Best Practices for Research Data Management

October 30, 2014

Presenters

Andrew Johnson
Research Data Librarian
University Libraries

Shelley Knuth
Research Data Specialist
Research Computing

Outline

- What is research data and why do we care about managing it?
- How do I write a data management plan to meet funding agency requirements?
- What are some resources on campus for research data management support?

Research data definitions

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

CU-Boulder Research Data Advisory Committee:
“digital outputs, which include content in structured forms including but not limited to: text files, word processing documents, spreadsheets, websites, calibration information, or simulation outputs; digital information artifacts, such as images, vector-based map products, audio and video products; digital outputs may also include the code used to decode those products, the metadata describing such information artifacts, and required source code that runs computer instructions.”

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

Funding agency requirements

Data Management Plan (DMP) requirements:

- National Science Foundation
- Department of Energy
- USGS
- Other agencies and foundations

More responses to the 2013 White House OSTP public access memo coming soon...

Internal competitions

- Innovative Seed Grants now require DMPs
- Best Digital Data Management Plans and Practices competition
  - 2014 winning plans: https://data.colorado.edu/cudmpguidance
  - Intending to run another competition in 2015

How to create a successful data management plan

- CU Boulder has many services available to you free of charge
- Research Data Services
  data.colorado.edu
  data-help@colorado.edu
  Twitter: @cu_data
  Facebook: CU Boulder Data
- DMP Tool: http://dmptool.org

DMPTool

- With the DMP Tool, you can:
  - Create a new DMP based on funding agency templates
  - Review public DMPs
  - Review requirements for DMPs from different funding agencies
  - Email your institution directly for help (once logged in)
- Let's login and look at a sample template and create a new DMP: http://dmptool.org

Sample DMP

- Let's cover a sample DMP we generated for a hypothetical NSF Division of Atm. and Geospace Sciences proposal

- Funding requirements: https://dmptool.org/guidance?utf8=%E2%9C%93&q=nsf+ags&commit=Search

- Sample plan: https://dmptool.org/plans/10130.pdf

Products of research – What does this mean?

- This section shows you've thought about what sort of data you will generate as part of your project
- How large will my files be?
- What can I expect for growth rates?

Sample DMP

- Section 2: Data formats (and metadata)
- Funding requirements: https://dmptool.org/guidance?utf8=%E2%9C%93&q=nsf+ags&commit=Search
- Sample plan: https://dmptool.org/plans/10130.pdf

Data format and metadata – what does this mean?

Data formats:

- Avoid proprietary formats
- Know what software can be used to read the data

Metadata:

- It’s data about data!
- Describes relevant data for re-creation and re-use

How do I create metadata?

- As simple as a text file! Example: http://www.usap-data.org/entry/NSF-ANT07-39464/2013-01-22_09-39-50/
- Other options: Standardized XML code
- Good metadata should follow community- or discipline-based standards: http://www.dcc.ac.uk/resources/metadata-standards
- Use consistent and documented conventions in the absence of standards

Sample DMP

- Section 3: Data access and sharing
- Funding requirements: https://dmptool.org/guidance?utf8=%E2%9C%93&q=nsf+ags&commit=Search
- Sample plan: https://dmptool.org/plans/10130.pdf

Data access and sharing – what does this mean?

- Funding agencies want to ensure they are getting the most out of their dollar
- Good for science, and good for you!
- Embargo periods are ok, within reason
- Should make data EASILY available
  - Not “by request” only
- Security issues?
  - Must consider privacy and intellectual property issues before making data available

Other ways to increase visibility and access to data

Publishing data sets

- Example: figshare

Publishing peer-reviewed data papers

- Example: http://www.earth-syst-sci-data.net/5/57/2013/essd-5-57-2013.pdf

Sample DMP

- Section 4: Policies for re-use and re-distribution

- Funding requirements: https://dmptool.org/guidance?utf8=%E2%9C%93&q=nsf+ags&commit=Search

- Sample plan: https://dmptool.org/plans/10130.pdf

Policies for re-use and re-distribution – what does this mean?

- Are there any conditions for people to re-use your data?
  - Proper citation is a good condition
- Any disclaimers?
- You must justify properly any limitations you have on who can use your data
- You must also describe how you advertise any restrictions

Sample DMP

- Section 5: Archiving of data
- Funding requirements: https://dmptool.org/guidance?utf8=%E2%9C%93&q=nsf+ags&commit=Search
- Sample plan: https://dmptool.org/plans/10130.pdf

Policies for archiving data – what does this mean?

- What will you do to ensure that the data collected as part of this important project is properly stored and preserved?
- You should have a sound plan in place for storage and preservation
- Store data, metadata, products, anything needed to re-use the data

Good practices for data archiving and preservation

- Trusted repository is best!
- Archiving data in a disciplinary repository
  - Example: Dryad
- Only storing data on thumb drives – bad
- Store multiple copies!
- Active management
- Backups!
- Review schedule for preservation

Data storage: PetaLibrary

- NSF Major Research Instrumentation grant
- Data collections from faculty and students
- Deposition and archiving of data
- Researchers pay for the medium (disk or tape)
- No HIPAA, FERPA, ITAR data
- Infrastructure guaranteed for 5 years

PetaLibrary options

Active: for data that is written or read frequently

- Always stored on spinning disk
- Mounted on CU-RC compute resources (NFS, GPFS)
- Accessible via certain protocols from outside CU-RC
- Option for second copy on tape or on disk in a different building

Archive: for data that is accessed infrequently

- Stored on a combination of disk and tape
- Not mounted on compute resources
- Accessible via certain protocols from outside CU-RC
- Option for additional copy on tape

(Some) data publishing: CU Scholar

- Website: http://scholar.colorado.edu
- Can be used to publish some data sets
- Data sets should be relatively small (<2 GB)
- Must be “publishable” (completed, well-documented)
- Contact Andrew Johnson (andrew.m.johnson@colorado.edu) for assistance with depositing data

Research Data Services

- Website: http://data.colorado.edu
- Email: data-help@colorado.edu
- DMP questions, reviews, assistance with creation
- Research data management questions, consultations, instruction sessions, customized workshops
- Support for DMPTool, PetaLibrary, CU Scholar

Thank you!