

Introduction to Data Management

Shelley L. Knuth, CU-Boulder Research Computing

https://github.com/ResearchComputing/Final_Tutorials/blob/master/intro_data_management.pdf

Outline

- What is research data and why do we care about managing it?
- How to manage data properly
 - Formats
 - Metadata
 - Storage
 - Sharing
- Competitive in research
- Data management plans
- 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

Data Workflow

- Many projects follow a similar workflow:
 - Data is produced or collected
 - Data needs to be cleaned/corrected to be used
 - Data is analyzed and visualized
 - Won't cover here as not a focus of data management
 - Data is stored
 - Data is shared
- How do we do each properly?

Production/Collection of Data

- Can you explain your dataset?
 - Think before you begin!
- What type/format of data is going to be generated?
 - Images? Text files?
- How large will the files/dataset be?
- What can I expect for growth rates?
 - Can I manage this dataset with my current computing resources?
- What products may be collected or generated?
 - Software
 - Code

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

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
- Very important!!

Data cleaning/correcting

- Very rarely is data that is newly collected perfect
- May need to:
 - Eliminate spurious data
 - Column headers?
 - Reformat dates
 - Convert file formats
 - Change units
- Could do an entire class on data cleaning and correcting
- Document your changes!!!

Identifying and Removing Outliers

- What is an acceptable range for your data?
 - Are you sure?
- Does the data fall within this range?
- Plot it!

- Data not in range? Why not?
 - Human error?
 - Instrument failure?
 - Data is possible but never seen it before?
 - Ozone hole satellite data
 - Must be sure before you delete data!!!!

- Confident in outliers?
 - Set missing data values

Data cleaning/correcting

- Use standard formats
 - netCDF is a good one
 - Standard format
 - Openly available
 - Can be used by many software programs
 - Can describe data within the file (metadata)
 - Don't know netCDF?
 - Text, CSV, etc.
 - Anything that can be used by multiple programs/people and will be around
- ALWAYS keep your original file safe
 - Multiple locations

Policies for backing up data

- What will you do to ensure that the data collected as part of this important project is kept for long term use?
- Keep in mind:
 - How not to lose the data
 - How can others reuse the data
 - Related to sharing...next...
 - How can you maintain the integrity of the data?
- Store data, metadata, products, anything needed to re-use the data
- During project and after project may be different

Data Backup

- Purpose: to make sure data doesn't disappear
 - Storage: Back up data on tape, disk, cloud
 - Archive: Data fixed to one location, keep secure, identifiers
 - Preservation: Specific items maintained over time
 - Ensures continued and reliable use of valued data
 - Curation: Continuous updating, interaction with data creators and users, adds value to dataset
- Most researchers do somewhere between archiving and preservation for final dataset

Good practices for data archiving, preservation, storage

- 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

Good practices for data backup

- Only storing data on thumb drives – bad
- Store multiple copies!
- Active management
- Backups!
- Review schedule for preservation

FAQ about Data Storage

- Q: Where should I store my data?
- A: Somewhere that, at a minimum, can ensure data won't disappear, won't be degraded, and can be accessed easily
- Q: How long do I need to store my data?
- A: Depends on viability of data. Good rule of thumb: 10 years
- Q: Do I need to store every little file I've ever collected?
- A: Depends on how important those files are to the reuse of data or re-creation of research

Data access and sharing

- 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

- 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

Re-use and re-distribution

- Things to consider:
- 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

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

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...

Data Management Plans

- Learn a lot about data management by looking at funding agency requirements
- Sample DMP for 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>

Successful DMPs

- Should include, at a minimum:
 - A brief description of the data, including size of dataset
 - A plan for storing the data long term
 - A description of what will be included in metadata
 - How you will share the data, without limits (if possible)
 - Making sure you can handle your dataset
 - Want to ensure you've thought about your data

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)

Using Data to Promote Research

- Publish or perish – data journals:
 - Geoscience Data Journal
 - Earth System Science Data
 - Scientific Data (Nature)
- Increasing consideration of “products”
 - NSF changed “publications” to “products” on CV
 - Explicitly mentions adding datasets
 - Many universities following suit
- DOIs increase exposure

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!

- Copyright 2014 by Andrew Johnson and Shelley Knuth
- This work is licensed under a [Creative Commons Attribution 3.0 Unported License](#).
- Questions? Email data-help@colorado.edu

