REMEDIATION DATA MANAGEMENT PLANS
A TOOL FOR RECOVERING RESEARCH DATA FROM MESSY, MESSY PROJECTS

IDCC 2018
Barcelona, 19-22 February 2018
The Watershed Research Cooperative story…

PROPOSAL

“it is apparent that there is a need for additional investment in data management”

“if the benefits from the WRC studies are to be realized the project needs help”

Watershedresearch.org
Context of DMPs

- geographically disperse,
- interdisiplinary,
- data based
- accountability for federal granted projects

“The Office of Science and Technology Policy (OSTP) hereby directs each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government. “
Context DMPs

Elements of a Data Management Plan

1. Data types and formats
2. Roles and responsibilities
3. Data organization
4. Data documentation
5. Storage, backup and data security
6. Archiving, preserving and sharing

A DMP is a plan. Its role is to prevent bad data management practices and encourage good data stewardship.
Context DMPs

What about data that has already been collected?

Data that is already messy?
How to manage data?

A remedial Data Management Plan

1. Data inventory
2. Data types and formats
3. Roles and responsibilities
4. Data organization
5. Data documentation
6. Storage and backup
7. Archiving, preserving and sharing
8. Implementation and priorities

For whom?

https://www.dataone.org/data-life-cycle
### pDMP vs rDMP

<table>
<thead>
<tr>
<th>Differences</th>
<th>pDMP</th>
<th>rDMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Audience and goal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Data inventory</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Implementation strategy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Commonalities</th>
<th>pDMP</th>
<th>rDMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Data organization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Data documentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Sharing and preservation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Differences pDMP rDMP: audience

- A lot of work to do: resources will be needed. Goal: plan best strategies, describe the magnitude of the problem, present priorities.

FOR ADMINISTRATORS AND MANAGERS

- Many researchers that have internal dynamics and conflicts. Goal: external consultation to start a conversation about data management practices and enforcing.

FOR ADMINISTRATORS AND MANAGERS

**rDMP**
- researchers

**pDMP:**
- researchers

OR
- administrators, managers

OREGON STATE UNIVERSITY 7
2. Diff pDMP rDMP: Data inventory

1. Relational databases: 3 relational databases managed by Data Manager
   - Climate
   - Hydrology
   - Temperature
   - Nutrients
   - Dissolved oxygen
   - Sediment

2. Trask tabular data: use of LTER templates

3. Fish database: managed by Researcher acting as Data Manager

4. Other digital data
   - Macroinvertebrates
   - Geospatial

5. Physical samples
Total of 409158 files and 344 GB of data

- 50000 excel files (73GB)
- 52000 images (79 GB)
- 334 databases (17 GB)
- 9000 documents pdf ppt (30 GB)
- 17000 geospatial (22GB)

Mix of
- Data
- Admin
- Personal folders
- Publications
- Backups
## 2. Diff pDMP rDMP: Data inventory

<table>
<thead>
<tr>
<th>Collect metadata about data</th>
<th>Detail according to the size of the project</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Subject</td>
<td>Small projects</td>
</tr>
<tr>
<td>- Location</td>
<td></td>
</tr>
<tr>
<td>- Responsibilities</td>
<td>File to file?</td>
</tr>
<tr>
<td>- Manager</td>
<td>Medium/large projects</td>
</tr>
<tr>
<td>- Versions</td>
<td></td>
</tr>
<tr>
<td>- Formats</td>
<td>Data collection categories</td>
</tr>
<tr>
<td>- Documentation</td>
<td></td>
</tr>
<tr>
<td>- Sensitivity</td>
<td>File analysis (automated).</td>
</tr>
<tr>
<td>- Sharing status</td>
<td></td>
</tr>
</tbody>
</table>

- Detail according to the size of the project

  - Small projects
    - File to file?
  
  - Medium/large projects
    - Management level categories
    - Data collection categories
    - File analysis (automated).
3. Diff pDMP rDMP: implementation

- **Priority 1**: Clean, document, and preserve in ScholarsArchive@OSU quality controlled datasets.
  - Use article publications as triggers to do the work.
- **Priority 2**: Clean, document and preserve data associated to past publications, including Hinkle.
- **Priority 3**: Triage data in shared drive folders.

Challenge: HOW? HOW MUCH? WHO?

**rDMP**

Something needs to change!
- Priorities
- Resources
- Individuals
- Motivate researchers: carrots or sticks?
Commonalities: Data organization; Documentation

- Levels of data quality:
  - Level zero (L0): raw data downloaded directly from instrument or model.
  - Level one (L1): Raw data in a format that is understandable by the researcher.
  - Level two (L2): Verified data that have undergone quality control
  - Level three (L3): L2 data that have been analyzed to answer specific research questions. Typically used for figures in a publication.

- Organization in datasets
- Folder structure
- File naming strategy
- Metadata standard: Ecological Metadata Language
- Define scenarios (minimum & ideal) for data documentation.
Commonalities:
Preservation strategy

- Preservation in OSU’s institutional repository ScholarsArchive@OSU.
- Tidy data, open format, documented
- Licenses

Welcome to the new ScholarsArchive@OSU! If you have any questions about how to upload your content, please see: http://bit.ly/2zyxAXI or use the Contact Form.
Conclusion

DMP useful structure
Some differences that need to be taken into account.
Also useful for other stages in the data life cycle? Maybe an archival Data Management Plan?

Clara Llebot Lorente
Data Management Specialist
clara.llebot@oregonstate.edu

https://www.dataone.org/data-life-cycle