DataONE Webinar Series

Analyzing, interpreting, and implementing data management plans

Amanda Whitmire
Heidi Imker
Sarah Jones
Enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it

DataONE network of Member Nodes: www.dataone.org/current-member-nodes

DataONE Search tool: https://search.dataone.org
Lesson 1: Data Management

Why manage data?

- The researcher perspective
  - Keep yourself organized
  - Find your own files
  - Track your processes for reproducibility
  - Better version control of data
  - More efficient data management
  - More backup to avoid data loss
  - Format your data for reuse by yourself and others
  - Document your data for understandability and reuse
  - Prepare it to share it and gain credibility and recognition for your scientific efforts

- Data management facilitates sharing and reuse.

The Case for Data Management

- If data are:
  - Well-organized
  - Documented
  - Preserved
  - Accessible
  - Verified as to accuracy & validity
- The results are:
  - High quality data
  - Data that is easy to share and reuse
  - Citation & credibility to researchers
  - Cost savings to further science

Lesson 2: Data Sharing

Describe data content, character, and process. Describe the research project, the data it will produce, and how these costs will be covered. How metadata will be created and/or captured?

The world of data around us

The digital deluge has created a surge of information that needs to be well-managed, discoverable, and accessible.

The amount of available storage is not keeping pace with the amount of data being produced.

Information vs. Available Storage

<table>
<thead>
<tr>
<th>Information</th>
<th>Available Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 GB</td>
<td>10 GB</td>
</tr>
<tr>
<td>1 TB</td>
<td>1 TB</td>
</tr>
</tbody>
</table>

Causes of data loss

- Natural disasters
- Facilities infrastructure failures
- Storage failures
- Server hardware or software failure
- Application software failure
- Human errors
- Malicious attack
- Format obsolescence
- Loss of competencies
- Loss of funding
- Loss of institutional commitment

Costs of not doing data management can be very high!

The Data Lifecycle

The stages through which research data pass from project inception to conclusion.

Data Reuse Example

Researchers reused and aggregated data from several different sources to determine migration routes for specific bird species.

Component 2: Metadata content & format

- What metadata are needed?
- How will metadata be created and/or captured?

Component 3: Data Management

- Information about data & data formats
  - Details, instruments, parameters, units, files, etc.
- How data will be acquired
- How data will be processed
- Description of data to be produced

Component 4: Long-term storage & management

- Anticipated costs for data
- Who will be responsible for the long term?
- What data transformations/formats will be used?
- Where will it be preserved?
- What data will be preserved?
- How will metadata be created and/or captured?

Component 5: Budget

- How these costs will be covered
- Anticipated costs for data

Concerns about data sharing

To the scientist

- Better education for new researchers on the importance of data management
- Ability to perform meta analyses
- Greater opportunity for data exchange
- Improved scientific network connections & potential collaborations

To the public

- Better access to information leads to better understanding and contributions toward solving big scientific problems
- Increased potential for new discoveries
- Increased awareness of important scientific projects

To the research sponsor/funder

- Increased efficiency and productivity
- More accurate research results
- Better version control of data
- More efficient data management
- Better version control of data
- More backup to avoid data loss
- Format your data for reuse by yourself and others
- Document your data for understandability and reuse
- Prepare it to share it and gain credibility and recognition for your scientific efforts
- Better version control of data
- More backup to avoid data loss
- Format your data for reuse by yourself and others
- Document your data for understandability and reuse
- Prepare it to share it and gain credibility and recognition for your scientific efforts

Metadata in context: containing and using it

Metadata is the glue in a data workflow. It contains information about the data so that others can understand and re-use your data in the present and the future.
DataONE Webinar Series
www.dataone.org/webinars
Upcoming Webinar Event
www.dataone.org/upcoming-webinar
Previous Webinar Events (Recording and Discussion)
www.dataone.org/previous-webinars

#DWS2016
@DataONEorg
Upcoming Webinar Event
www.dataone.org/upcoming-webinar

DPM “Stack”: A Management Infrastructure Frame for Digital Preservation that Parallels Technical Infrastructure

December 13
Nancy McGovern
Massachusetts Institute of Technology
If you attending as part of a group, please enter the number of people listening within the “questions” box. Thanks!
Analyzing, Interpreting, and Implementing Data Management Plans

Amanda Whitmire
Stanford University

Heidi Imker
University of Illinois, Urbana-Champaign

Sarah Jones
Digital Curation Centre
DataONE Webinar Series
www.dataone.org/webinars
Upcoming Webinar Event
www.dataone.org/upcoming-webinar
Previous Webinar Events (Recording and Discussion)
www.dataone.org/previous-webinars

#DWS2016
@DataONEorg
Data Management Plans
What good are they to us?

Amanda L. Whitmire, Ph.D.
Head Librarian & Bibliographer
Harold A. Miller Library
Hopkins Marine Station
Stanford University

@AWhitTwit
What are DMPs good for?

1. Information about researcher habits for data services development - Amanda

2. DMP consultation as gateway service to launch more meaningful interactions - Heidi

3. Overseas perspective; how universities have embedded DMP services into existing workflows & systems - Sarah
DMPs as source of researcher intel
Data management plan As Research Tool (DART Project)

Amanda Whitmire | Stanford University Libraries
Jake Carlson | University of Michigan Library
Patricia M. Hswe | Update
Lizzy Rolando | MailChimp
Susan Wells Parham | Georgia Institute of Technology Library
Brian Westra | University of Oregon Libraries

This project was made possible in part by the Institute of Museum and Library Services grant number LG-07-13-0328.

@DMPResearch
<table>
<thead>
<tr>
<th>General Assessment Criteria</th>
<th>Performance Level</th>
<th>Complete / detailed</th>
<th>Addressed issue, but incomplete</th>
<th>Did not address issue</th>
<th>Directorates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Describes what types of data will be captured, created or collected</strong></td>
<td></td>
<td>Clearly defines data type(s). E.g. text, spreadsheets, images, 3D models, software, audio files, video files, reports, surveys, patient records, samples, final or intermediate numerical results from theoretical calculations, etc. Also defines data as: observational, experimental, simulation, model output or assimilation</td>
<td>Some details about data types are included, but DMP is missing details or wouldn’t be well understood by someone outside of the project</td>
<td>No details included, fails to adequately describe data types.</td>
<td>All</td>
</tr>
<tr>
<td><strong>Describes how data will be collected, captured, or created (whether new observations, results from models, reuse of other data, etc.)</strong></td>
<td></td>
<td>Clearly defines how data will be captured or created, including methods, instruments, software, or infrastructure where relevant.</td>
<td>Missing some details regarding how some of the data will be produced, makes assumptions about reviewer knowledge of methods or practices.</td>
<td>Does not clearly address how data will be captured or created.</td>
<td>GEO AGS, GEO EAR SGP, MPS AST</td>
</tr>
<tr>
<td><strong>Identifies how much data (volume) will be produced</strong></td>
<td></td>
<td>Amount of expected data (MB, GB, TB, etc.) is clearly specified.</td>
<td>Amount of expected data (GB, TB, etc.) is vaguely specified.</td>
<td>Amount of expected data (GB, TB, etc.) is NOT specified.</td>
<td>GEO EAR SGP, GEO AGS</td>
</tr>
</tbody>
</table>
500 DMPs
100 from each institution

Used Qualtrics survey to collect data

Distribution across NSF directorates followed distribution of funded proposals
The DART Project: using data management plans as a research tool

Contributors: Amanda Whitmire, Jacob Carlson, Brian Westra, Patricia Hawe, Susan Parham
Date created: 2015-10-15 02:44 PM | Last Updated: 2016-06-22 04:36 PM
Category: Project

Description: This is a three-year National Leadership Grant for Libraries Demonstration Project to facilitate a multi-university study of faculty data management plans (DMPs). The primary outputs of this project will be an analytic rubric to standardize the review of data management plans as a means to inform targeted expansion or development of research data services at academic libraries; and a study utilizing the rubric that presents the results of data management plan analyses at five universities. This project was made possible in part by the Institute of Museum and Library Services grant number LG-07-13-0328.

License: CC0 1.0 Universal

Find the rubric
See the survey we used to collect assessment data
Look at our DMP assessment data

https://osf.io/kh2y6/
Describes what type(s) of data produced

<table>
<thead>
<tr>
<th></th>
<th>Complete/detailed</th>
<th>Addressed issue, but incomplete</th>
<th>Did not address</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>68</td>
<td>22</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>BIO</td>
<td>83</td>
<td>11</td>
<td>6</td>
<td>75</td>
</tr>
<tr>
<td>CISE</td>
<td>53</td>
<td>32</td>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td>ENG</td>
<td>74</td>
<td>20</td>
<td>7</td>
<td>45</td>
</tr>
<tr>
<td>GEO</td>
<td>64</td>
<td>29</td>
<td>7</td>
<td>30</td>
</tr>
<tr>
<td>MPS</td>
<td>61</td>
<td>24</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>SBE</td>
<td>82</td>
<td>12</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>

Describes how data will be shared

<table>
<thead>
<tr>
<th></th>
<th>Complete/detailed</th>
<th>Addressed issue, but incomplete</th>
<th>Did not address</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>50</td>
<td>41</td>
<td>9</td>
<td>70</td>
</tr>
<tr>
<td>BIO</td>
<td>67</td>
<td>31</td>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>CISE</td>
<td>41</td>
<td>44</td>
<td>15</td>
<td>50</td>
</tr>
<tr>
<td>ENG</td>
<td>37</td>
<td>53</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>GEO</td>
<td>61</td>
<td>35</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>MPS</td>
<td>48</td>
<td>44</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>SBE</td>
<td>58</td>
<td>34</td>
<td>8</td>
<td>0</td>
</tr>
</tbody>
</table>

### Where will they share data?

<table>
<thead>
<tr>
<th>Category</th>
<th>All</th>
<th>BIO</th>
<th>CISE</th>
<th>ENG</th>
<th>GEO</th>
<th>MPS</th>
<th>SBE</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal / supplement</td>
<td>36</td>
<td>27</td>
<td>23</td>
<td>45</td>
<td>35</td>
<td>54</td>
<td>18</td>
<td>80</td>
</tr>
<tr>
<td>Data center or repository</td>
<td>34</td>
<td>75</td>
<td>14</td>
<td>8</td>
<td>66</td>
<td>25</td>
<td>42</td>
<td>70</td>
</tr>
<tr>
<td>On request</td>
<td>30</td>
<td>23</td>
<td>30</td>
<td>38</td>
<td>24</td>
<td>34</td>
<td>26</td>
<td>60</td>
</tr>
<tr>
<td>Personal website</td>
<td>25</td>
<td>13</td>
<td>44</td>
<td>31</td>
<td>25</td>
<td>20</td>
<td>12</td>
<td>50</td>
</tr>
<tr>
<td>Other method</td>
<td>22</td>
<td>27</td>
<td>30</td>
<td>15</td>
<td>23</td>
<td>18</td>
<td>22</td>
<td>40</td>
</tr>
<tr>
<td>Institutional repository</td>
<td>17</td>
<td>6</td>
<td>12</td>
<td>20</td>
<td>6</td>
<td>28</td>
<td>20</td>
<td>35</td>
</tr>
<tr>
<td>Conference / proceedings</td>
<td>13</td>
<td>8</td>
<td>11</td>
<td>23</td>
<td>8</td>
<td>13</td>
<td>8</td>
<td>30</td>
</tr>
<tr>
<td>Did not specify</td>
<td>8</td>
<td>0</td>
<td>18</td>
<td>9</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>Thesis / Dissertation</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Not planning to share</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Book</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

What is going on with Biology?
They have infrastructure!

**BIO:** Repositories mentioned (frequency)

- GenBank (14)
- Dryad (12)
- SRA (11)
- iDigBio (3)
- Knowledge Network for Biocomplexity (3)
- MorphBank (3)
- NCBI (3)
- TreeBASE (2)
-Question-
RESEARCH ARTICLE

Water, Water, Everywhere: Defining and Assessing Data Sharing in Academia

Steven Van Tuyl1,*, Amanda L. Whitmire2

1 Center for Digital Scholarship and Services, University Libraries, Oregon State University, Corvallis, Oregon, United States of America, 2 Harold A. Miller Library, Hopkins Marine Station, Stanford University, Pacific Grove, California, United States of America

These authors contributed equally to this work.

* steve.vantuyl@oregonstate.edu


What we did

1. Define criteria for assessing the effectiveness of data sharing
   1. Discoverable?
   2. Accessible?
   3. Transparent?
   4. Actionable?

2. Used DMPs & publications from NSF-funded work to look for associated datasets
Fig 1. Total DATA scores from 25 NSF-funded projects, as located via data management plans

http://journals.plos.org/plosone/article?id=info:doi/10.1371/journal.pone.0147942
DMP review deepens our understanding & allows for more targeted support.
A Case Study of DMP Implementation

- Peg Burnette
  Biomedical Librarian
  Social Sciences, Health, and Education Library

- Sarah Williams
  Life Sciences Data Services Librarian
  Funk ACES Library

- Heidi Imker (presenting)
  Director
  Research Data Service
  University Library
Illinois Research Data Service

- Funded by campus administration in 2013 and based at the University Library
- Core staff of 4 FTE + “voluntary” efforts of many others
- Regular interactions with data-related campus groups, e.g. central IT, supercomputing, IRB, security, OVCR, etc.
- Depend on the expertise of our library colleagues for on-the-ground interactions, including DMP reviews and data management consultations
Set-up

- Contacted in 2013 for a DMP review (pre-RDS)
- Contacted in 2014 when the grant was funded
- Did one preliminary consultation and based on the questions, pulled in 3 others and followed-up with a report and several “check-point” consultations

- Seemed to be going well! How? Why?
- http://dx.doi.org/10.7191/jeslib.2016.1101
Theme - Basic Elements

- Communication
- Planning for data management
- Documentation
- Failsafe efforts and checks
Theme - Staffing

• Hired a project coordinator that served as the data point person
• Effort on setting up protocols and documentation for data management was front loaded, and thereafter was just checking.
• This person served many roles!

➢ Smaller projects (e.g. not enough need or funds for a dedicated project manager), explicitly assign duties and expectations.
Theme - Data Quality

• Good data is core good science
• Getting better at data management is part of getting better at research itself
• A sort of experiential learning
Theme - Mentoring

Who taught you how to do this stuff?

Box of unlabeled USB drives found in drawer

https://www.youtube.com/watch?v=KUXb7do9C-w
Theme - Mentoring

- Leader came from a background in strong data management
- Co-PIs both expressed that they’re learning from colleagues and as they go
- Project manager actually seemed baffled when I asked if she has trouble getting people to follow data management protocols

- Creating those expectations and accountability is work
Theme - Peace of Mind

• Expression of emotional distress over possibility of poorly managed data
• High need to be efficient as faculty
• Thought of as an investment with hard-to-measure returns – and that was okay
What have we learned?

• Basic elements are present and core
• Some of the “softer” skills were a little surprising
• Hadn’t occurred to us to think so explicitly about the role of mentoring in data management
  • Can we use mentorship as a way to frame how to set up proactive data management practices?
  • Write into DMPs that researchers on the grant will attend data management training from the RDS.
What have we gained?

• New perspectives
• Validation for what we’re trying to accomplish
• Ammo: if you don’t believe us – believe *them*, your own colleagues at your own institution. It can be done.
Thank you!

Heidi Imker imker@illinois.edu
Peg Burnette phburn@illinois.edu
Sarah Williams scwillms@illinois.edu
Supporting DMPs: lessons from Europe

Sarah Jones
Digital Curation Centre, Glasgow
sarah.jones@glasgow.ac.uk
Twitter: @sjDCC

Analyzing, interpreting & implementing DMPs, DataOne webinar, 8 November 2016
Heavy requirements landscape

• Research Councils and charity funders require DMPs.
• 74% of uni RDM policies also mandate DMPs*

* See Laurence Horton analysis at www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies
DMP trends

- Increasing drive towards openness
- DMPs as living documents
- FAIR data management
Basic uni support

- Many unis offer custom guidance (at institutional and in some cases school / department level) as well as example answers.
DMP consultations

Consulting, supporting and networking with researchers & all other interest groups

Slide content courtesy of Mari Elisa Kuusniemi (MEK), University of Helsinki Library
DMP feedback and review

www.data.cam.ac.uk/DMPsupport
UK adoption of DART approach

• Community-led initiative to develop evaluation rubrics based on key funder requirements

• Emphasis on funder specifics to check compliance and provide feedback pre-submission

• Example BBSRC rubric:
  • https://research-data-network.readme.io/docs/bbsrc-dmp-compliance-rubric
Review functionality in DMP tools

• Revising the review functionality in DMPRoadmap (joint codebase for DMPonline & DMPTool)

• Opting for simpler, less-formal process

• Use of substance editor to annotate text directly

• Share what you need!
Collaboration with research offices

- Research offices play a key role as the first point of contact when PIs are preparing grants
  - Include links to DMP tools in research office mailings e.g. replies to costing requests
  - RO provides list of new awards so RDM team can contact PIs about DMPs at outset
  - Hot-desking / co-location so teams work more closely and share expertise
  - Collaboration on costing data management
Integrating DMPs into workflows

Example of embedding flags into the grant costings system at the University of Leicester
DMPs to define / allocate storage

• The University of Manchester requires an outline DMP prior to a grant application.

• This asks questions about the storage requirements to plan resourcing and allocate space.

• The outline DMP generates an RDM Plan Reference Number to include in the Research Application form. Proposals can't proceed without this.

• [www.library.manchester.ac.uk/services-and-support/staff/research/services/research-data-management/data-management-planning-tool](http://www.library.manchester.ac.uk/services-and-support/staff/research/services/research-data-management/data-management-planning-tool)
Desire for pre- or post- phases

Idea to blend institutional and funder requirements, ensure costings are included and plans implemented

- Resource planning
  - Institutional questionnaire
    - Data volumes
    - Sensitive data
    - Costings
  - Proposal
    - Funder DMP requirement (usually grant application)
  - Implementing
    - In-project planning
      - Ethics process
      - Service requests
      - Internal audit

Idea to blend institutional and funder requirements, ensure costings are included and plans implemented
Thanks for listening

- DCC resources on DMPs:
  - www.dcc.ac.uk/resources/data-management-plans

- Follow us on twitter:
  - @DMPonline and #ukdcc