Best Practices for Good Data Management

February 19, 2015
Presenters

Andrew Johnson
Research Data Librarian
University Libraries
University of Colorado Boulder
andrew.m.johnson@colorado.edu

Shelley Knuth
Senior Research Data Specialist
Research Computing
University of Colorado Boulder
shelley.knuth@colorado.edu
Outline

- Introduction and definitions
- Creating a DMP
- Basic elements of a DMP
- Additional DMP resources
- Q&A
What is a DMP?

“A data management plan is a formal document that outlines what you will do with your data during and after you complete your research.”

University of Virginia Research Data Services,
http://data.library.virginia.edu/data-management/whymanage/
What is research data?

White House Office of Management and Budget:
“The recorded factual material commonly accepted in the scientific community as necessary to validate research findings."

Why do we care about managing research data?

Good for science:
- Reproducibility
- Efficiency
- Innovation

Good for you:
- More usage (including citations)
- More exposure to potential collaborators
- More competitive grant applications
How do you get started creating a DMP?

Try the DMPTool: [http://dmptool.org](http://dmptool.org)

- Tool for creating custom data management plans to meet specific funder requirements
- No need to register to use:
  - Click “Get Started”
  - Select “University of Colorado Anschutz Medical Campus”
  - Log in with official university password
What are the basic elements of a DMP?

Basic NIH data management plans (i.e., data sharing plans) include answers to the following questions:

- What types of data will you produce?
- When will you make the data available?
- What file formats will you use for your data, and why?
- What transformations will be necessary to prepare data for preservation/data sharing?
- What metadata/documentation will be submitted alongside the data?
- Will a data-sharing agreement will be required? What will the agreement state?
- What are your plans for providing access to your data?
- Which archive/repository/central database have you identified as a place to deposit data?
What types of data will you produce?

- What data is being created or used as part of this project?
- Is this new data that is difficult to reproduce, or is it data being reused from another source?
- What is the anticipated size of the dataset and expected rate of growth?
When will you make the data available?

- As soon as possible and for as long as possible.
  - Available upon publication of paper that uses data
- Embargo period for authors
  - Allows for people who collected data to get first chance at publication
- Expected that data will be made available by the end of the project if possible
What file formats will you use for your data, and why?

- Non-proprietary file formats are the most appropriate to use to ensure access to the data in the future.

- Proprietary formats:
  - .doc, .docx
  - .ppt, .pptx
  - .xls, .xlsx
  - .mat
  - Avoid their use

- Non-proprietary formats:
  - .txt
  - .jpeg
  - .tiff
  - .csv
What transformations will be necessary to prepare data for preservation/data sharing?

- Will data need to be converted before sharing?
  - Example: use proprietary formats when working with data
  - Convert to non-proprietary before sharing
- Changes from raw data?
  - Example: confidential data needs to be deidentified to ensure privacy
What metadata/documentation will be submitted alongside the data?

- Metadata is data about data
  - Purpose to document important details about data for reuse

- To include in metadata:
  - Project lead name and contact information
  - Date(s) of project
  - Collection protocols
  - Software/instruments used
  - Location of collection

- Can do in a text document or XML

- Follow standard metadata practices when possible
Will a data-sharing agreement be required?

- Are there any restrictions for sharing your data?
  - Do you have confidential data that cannot be shared?
- Are you restricted by proprietary or patent issues?
- If applicable outline these limitations
- Describe any agreement that will be required for others to access your data
What are your plans for providing access to your data?

- How will you share your data?
- How long will your data be available?
- How will people access your data?
  - Dedicated repositories (ideal)
    - List policies of repository in plan
  - Website that you manage
    - Must have clear plan for maintenance
    - Who is in charge?
  - Upon request (not ideal)
Which archive/repository/central database have you identified as a place to deposit data?

- Choose a repository
  - Best will promote open access
  - Have data standards that allows data to be easily shareable
  - Also consider what best suits your data/financial/project needs
- NIH-supported options are listed here:
Where can you get additional help?

NIH Help:

Further information on NIH DMPs:


Suggested repositories:

Thank you! Questions?

- Copyright 2015 by Andrew Johnson and Shelley Knuth
- This work is licensed under a [Creative Commons Attribution 3.0 Unported License](https://creativecommons.org/licenses/by/3.0/).

Twitter: @cu_data
Subscribe to our data events email list: [https://lists.colorado.edu/sympa/info/data_community](https://lists.colorado.edu/sympa/info/data_community)