Data Management Essentials

Jean Aroom, Lisa Spiro, Kathy Weimer and Chandler Wilkerson

researchdata@rice.edu
Why manage your data?
Why manage your data?

• Prevent loss of data.
• Better organize your data, saving time and enabling analysis.
• Promote good research practices, such as replication of results.
• More easily share your data, so that it can then be cited and foster future discoveries.
• Comply with the requirements of funding organizations.
Examples of Funding Agency Guidelines

- NSF, such as
  - Biological Science
  - Geosciences
- NIH Data Sharing Policy
- Department of Energy
- NEH Office of Digital Humanities
- NASA Earth Science
- NOAA
- Gordon & Betty Moore Foundation
Some Principles Underlying Data Management/ Sharing Requirement

• Values openness as important to scientific progress and trustworthiness.
• Respects norms of disciplinary communities.
• Recognizes constraints such as privacy and 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 research data and 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
Case Study: Rio Grande Basin Hydrologic Geodatabase Compendium

- General NSF Data Management Plan from DataOne.
- Focuses on compiling a geodatabase compendium of water resources data for the Rio Grande.
- Supports visualizing and analyzing water quality data at specific sites.
Components of a Data Management Plan (NSF-Generic)

1. Types of Data Produced
2. Data and Metadata Standards
3. Policies for Access and Sharing
4. Policies for Re-use and Distribution
5. Plans for Archiving and Preservation

Questions included with each section were drawn from DMP Tool.
1. Types of Data Produced

What data will be generated in the research?

Surface water, ground water, and water quality data for the middle Rio Grande basin study area.

How will you capture or create the data?

All datasets are acquired from the original data source (for example, EPA websites). Data will be processed using dataset-specific VBA programs.

Which file formats will you use for your data, and why?

A relational spatially-enabled Microsoft Access database, Visual Basic for Applications (VBA) code, and Structured Query Language (SQL) code.

How will sensitive or proprietary data be handled?
2. Metadata Standards

What contextual details (metadata) are needed to make the data you capture or collect meaningful?

- A data dictionary.
- An entity-relationship diagram defining database structure.
- Program file comment headers.
- Annotated code to promote code readability.

What metadata standards will be implemented?

A metadata record will be produced using the FGDC standard that describes the entire geodatabase. The FGDC standard was chosen due to required Federal government standards.
3. Public Access & Data Sharing

How, when, and where will you make the data available?

*Data are public and will be obtainable thru the New Mexico Interstate Stream Commission (NMISC).*

Who are the foreseeable data users?


How should your data be cited?

*USGS publications will be released describing the methods and data sources and can be used as documentation for the data and to cite the data.*
Ways to Share Data

• Deposit in an open repository [preferred]
  – Disciplinary (e.g. ICPSR, GenBank)
  – Institutional (e.g. Rice Digital Scholarship Archive)

• Publish as supplementary data with your article [also recommended]
  – e.g. Dryad

• Post to a personal/ research group website
  – How (long) will website access be maintained?

• Make available upon request (e.g. via email)
  – Downsides: harder for other researchers to discover and cite and for PI to manage
Note Any Restrictions on Data Sharing

- **Private/confidential data** (e.g. Human Subjects, governed by IRB, HIPAA, FERPA)
  - Consider obtaining informed consent for sharing data
  - De-identify/clean data

- **Intellectual property**, e.g. patent & copyright
  - Data can be embargoed for a period of time

- **Sensitive data**, e.g. locations of endangered species

- **Export controls**
4. Policies for Re-use and Distribution

What are the foreseeable uses of the data?

Access to databases and associated software tools will be available for educational, research and non-profit purposes.

Will any permission restrictions need to be placed on the data?

Materials generated under the project will be disseminated in accordance with University/Participating institutional and NSF policies. Materials may be transferred to others under the terms of a material transfer agreement.
Make Explicit **Re-use Policies**

- Data can’t be copyrighted, but expressions of it (e.g. tables) can be if sufficiently creative.
- Data can be [licensed](https://www.creativecommons.org/licenses/by/), so that you set terms of use (e.g. requiring attribution).
- You can release data into the public domain through the [CC-0 license](https://creativecommons.org/publicdomain/zero/1.0/).
5. Plans for Archiving and Preservation

What procedures does your intended long-term data storage facility have in place for preservation and backup?

*Data versioned over time & maintained in a date-stamped file structure with text files documenting the provenance.*

Which archive have you identified as a place to deposit data?

1) New Mexico Interstate Stream Commission Central Office
2) NMISC field office
3) Consortium of Universities for the Advancement of Hydrologic Science

What transformations will be necessary to prepare data for preservation and sharing?

*Database exported to a delimited text format.*
HOW RICE SUPPORTS DATA MANAGEMENT
Consult with Rice’s Research Data Management Team

http://researchdata.blogs.rice.edu/
Rice’s Research Data Management Team Can Advise On:

- Developing data management plans
- Appropriate metadata and data formats
- Best practices for curating your data
- Storage options
- Disciplinary repositories where you can deposit your data
- Making your final data publicly accessible through the Rice Digital Scholarship Archive
Create Your Plan Using the **DMPTool**

Log in using your Rice NetID and password

[Image of DMPTool interface]

https://dmptool.org/
## Options for Storing Your Data

### Rice Storage Security Levels

<table>
<thead>
<tr>
<th>Category</th>
<th>Google Drive</th>
<th>Indiv. User Drive</th>
<th>Dept Share</th>
<th>RNAS</th>
<th>OWL-Space</th>
<th>Box</th>
<th>SPICE</th>
<th>Crate</th>
<th>Archive</th>
<th>Rice Dropbox File Drop</th>
<th>Crashplan Back up</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty</td>
<td>30GB</td>
<td>5GB</td>
<td>40GB shared</td>
<td>varies</td>
<td>by request</td>
<td>30GB</td>
<td>varies</td>
<td>500GB per research award</td>
<td>500GB per research award</td>
<td>1GB per file</td>
<td>Unlimited</td>
</tr>
<tr>
<td>Staff</td>
<td>30GB</td>
<td>5GB</td>
<td>40GB shared</td>
<td>varies</td>
<td>by request</td>
<td>30GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grad Student</td>
<td>30GB</td>
<td>5GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Undergrads</td>
<td>30GB</td>
<td>2GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Yellow** = Located in Rice, most secure if devices secured.  
**Red** = Covered by Rice contract, semi-secure
Come to Our Full Data Management Workshop on 4/7/15

Creating a Data Management Plan

This hands-on workshop will introduce participants to the process of writing a data management plan, as required by the NSF and other funding agencies. It will also share some tips for effectively organizing, storing and sharing research data.

<table>
<thead>
<tr>
<th>When</th>
<th>Apr 07, 2015 from 02:00 PM to 03:30 PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where</td>
<td>Fondren Library Basement B43A</td>
</tr>
<tr>
<td>Contact Name</td>
<td>Lisa Spiro</td>
</tr>
<tr>
<td>Contact Phone</td>
<td>713-348-2480</td>
</tr>
<tr>
<td>Add event to calendar</td>
<td>vCal, iCal</td>
</tr>
</tbody>
</table>

https://library.rice.edu/events/creating-a-data-management-plan
Resources on Research Data Management

- DataONE Primer on Data Management
- Penn State Data Management Plan Tutorial
- ICPSR Data Management & Curation
- University of Minnesota Data Management