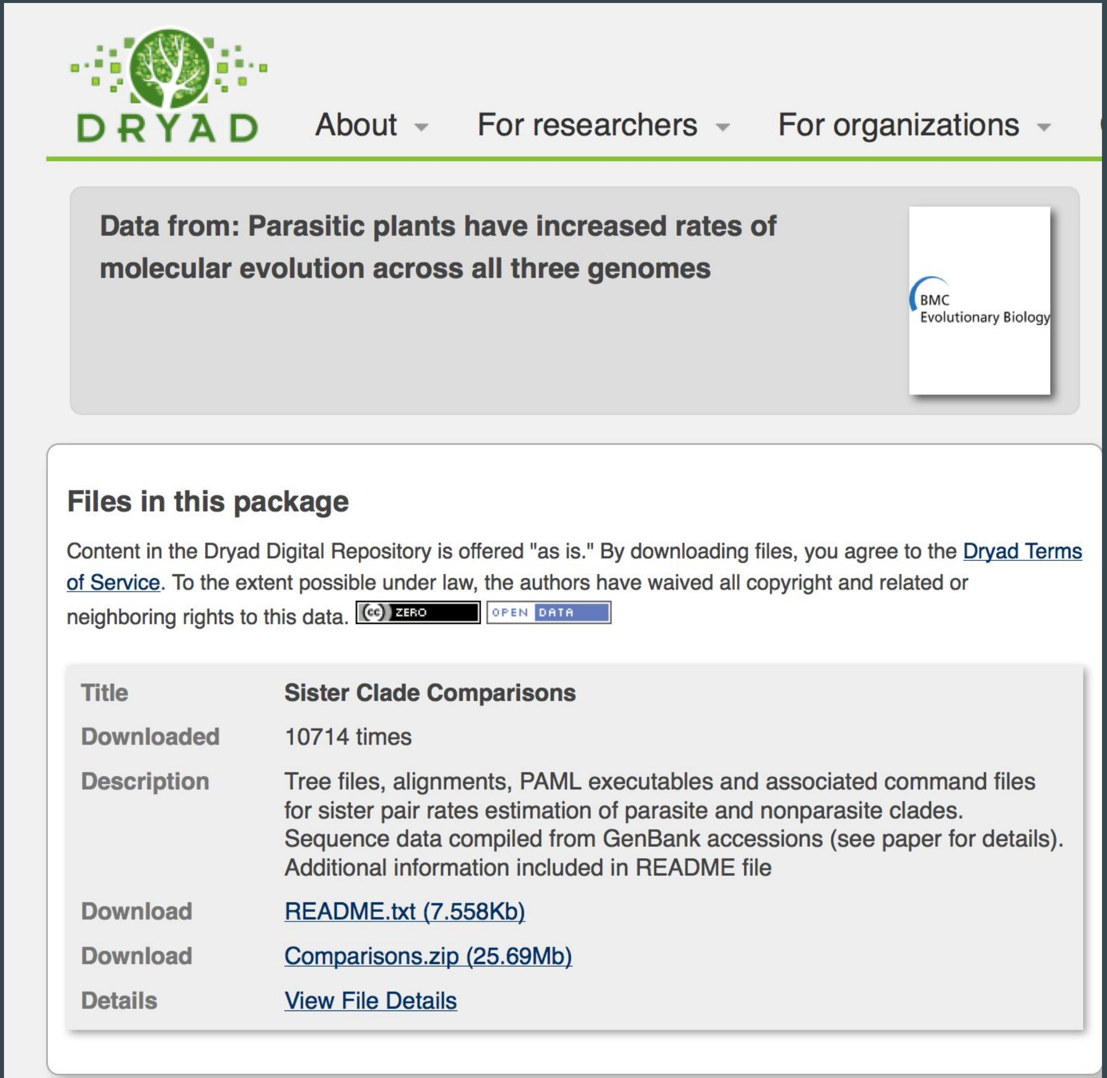
3. Describe your data.

- Document your data, recording information like title, creator, dates, subject, context & methods.
- Use established metadata standards so data are discoverable & interpretable.
    - e.g. Ecological Metadata Language or Data Documentation Initiative [DDI]

Example of Metadata for Data: Dryad

Based on Dublin Core standard

http://datadryad.org/resource/doi:10.5061/dryad.fc74k



Data from: Parasitic plants have increased rates of molecular evolution across all three genomes

Files in this package

Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the Dryad Terms of Service. To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data. (cc) ZERO OPEN DATA

Title	Sister Clade Comparisons
Downloaded	10714 times
Description	Tree files, alignments, PAML executables and associated command files for sister pair rates estimation of parasite and nonparasite clades. Sequence data compiled from GenBank accessions (see paper for details). Additional information included in README file
Download	README.txt (7.558Kb)
Download	Comparisons.zip (25.69Mb)
Details	View File Details

4. Use effective storage strategies.

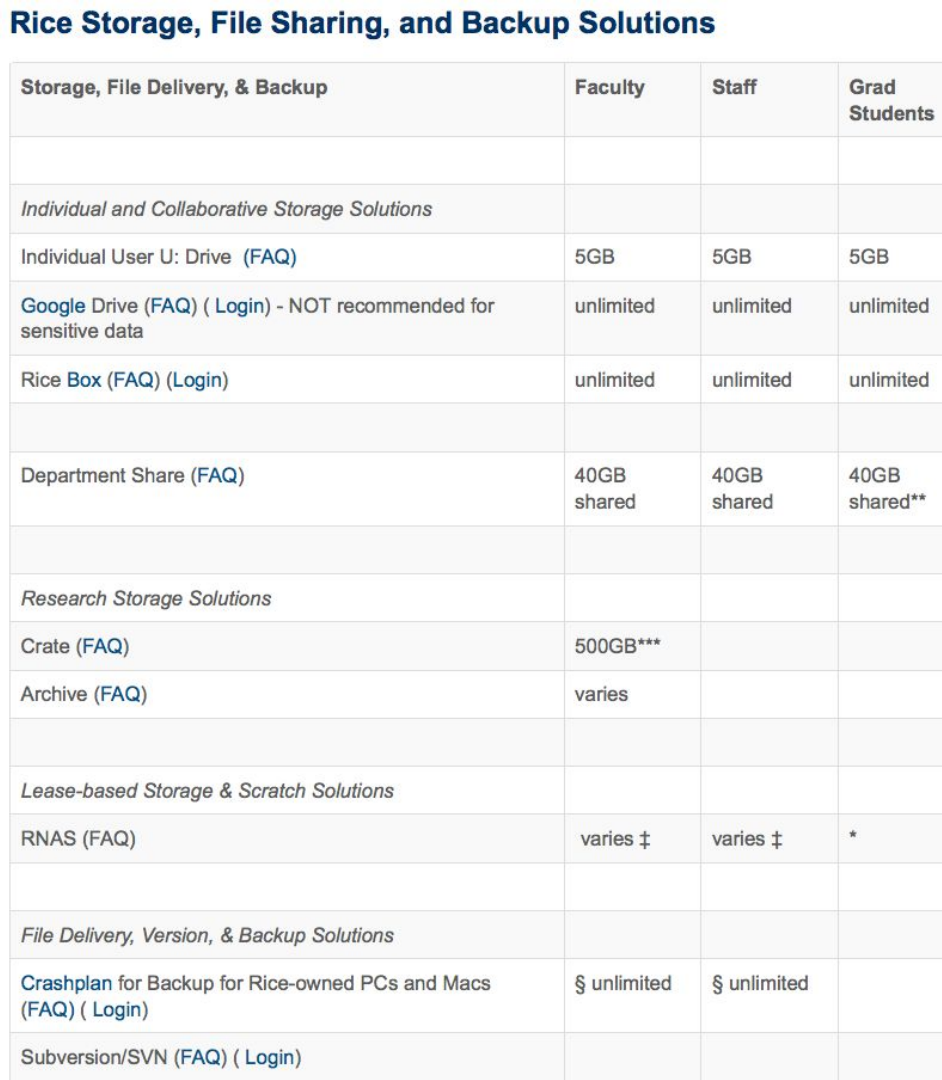
- Keep 3 copies of data in multiple locations: “original, near and far” (e.g. hard drive, external drive, server)
- Manage versions of files (e.g. using Subversion or GitHub)
- Determine who needs access to files & ensure they are trained in properly handling them.
- Provide appropriate security for data (e.g. anti-virus protection, access control, encryption, de-identification of data).
- Store data in non-proprietary formats (e.g. .txt not .doc)

Storage Options at Rice

Crate: “research storage solution for Rice researchers; 500GB per research award”

Archive: “research solution for long-term retention of completed work”

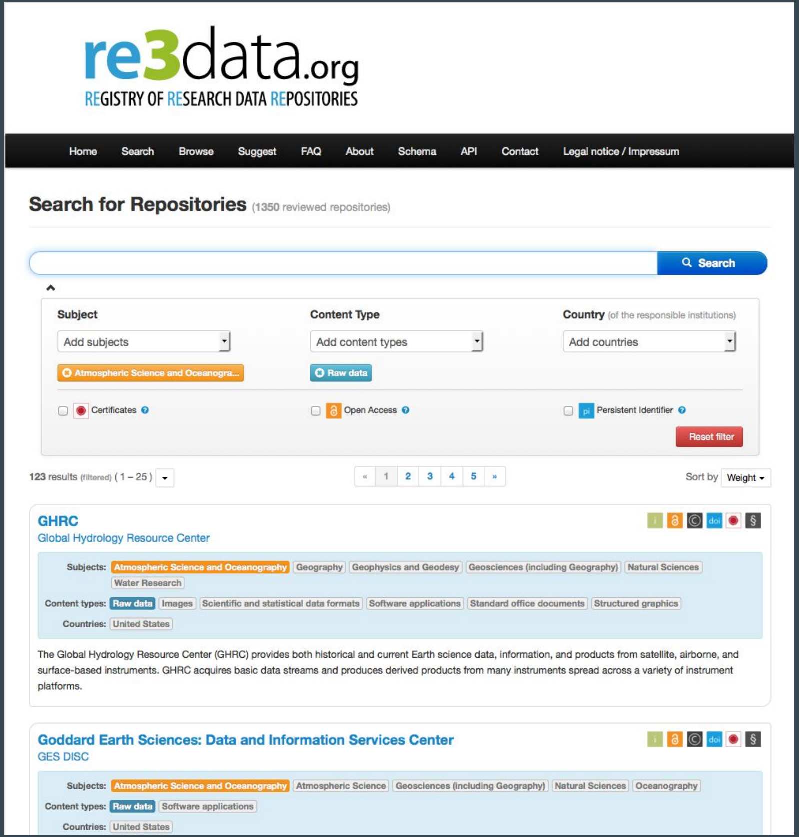
Box: “enterprise cloud-based storage & collaboration service”



Rice Storage, File Sharing, and Backup Solutions

Storage, File Delivery, & Backup	Faculty	Staff	Grad Students
Individual and Collaborative Storage Solutions			
Individual User U: Drive (FAQ)	5GB	5GB	5GB
Google Drive (FAQ) (Login) - NOT recommended for sensitive data	unlimited	unlimited	unlimited
Rice Box (FAQ) (Login)	unlimited	unlimited	unlimited
Department Share (FAQ)	40GB shared	40GB shared	40GB shared**
Research Storage Solutions			
Crate (FAQ)	500GB***		
Archive (FAQ)	varies		
Lease-based Storage & Scratch Solutions			
RNAS (FAQ)	varies ‡	varies ‡	*
File Delivery, Version, & Backup Solutions			
Crashplan for Backup for Rice-owned PCs and Macs (FAQ) (Login)	§ unlimited	§ unlimited	
Subversion/SVN (FAQ) (Login)			

5. Share data through an appropriate data archive.



Search for Repositories (1350 reviewed repositories)

Subject: Add subjects — Atmospheric Science and Oceanogra...

Content Type: Add content types — Raw data

Country (of the responsible institutions): Add countries

□ Certificates □ Open Access □ Persistent Identifier

123 results (filtered) (1 – 25) « 1 2 3 4 5 » Sort by Weight

GHRC
Global Hydrology Resource Center

Subjects: Atmospheric Science and Oceanography | Geography | Geophysics and Geodesy | Geosciences (including Geography) | Natural Sciences | Water Research

Content types: Raw data | Images | Scientific and statistical data formats | Software applications | Standard office documents | Structured graphics

Countries: United States

The Global Hydrology Resource Center (GHRC) provides both historical and current Earth science data, information, and products from satellite, airborne, and surface-based instruments. GHRC acquires basic data streams and produces derived products from many instruments spread across a variety of instrument platforms.

Goddard Earth Sciences: Data and Information Services Center
GES DISC

Subjects: Atmospheric Science and Oceanography | Atmospheric Science | Geosciences (including Geography) | Natural Sciences | Oceanography

Content types: Raw data | Software applications

Countries: United States

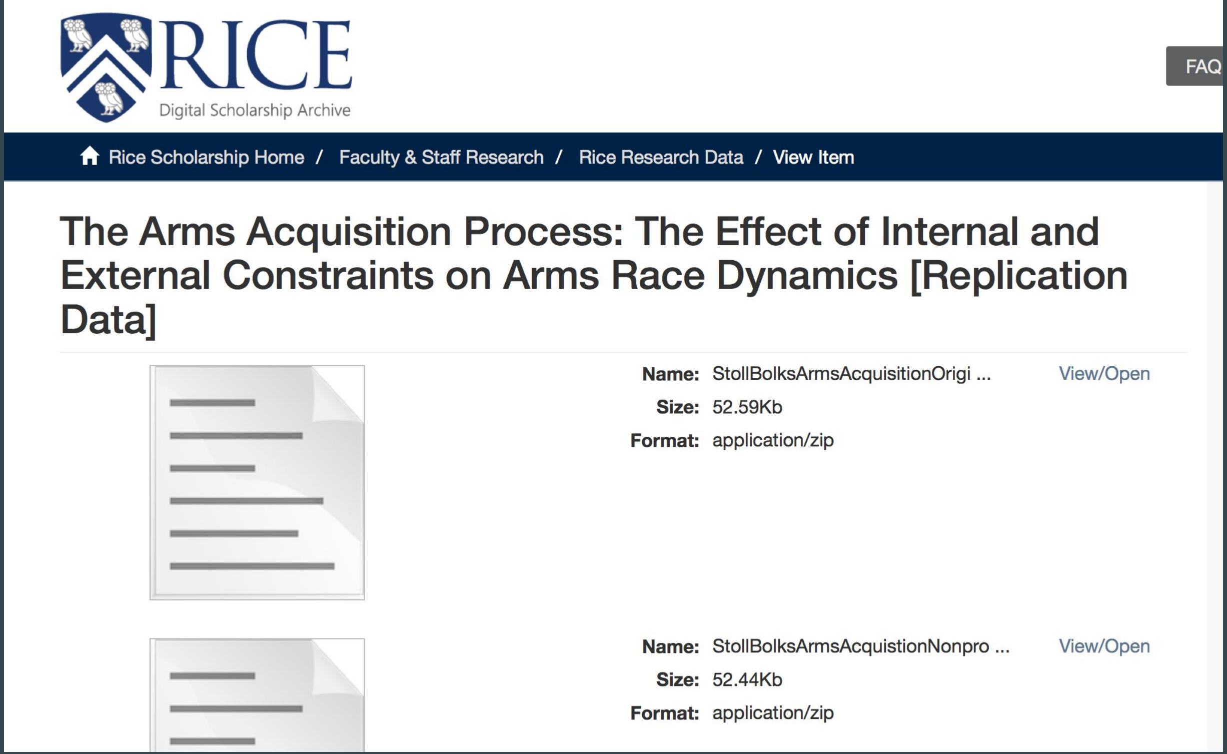
http://www.re3data.org/

Agencies permit different approaches to data sharing. Perhaps the best is to use a national data archive.

Why share?

- Increase citations
- Meet reproducibility & data sharing standards
- Facilitate future research

Share Small to Medium Datasets through the Rice Digital Scholarship Archive



RICE
Digital Scholarship Archive
FAQ
Rice Scholarship Home / Faculty & Staff Research / Rice Research Data / View Item

The Arms Acquisition Process: The Effect of Internal and External Constraints on Arms Race Dynamics [Replication Data]

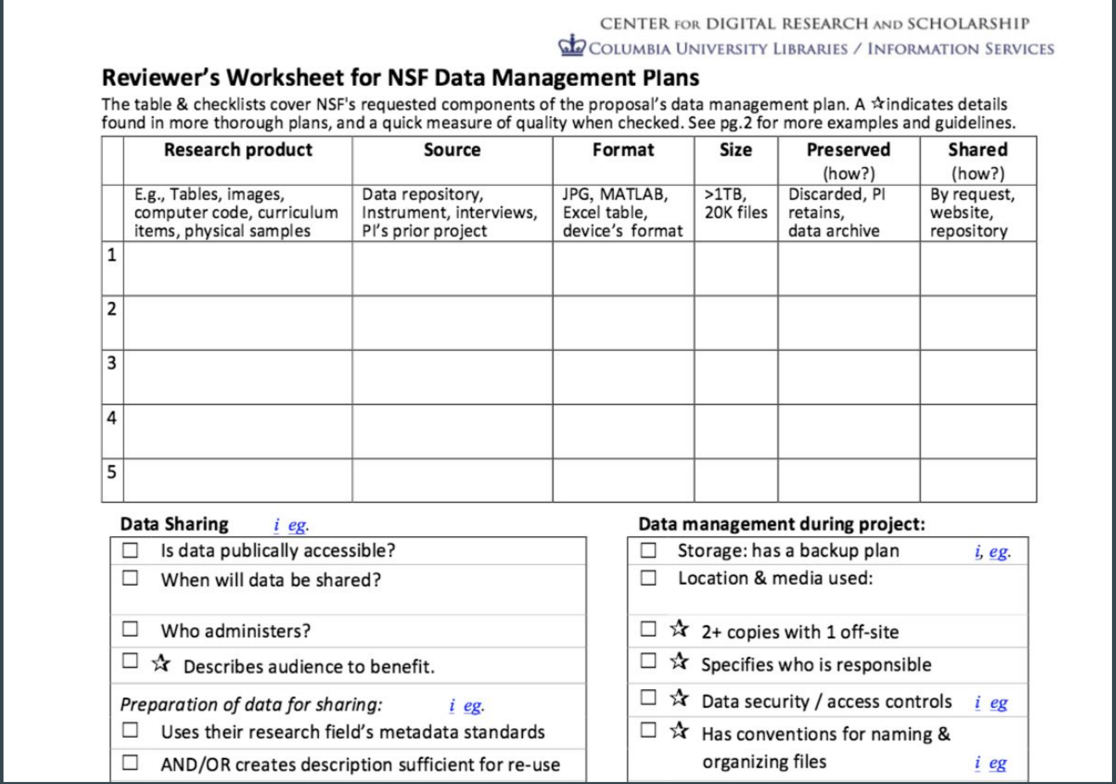
Name: StollBolksArmsAcquisitionOrigi ... View/Open
Size: 52.59Kb
Format: application/zip

Name: StollBolksArmsAcquistionNonpro ... View/Open
Size: 52.44Kb
Format: application/zip

https://scholarship.rice.edu/handle/1911/77660

4. Evaluate a sample plan

How to Evaluate a DMP



CENTER FOR DIGITAL RESEARCH AND SCHOLARSHIP
COLUMBIA UNIVERSITY LIBRARIES / INFORMATION SERVICES

Reviewer's Worksheet for NSF Data Management Plans

The table & checklists cover NSF's requested components of the proposal's data management plan. A ☆indicates details found in more thorough plans, and a quick measure of quality when checked. See pg.2 for more examples and guidelines.

	Research product	Source	Format	Size	Preserved (how?)	Shared (how?)
	E.g., Tables, images, computer code, curriculum items, physical samples	Data repository, Instrument, interviews, PI's prior project	JPG, MATLAB, Excel table, device's format	>1TB, 20K files	Discarded, PI retains, data archive	By request, website, repository
1						
2						
3						
4						
5						

Data Sharing *i eg.*

□ Is data publically accessible?
□ When will data be shared?
□ Who administers?
□ ☆ Describes audience to benefit.

Preparation of data for sharing: *i eg.*

□ Uses their research field's metadata standards
□ AND/OR creates description sufficient for re-use

Data management during project:

□ Storage: has a backup plan *i, eg.*
□ Location & media used:
□ ☆ 2+ copies with 1 off-site
□ ☆ Specifies who is responsible
□ ☆ Data security / access controls *i eg*
□ ☆ Has conventions for naming & organizing files *i eg*

Center for Digital Research & Scholarship, Columbia University Libraries, "Reviewer's Worksheet for NSF Data Management Plans"

Exercise: Let’s evaluate a sample plan

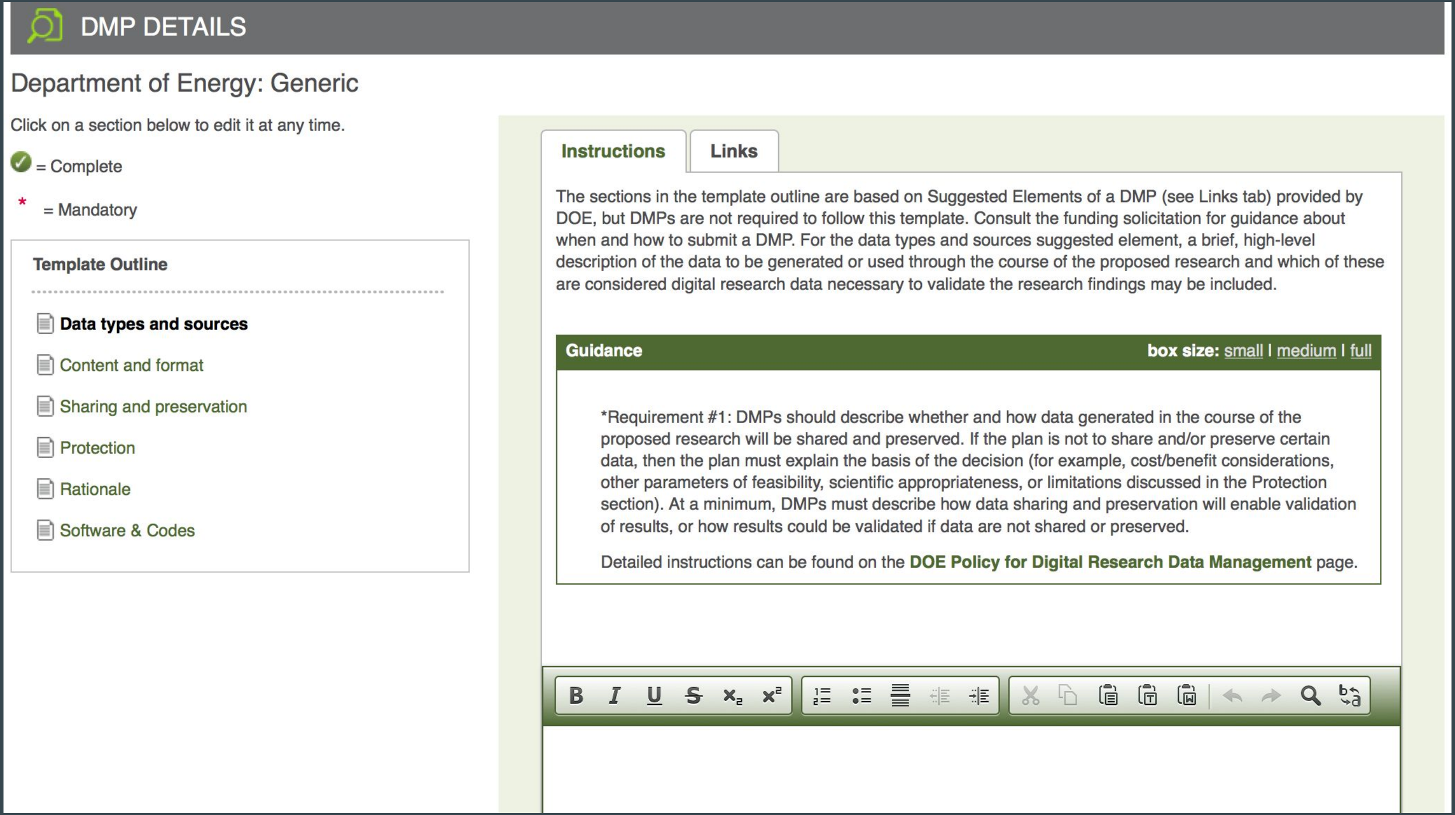
Use the “Reviewers’ Worksheet” to evaluate either “Rio Grande Basin” or the workshop on Afro-Caribbean Labor (NEH) [10 minutes]

Consider:

- What are this plan’s strengths? Weaknesses?
- What is your overall evaluation?

5. Experiment with DMP Tool

Creating DMPs Using DMPTool



DMP DETAILS

Department of Energy: Generic

Click on a section below to edit it at any time.

✓ = Complete

* = Mandatory

Template Outline

- **Data types and sources**
- Content and format
- Sharing and preservation
- Protection
- Rationale
- Software & Codes

The sections in the template outline are based on Suggested Elements of a DMP (see Links tab) provided by DOE, but DMPs are not required to follow this template. Consult the funding solicitation for guidance about when and how to submit a DMP. For the data types and sources suggested element, a brief, high-level description of the data to be generated or used through the course of the proposed research and which of these are considered digital research data necessary to validate the research findings may be included.

Guidance **box size:** small | medium | full

*Requirement #1: DMPs should describe whether and how data generated in the course of the proposed research will be shared and preserved. If the plan is not to share and/or preserve certain data, then the plan must explain the basis of the decision (for example, cost/benefit considerations, other parameters of feasibility, scientific appropriateness, or limitations discussed in the Protection section). At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

Detailed instructions can be found on the **DOE Policy for Digital Research Data Management** page.

https://dmptool.org

Exercise: Sketch out a DMP

- Log into https://dmptool.org
- Select the NSF-Earth Sciences template.
- Create a draft DMP for “Rio Grande...” Try to improve upon the plan that you’ve been provided.
- Alternatively, you can create a DMP for your own (real or imagined) project using the appropriate template.

6. Data Management Resources at Rice & Beyond

Help Provided by the Rice Research Data Management Team

- Assistance developing data management plans.
- Consultation on organizing and managing data.
- Assistance identifying appropriate data repositories.

>> W: http://researchdata.rice.edu/

>> E: researchdata@rice.edu

Help Provided by the Office of Proposal Development

- Assist in developing your proposal, including the DMP
- Identify components that should be included in the DMP
- Draft the non-technical parts of the DMP
- Review, edit, and format the final version of the DMP
- Connect you with other data management resources on campus and online

>>Office of Proposal Development

DMP Components*

NSF - program solicitation or NSF GPG

NIH - FOA or Application Guide

DOE - FOA or Statement of Digital Data Management

*good idea to reference elements of research plan

Another Resource: Office of Research Compliance

Help Provided by Rice's Center for Research Computing

- "Operating best-in class on-premise shared compute, visualization and data-storage facilities;
- Facilitating access to on-premise, regional, national and commercial cloud facilities;
- Delivering user services and training for best use of shared facilities;
- Offering application and proposal consulting support-services."

Helpful Resources

- Borer, Elizabeth T., et al "Some Simple Guidelines for Effective Data Management." *Bulletin of the Ecological Society of America* (2009): 205–14. doi:10.1890/0012-9623-90.2.205.
- Data Carpentry and Software Carpentry
- Data One, Primer on Data Management
- NISO Primer, Research Data Management
- U of Oregon Libraries, Research Data Management Best Practices
- UK Data Service Costing Tool
- UNC Research Data Toolkit: Example Language
- USGS Data Management

More Helpful Resources

- DataOne Primer on Data Management
- Dataverse, Data Management Plans
- ICPSR Guide to Social Science Data Preparation and Archiving
- Oak Ridge National Lab Distributed Active Archive Center, Best Practices for Preparing Environmental Data Sets to Share and Archive
- Svend Juul et al, “Take good care of your data”
- UK Data Archive, Managing and Sharing Data: Best Practices for Researchers
- White, Ethan P., et al “Nine Simple Ways to Make It Easier to (re)use Your Data.” *Ideas in Ecology and Evolution* (8/30/ 2013).