Data Management at CU Boulder

Presenter: Andy Monaghan

Content developed by Aditya Ranganath and Dylan Perkins

https://www.colorado.edu/crdds/

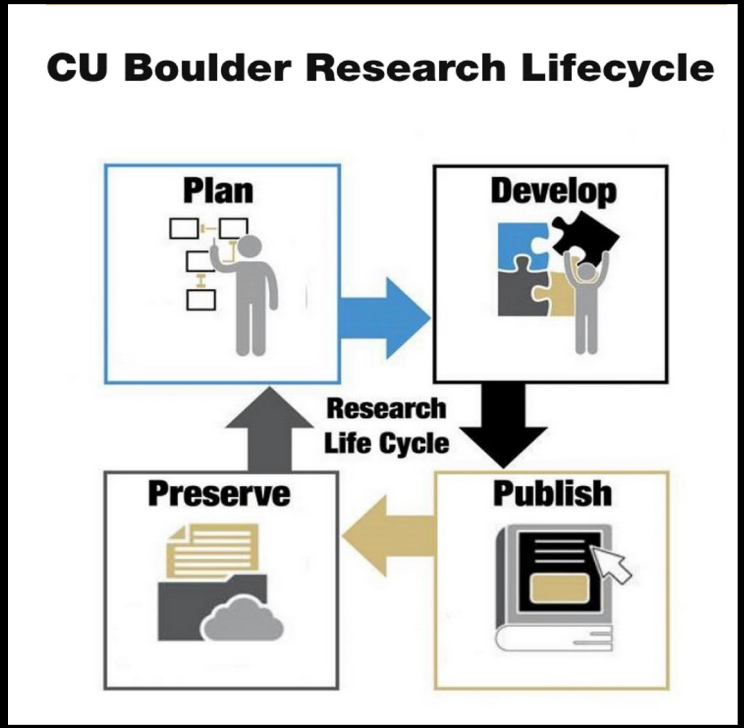
crdds@colorado.edu

Center for Research Data & Digital Scholarship

UNIVERSITY OF COLORADO **BOULDER**

Data Management and the Research Lifecycle

Data management is implicated in every stage of the research lifecycle



Presentation Scope

- Our discussion about data management today will be cast in fairly broad terms, focusing on general principles that apply across projects
- But each project has its own idiosyncrasies, and these are best discussed in the context of a 1:1 consultation
    - If you'd like to discuss your specific data management needs, please contact us! We host drop-in consultation hours on Tuesdays and Thursdays, and you can also make an appointment: crdds@colorado.edu

Presentation Roadmap

- Why is data management important?
- Data management planning
- Data management during a research project
- Data management after a research project
- Data Management, Big Data, and High-Performance Computing

The Value of Research Data Management

- Saves time
- Facilitates continuity and communication in collaborative research settings
- Facilitates reproducibility and the data publication process
- Improves the quality of published data

Data Management Planning

- Putting in place a framework for how you will manage your research data over the lifecycle of the project
- Facilitates addressing possible issues and approach things in a deliberate and intentional way
- Data management plans can be internally facing, as well as externally facing; they are increasingly a requirement of funding agencies

The Elements of a Data Management Plan

- The NSF's Data Management plan guidelines for engineering fields (https://nsf.gov/eng/general/ENG_DMP_Policy.pdf) require discussion of the following:
    - Research products
    - Data formats and standards
    - Dissemination, Access, and Sharing of Data
    - Re-use, Redistribution, and Production of Derivatives
    - Archiving of Data

Support for Data Management Planning at CU-Boulder

- One of CRDDS's roles is to assist researchers with writing DMPs
- CU subscribes to a useful tool, called "dmptool" that has pre-formatted DMP templates from various funding agencies, and which walks you through the process of completing one: https://dmptool.org/
  - When signing in, indicate that you're from CU Boulder
- The data librarians are happy to read drafts and provide feedback on your draft DMPs, so please send them their way! (crdds@colorado.edu)

Data Management During a Project

- Storage
- Documentation
  - Metadata! See https://libraries.mit.edu/data-management/store/documentation/
- File management
  - Pprimer on file naming and organization conventions: https://researchdata.wisc.edu/file-naming-and-versioning/

Data Management Tools for Managing Data and Files During a Project

- The Unix Shell/Command Line
- Git and GitHub
  - https://github.com
- Open Science Framework
  - https://osf.io/
- CRDDS offers workshops on all of these tools (and more!)

Data Management in a Project's Afterlife: Dissemination and Archiving

- Almost all funding agencies require data to be publicly disseminated and stored over a long time horizon

- In a DMP, stating that data will be shared "upon request" or on personal websites is generally inadequate; the norm is that data will be disseminated through a dedicated online repository with a digital object identifier (DOI)

- Will require metadata and documentation to facilitate reuse; easier to generate this information if you've paid attention to data management from the start

Data Management and Big Data

- As the size and complexity of your data increases, storage and data management issues are likely to become increasingly complex

- Such “big” datasets (on the terabyte or petabyte scale) require specialized tools and infrastructure for data management

- CU Boulder’s offers:
  - Microsoft OneDrive (up to 5 TB per CU Boulder person)
  - “PetaLibrary” (unlimited TB, accessed via CU Research Computing)

Gateways for PetaLibrary Access

- OnDemand
  - https://www.ondemand.rc.colorado.edu (Currently CU-Boulder only)
  - https://www.ondemand-rmacc.rc.colorado.edu (Everyone else)
- Globus
  - https://www.globus.org
- Command Line
- URL (for sharing)

PetaLibrary

- Active disk storage
  - $45/TB/year
  - Accessible from all RC compute nodes
- Archive tape storage
  - $20/TB/year
  - iRODS

Questions?

Email: crdds@colorado.edu

Website: www.colorado.edu/crdds