Data Management at CU Boulder

Presenters: Aditya Ranganath and
Dylan Perkins

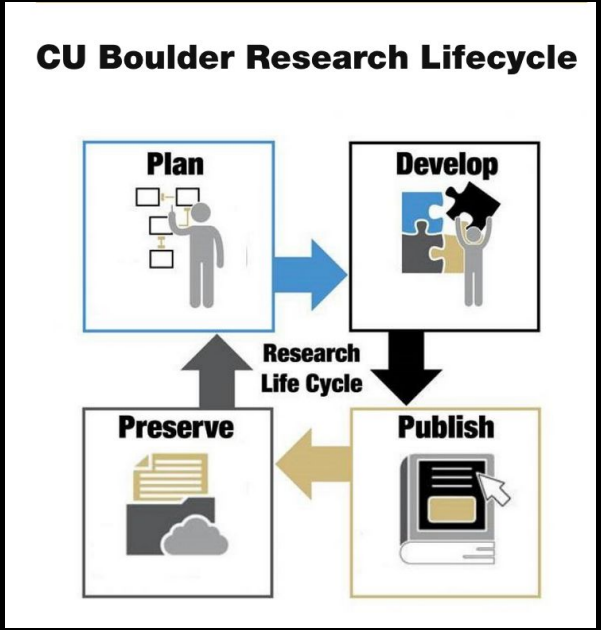
crdds@colorado.edu

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:
  - “Research data management (or RDM) is a term that describes the organization, storage, preservation, and sharing of data collected and used in a research project. It involves the everyday management of research data during the lifetime of a research project (for example, using consistent file naming conventions). It also involves decisions about how data will be preserved and shared after the project is completed (for example, depositing the data in a repository for long-term archiving and access).” (Source: https://pitt.libguides.com/managedata)



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?
  - As researchers, why is this data management something you should invest your time in?
- Data management planning
  - Increasingly, funding agencies require researchers to put together data management plans (DMPs) as part of their grant applications. What exactly are DMP's and what should you consider when writing them?
- Data management during a research project
  - What are some tools and strategies you can use to keep all of your data organized while you are collecting and analyzing it?
- Data management after a research project
  - The norms of open science, as well as funding agency and journal policies, are increasingly requiring research data to be publicly disseminated after research is published. How can CRDDS assist you with the process of making your data publicly available, so that it can be reused and cited by other researchers?
- Data Management, Big Data, and High-Performance Computing

The Value of Research Data Management

- Saves time
- Facilitates continuity and communication in collaborative research settings
  - If people cycle through a lab or research group over time, data management ensures that new members can quickly get up to speed and pick up where departing members left off
  - In the context of team member turnover, attention to data management minimizes the risk of data becoming misused or becoming unusable because of missing context or metadata
- Facilitates reproducibility and the data publication process
- Improves the quality of published data
  - This benefits you as the producer of data, since it increases the likelihood that your data will be useful to others, which leads to citations and influence
  - It benefits you as a consumer of data, since using well-managed and well-documented data from other labs to saves you time

Data Management Planning

- Data management planning is about putting in place a framework for how you will manage your research data over the lifecycle of the project
- Putting things in writing helps you to think through various issues that might arise, and approach things in a deliberate and intentional way, rather than figuring things out as you go.
- Data management plans can be internally facing, as well as externally facing; they are increasingly a requirement of funding agencies, which ask you to submit formal data management plans as a part of funding applications.
- When starting a project consider drafting two plans:
    - One for funding agencies
    - Another for internal use, which can be based on the externally facing plan, but which goes into greater detail on operational issues, and thereby facilitates project continuity

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: What data products will result from the research being proposed? What metadata will be collected and distributed to facilitate the data's legibility to others?
  - Data formats and standards: What are the file formats in which the data will be stored and distributed, and what issues arise with respect to issues such as accessibility and interoperability?
  - Dissemination, Access, and Sharing of Data: How will the data be disseminated to the broader research community after the research is complete (i.e. disciplinary repository? Institutional repository?). Are there privacy-related, or legal or ethical constraints on the ability to share data?
  - Re-use, Redistribution, and Production of Derivatives: What can other parties do with the research data that you will share? Are there any constraints on how the data you produce can be used by these external parties?
  - Archiving of Data: What data will be archived, and how will it be preserved over time?

The Elements of a Data Management Plan, continued

- Beyond the the NSF's guidelines/required sections, the UK's Digital Curation Center (DCC) has a more comprehensive list of issues to consider when data management planning that you might consider when writing an internally facing plan: https://www.dcc.ac.uk/sites/default/files/documents/resource/DMP_Checklist_2013.pdf

Support for Data Management Planning at CU-Boulder

- One of CRDDS's roles is to assist researchers with writing DMPs, particularly DMPs that are required as part of grant and funding proposals
- We subscribe 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 our way!

DMPTool

- Brief DMPTool tour
- Brief DMPTool Activity
  - Select a DMP Template relevant for your field (if you're having trouble choosing, pick the generic NSF template)
  - Look through the prompts
  - Think of a project you plan to initiate in the near future that will require the use of data. Attempt to answer the prompts in the context of your proposed project, and use the tool to generate your draft DMP.
  - Were any of the prompts challenging or confusing? Is there anything you would like clarification on?

Data Management During a Project

- Storage
  - Where will data be stored? One Drive, PetaLibrary etc.
  - How will data be backed up?
  - What security measures are in place to protect the data?
- Documentation
  - Don’t wait until the end of a project to generate metadata! Documenting your data workflows as you go will make life much easier. For guidance on best practices for data documentation and metadata, this document (from MIT) is a good place to start: https://libraries.mit.edu/data-management/store/documentation/
- File management
  - Data files can quickly proliferate, and it’s important to have a framework in place that regulates how files are named, where they are stored, and how they are versioned. For a useful primer on file naming and organization conventions, see this resource: https://researchdata.wisc.edu/file-naming-and-versioning/

Data Management Tools for Managing Data and Files During a Project

- The Unix Shell/Command Line:
  - Create, name, delete, and move around files and directories,programmatically (For an excellent tutorial on using the Unix shell for file management, see here: https://swcarpentry.github.io/shell-novice/.
- Git and GitHub
  - Among other things, helps to track changes made to files containing data and code over time.
- Open Science Framework
  - Designed to be a one-stop shop for all of your data and project management needs.
  - It integrates various data storage, analysis, versioning, and management applications into one unified platform
  - https://osf.io/
- CRDDS offers workshops on all of these tools (and more!)

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

- Funding agencies increasingly require data to be publicly disseminated and stored over a long time horizon
- The DMPs required by these agencies ask researchers about their plans in these areas
- Saying that data will be shared "upon request" or on personal websites is increasingly seen as inadequate; the norm taking hold is that data will be disseminated through a dedicated online repository
- These repositories can be discipline-specific, or institutional repositories (generally, either is seen as acceptable)
- Will require metadata and documentation to facilitate reuse; easier to generate this information if you've paid attention to data management from the start
- More on this in a presentation tomorrow

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
- In the remainder of the presentation, we will survey some of these tools and resources

What is Research Computing?

- Provide services for researchers that include:
  - Large scale computing
  - Data storage
  - High speed data transfer
  - Data management support
  - Consulting
  - Training

- We are likely best known for:
  - Alpine Supercomputer
  - PetaLibrary storage

Alpine

- 300+ compute nodes
  - Majority 64 cores / node
- 18,000+ cpu cores
- High-speed interconnects – can run parallel jobs across lots of nodes
- Super fast scratch file system

What Would I Use Alpine For?

- Research Computing is more than just Alpine
- What would you use Alpine For?
  - Solving large problems that require more:
    - Memory than you have on your personal computer
    - Cores/nodes/power than you have on your personal computer
  - High performance GPU computing
  - High memory jobs
  - Visualization rendering
- Not a place for:
  - Large data storage

Data Management on HPC

- Filesystems
  - /home
    - Small 2GB in size - Best for user configuration files (bash, ssh). Not to be shared with anyone
  - /projects
    - Larger 250GB in size - Best for sharing with others including data, code etc.
  - /scratch
    - Largest 10TB in size - Best for fast read/write during analysis. 90 purge policy, not a permanent place for data

PetaLibrary

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

Gateways

Open OnDemand

- Browser-based
- Supports user-provided custom kernels
- Access to All CURC systems
  - https://www.ondemand.rc.colorado.edu (Currently CU-Boulder only)
  - https://www.ondemand-rmacc.rc.colorado.edu (Everyone else)

…and More

- “Condo” computing available (groups buy dedicated nodes) - Blanca
- New on-premise cloud system CUmulus
- Public Cloud Support
- User support team can help you get started on CURC resources

Questions?

Email: crdds@colorado.edu

Website: www.colorado.edu/crdds