Best Practices for Good Data Management

Shelley Knuth
shelley.knuth@colorado.edu

Andrew Johnson
andrew.m.johnson@colorado.edu

www.data.colorado.edu

Link to survey on this topic: http://goo.gl/forms/8VidcwOhRT

Slides: https://github.com/ResearchComputing/Final_Tutorials/

Outline

- What is research data and why do we care about managing it?
- How do I write a good data management plan?
  - Examples
  - DMP Tool
- Resources

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."

- Data itself can really, can be anything!
  - Anything that can be stored on your system

Why do we care about managing research data?

- Good for science:
  - Reproducibility
  - Efficiency
  - Innovation
- Good for you:
  - Let’s keep that data safe!
  - More usage (including citations)
  - More exposure to potential collaborators
  - More competitive grant applications
- Becoming increasingly required
  - Funding agencies, DMPs

Successful DMPs

- Should include, at a minimum:
  - A description of the data, including type(s) and size
  - A plan for preserving the data long term
  - How you will describe the data so that others can reuse it
  - How you will provide as widespread access to the data as possible

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
- Contact your institution directly for help or feedback (once logged in)

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

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

Products of research – What does this mean?

- Section shows you've thought about your data
- How large will my files be?
- What can I expect for growth rates?
  - Manage this dataset with current resources?
- How will I collect my data?
- Existing data?
- What products may be collected or generated?

- Your data?
- 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

Data formats

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

- Proprietary formats:
  - .docx
  - .pptx
  - .xlsx
  - .psd
  - .mov

- Non-proprietary formats:
  - .txt
  - .pdf
  - .csv
  - .tif
  - .mp4

- Know what software can be used to read the data

Metadata

- Data about data!

- Describes relevant data for re-creation and re-use

- Information to include:
  - Contact information about who is in charge of data
  - How the data was collected
  - Important information in collection process
  - Date, location of collection, etc
  - Units
  - Other relevant information

- Your data?
- https://dmptool.org/plans/10130.pdf

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

Data access and sharing – what does this mean?

- Data sharing becoming very important to funding agencies
  - Reproduce existing research
  - Promote further research

- To share data, must properly manage it
  - Proper formats
  - Metadata
  - Stored properly
    - Might be able to combine sharing and storage in one

Data access and sharing – what does this mean?

- Proper ways to share data:

- Data must be made easily available
  - Not "by request" only

- Share with a place that has a digital object identifier (DOI)

- Embargo periods are ok, within reason
  - Data should be published when articles using data are published

- Security issues?
  - Must consider privacy and intellectual property issues before making data available

Where can I share my data?

- Trusted repositories
  - Can store and share data
  - Some charge a fee, some are free
  - Want one with a DOI

- Free example: figshare

- Disciplinary repository
  - http://www.re3data.org/browse/by-subject/

- Generic
  - Dryad

- Personal website?
  - Not great
  - If choose must come up with a schedule for maintenance

- Your data?
- 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

- Your data?
- 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
  - Who? How long? Where? What?

- Store data, metadata, products, anything needed to re-use the data

- Before and after project may be different

Good practices for data archiving and preservation

- Trusted repository is best!
  - Somewhere people make sure it's safe so you don't have to
  - Disciplinary repository
    - http://www.re3data.org/browse/by-subject/
  - Otherwise somewhere more generic
    - Dryad
  - Or somewhere more local
    - University/industry/research group storage facilities
      - At CU: PetaLibrary

Data storage: PetaLibrary

- NSF Major Research Instrumentation grant
- Large data collections from faculty and students
- Deposition and storing of data
- Researchers pay for the medium (disk or tape)
- No HIPAA, FERPA, ITAR data
- Infrastructure guaranteed for 4 years

(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

- Your data?
- https://dmptool.org/plans/10130.pdf

Available Resources

- CU Boulder has many services available 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

Thank you!

- Questions? Email data-help@colorado.edu
- Link to survey on this topic: http://goo.gl/forms/8VidcwOhRT
- Slides: https://github.com/ResearchComputing/Final_Tutorials/blob/master/intro_data_management.pdf