RESEARCH DATA MANAGEMENT FUNDAMENTALS

Bill Corey
Research Data Management Librarian
Research Data Services & Sciences
University of Virginia Library
March 19, 2019

http://library.soton.ac.uk/researchdata

This workshop provides an overview of data management topics and practices. The emphasis is on strategies researchers can implement to make their data more findable, accessible, interoperable, and reusable — for themselves or others.

- **file organization** and **formats**
- creating **documentation** and **metadata**
- **data security** and **backups**
- **data sharing** and **publishing**

responsible data reuse

- **citation**
- **credit**
- **Copyright**

Why should you be concerned about making your data more findable, accessible, interoperable, and reusable? Because it:

- Increases the impact and visibility of research
- Promotes innovation and potential new data uses
- Leads to new collaborations between data users and creators
- Maximizes transparency and accountability
- Enables scrutiny of research findings
- Encourages improvement and validation of research methods
- Reduces cost of duplicating data collection
- Provides important resources for education and training

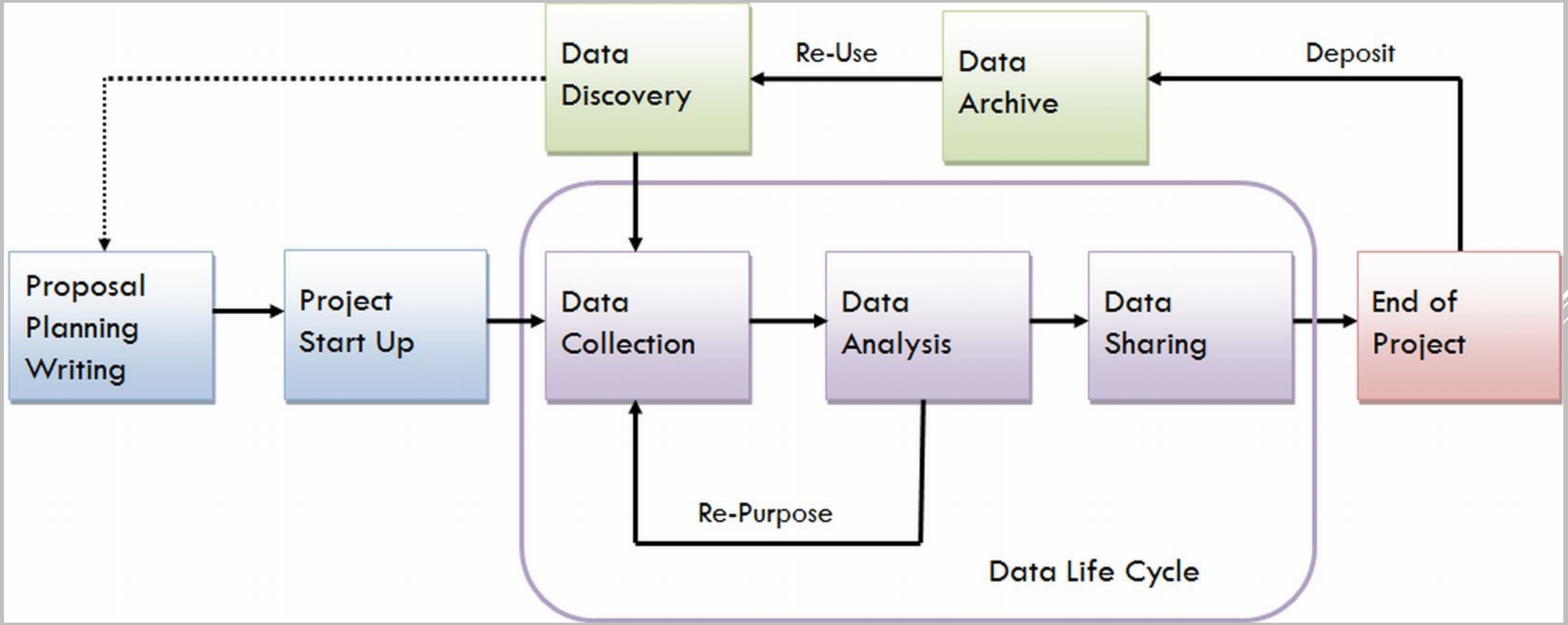
https://www.ukdataservice.ac.uk/manage-data/plan/why-share.aspx

Data Sharing and Management Snafu in 3 Short Acts

NYU Health Sciences Library

What is the Data Life Cycle?

The life cycle illustrates steps through which well managed data moves from creation to conclusion in a research project.



If your data are:

- well-organized
- documented
- preserved
- accessible
- verified as to accuracy and validity

Then the result will be:

- high-quality data
- easy to share and re-use in science
- citation and credibility to the researcher
- cost-saving to science

courtesy DataONE

Steps in the Data Life Cycle

Proposal Planning & Writing:

- Review of existing data sources, determine if project will produce new data or combine existing data
- Investigate archiving challenges, costs, consent and confidentiality
- Identify potential users of your data
- Contact Archives for advice

Project Start Up:

- Create a data management plan
- Make decisions about documentation form and content
- Conduct pretest of collection materials and methods

Steps in the Data Life Cycle

Data Collection:

- Organize files, backups & storage, QA for data collection
- Think about access control and security

Data Analysis:

- Document analysis and file manipulations
- Manage file versions

Steps in the Data Life Cycle

Data Sharing:

- Determine file formats
- Verify institutional and funder requirements or restrictions
- Contact Archive for advice
- Further document and clean data

End of Project:

- Deposit data in data archive (repository)

File Organization

Best practices:

- File Version control
- Directory structure
- File naming conventions (including discipline-specific)
- File structure
- Use same structure for Backups

File Naming

Best practices:

- Descriptive names
- Unique identifier or project name/acronym
- Primary investigator (PI)
- Location and/or spatial coordinates
- Year of study
- Data type
- version number
- File type

File Formats

Best practices:

- Non-proprietary
- Unencrypted
- Uncompressed
- Open, documented standard
- Commonly used by your research community
- Use common character encodings – ASCII, Unicode, UTF-8

Research Data Management Subject Guide

Documentation and Metadata

Why you should document your data:

- Enables efficient organization of the research data
- Facilitates discovery
- Facilitates research data sharing
- Identifies the creator(s)of the data
- Provides permanent identifiers for the data
- Links the data to other related products – articles and other datasets
- Supports archiving and preservation

Research Data Management Subject Guide

Documentation and Metadata

Research Project Documentation:

- Context of data collection
- Data collection methods
- Structure and organization of data files
- Data sources used
- Data validation and quality assurance
- Transformation of data from the raw data through analysis
- Information on confidentiality, access and use conditions

Research Data Management Subject Guide

Documentation and Metadata

Dataset Documentation:

- Variable names and descriptions
- Explanation of codes
- Explanation of classification schemes used
- Algorithms used to transform data
- File format
- Software used in collection – version, OS
- Software used in analysis – version, OS

Research Data Management Subject Guide

Documentation and Metadata

Types of documentation:

- Data dictionaries
- Permanent identifiers - DOI
- Code books
- File directories
- Methodologies
- Glossary
- ReadMe files
- Data definition files

[Research Data Management Subject Guide]

Metadata:

- Schema
- Standards (general) – Dublin Core
- Standards (discipline-specific)

Data Security

Best Practices:

- **Network Security**: Keep confidential data off of the internet. Put highly sensitive materials on computers not connected to the internet.
- **Physical Security**: Restrict access to buildings and rooms where computers or media are kept. Only let trusted individuals troubleshoot computer problems.
- **Computer Systems and Files**: Keep virus protection up top date. Don't send confidential data via e-mail or FTP. Use Encryption if you must. Use strong passwords on files and computers.

Research Data Management Subject Guide

Backups

Best Practices:

Accidents DO happen: hardware failures, media deteriorates, drives are lost, computers are stolen, data files are corrupted by viruses, power failures damage drives, and human errors are not uncommon.

- 3-2-1 Rule: Keep 3 copies of your files in 2 different locations, with 1 copy off-site, ideally in a different geographic zone.
- Backup often. Schedule backups frequently, and follow the schedule.
- Use a reliable medium. Test your backups periodically by testing files restores. Check the integrity of the data using checksum validation.

Research Data Management Subject Guide

Data Sharing

Why you should share your research data:

- Enabling others to replicate and verify results as part of the scientific process
- Allows researchers to ask new questions and conduct new analysis
- Linking to research products like publications and presentations
- Creating a more complete understanding of a research study
- Meeting sponsor, funder, publisher, and institution expectations
- Receiving credit for data creation for career advancement
- Reduces the costs of duplicating data collection

Research Data Management Subject Guide

Data Sharing

How you should share your research data:

- Deposit it a discipline-specific repository, general repository, or archive
- Deposit in UVa's Data Repository – LibraData (your final, publishable products of research)
- Disseminate through a project, personal, or department website
- Submit as supplemental material to a journal in support of an article
- Peer-to-peer exchange

Research Data Management Subject Guide

Data Sharing

Advantages of using a data repository:

- Persistent identifiers – unique and citable
- Access controls
- Terms of use and licenses
- Repository guidelines for deposit
- Data preservation – migrating to new formats or emulating old formats
- Professional backup and documentation
- Repository Standards ensure commitment and quality

Research Data Management Subject Guide

Data Sharing Repository Search

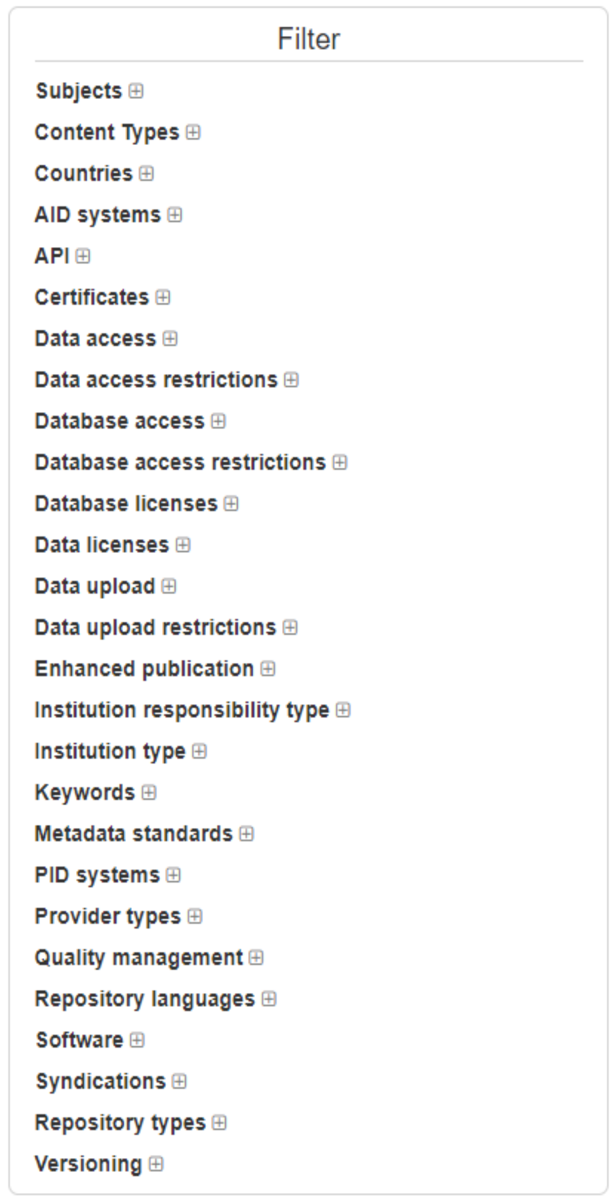
2297 repositories

1038 in US

Browse by

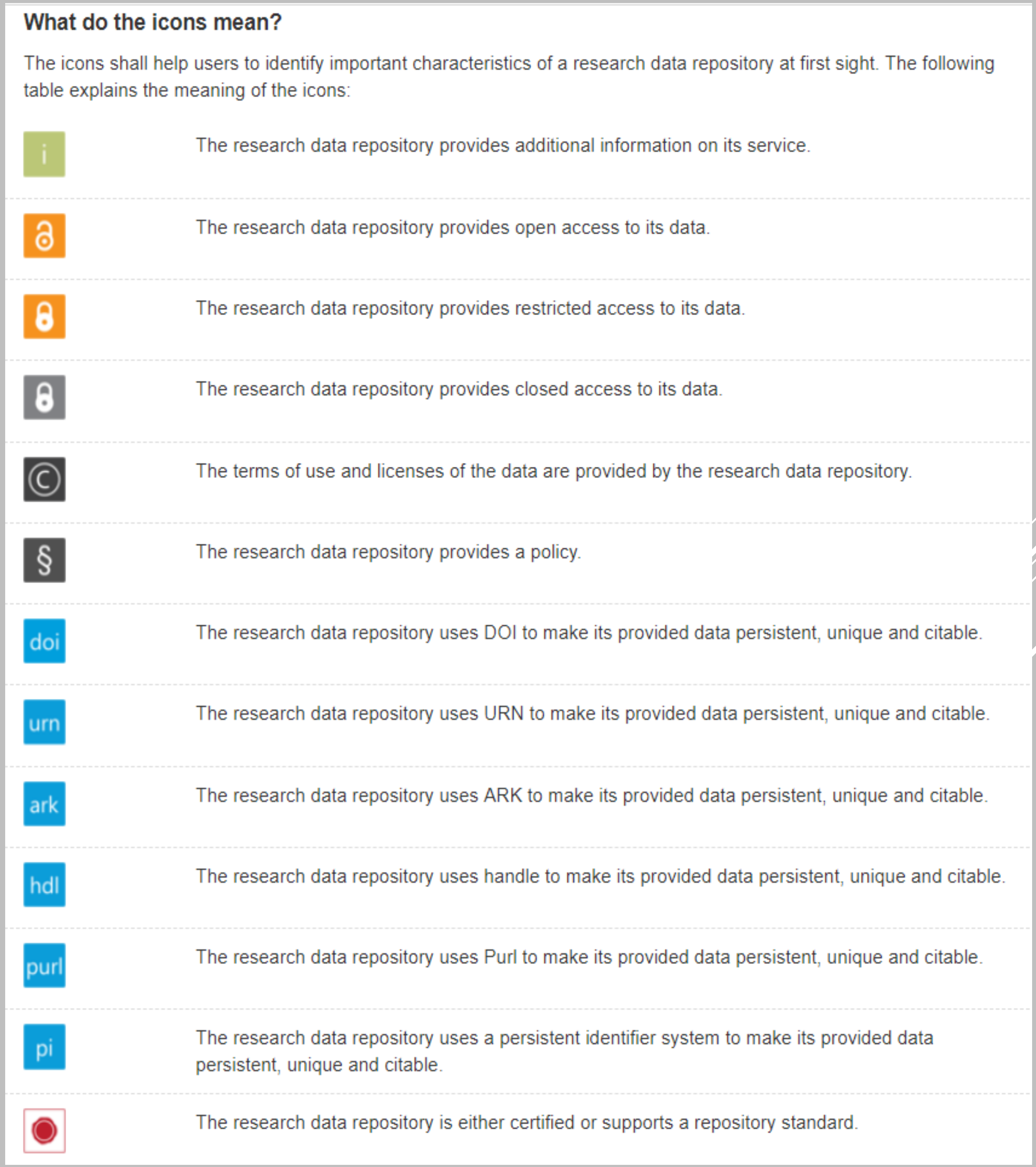
- Subject
- Content type
- country

re3data.org



Filter

- Subjects ⊞
- Content Types ⊞
- Countries ⊞
- AID systems ⊞
- API ⊞
- Certificates ⊞
- Data access ⊞
- Data access restrictions ⊞
- Database access ⊞
- Database access restrictions ⊞
- Database licenses ⊞
- Data licenses ⊞
- Data upload ⊞
- Data upload restrictions ⊞
- Enhanced publication ⊞
- Institution responsibility type ⊞
- Institution type ⊞
- Keywords ⊞
- Metadata standards ⊞
- PID systems ⊞
- Provider types ⊞
- Quality management ⊞
- Repository languages ⊞
- Software ⊞
- Syndications ⊞
- Repository types ⊞
- Versioning ⊞



What do the icons mean?

The icons shall help users to identify important characteristics of a research data repository at first sight. The following table explains the meaning of the icons:

i	The research data repository provides additional information on its service.
🔓	The research data repository provides open access to its data.
🔒	The research data repository provides restricted access to its data.
🔒	The research data repository provides closed access to its data.
©	The terms of use and licenses of the data are provided by the research data repository.
§	The research data repository provides a policy.
doi	The research data repository uses DOI to make its provided data persistent, unique and citable.
urn	The research data repository uses URN to make its provided data persistent, unique and citable.
ark	The research data repository uses ARK to make its provided data persistent, unique and citable.
hdl	The research data repository uses handle to make its provided data persistent, unique and citable.
purl	The research data repository uses Purl to make its provided data persistent, unique and citable.
pi	The research data repository uses a persistent identifier system to make its provided data persistent, unique and citable.
●	The research data repository is either certified or supports a repository standard.

Data Sharing

Things to consider in preparing your data for sharing and archiving:

- File formats for long-term access: non-proprietary or open formats
- Documentation: document your research and data so others can interpret the data.
- UVa Data Retention Policy: University faculty and researchers have a responsibility to maintain research data and make the data available for preservation by the University both as a matter of research integrity, and because of the University's ownership rights.
- Ownership and Privacy: Carefully consider the implications of sharing your data, in terms of copyright and IP ownership, and ethical requirements like privacy and confidentiality.

Research Data Management Subject Guide

Data Publishing

Advantages to Publishing Research Data:

- Increased exposure of a dataset
- Validation – strengthens the credibility of the study relying on the data
- Element of peer-review of the dataset
- Academic accreditation for the researcher
- Sharing of datasets not tied to publications
- Increased citation counts for related articles
- Faster pace of science progress – maximize opportunities for reuse

Responsible Data Reuse

Copyright and Intellectual Property Rights

Strategies to consider in preparing your data for sharing and archiving:

- Data is not copyrightable. A particular expression of data, such as a chart or a table in a book, can be.
- Data can be licensed. Some data providers apply licenses that limit how the data can be used.
- Data can be considered to be IP if it is used to create a patentable object or process that has commercial application.

Responsible Data Reuse

Privacy and Confidentiality

Strategies for using shared sensitive and confidential data:

- Gaining informed consent that includes consent for data sharing (via deposit in a repository or archive).
- Protecting privacy through anonymizing data
- Considering controlling access to the data (via embargoes or access/licensing terms and conditions).

Responsible Data Reuse

Data Citation

Primary Elements to include in all data citations:

- Creator: Author(s) of the dataset
- Title: Name of the dataset
- Publisher (or Distributor): Repository name
- Publication Year: Date the dataset was released or published
- Version: If you have multiple versions of a specific dataset.
- Persistent Identifier: Unique Identifier. This is often a DOI, but can also be an URN or Handle System.

Research Data Management Subject Guide

Responsible Data Reuse

Data Citation

Example citations:

- Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. Geological Institute, University of Tokyo.http://dx.doi.org/10.1594/PANGAEA.726855
- Sidlauskas B (2007) Data from: Testing for unequal rates of morphological diversification in the absence of a detailed phylogeny: a case study From characiform fishes. Dryad Digital Repository. doi:10.5061/dryad.20
- Barnes, Samuel H. Italian Mass Election Survey, 1968. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1992-02-16. https://doi.org/10.3886/ICPSR07953.v1

Research Data Management Subject Guide

Thanks for attending!

If you have any questions or concerns, please contact me.

Bill Corey

Research Data Management Librarian

wtc2h@virginia.edu

434-243-5882

Research Data Management Subject Guide

Research Data Services and Sciences

Research Data Management