RESEARCH DATA MANAGEMENT BEST PRACTICES

Bill Corey
Research Data Management Librarian
Research Data Services & Sciences
University of Virginia Library
September 25, 2019

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

This workshop provides an overview of research data management best 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 storage**, **security** and **backups**
- **data sharing**

responsible data reuse

- **citation**
- **credit**
- **copyright**

Data Deluge!

https://www.meetingsnet.com/sites/meetingsnet.com/files/MNDec_funnel_guy.jpg

What is Research Data?

"The recorded factual material commonly accepted in the scientific community as necessary to validate research findings."

https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/circulars/A110/2cfr215-0.pdf

"Research data is any information that has been collected, observed, generated or created to validate original research findings."

https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained

Data can be digital or analog.

There are 5 categories of data:

- **Observational**: Captured in real-time, can't be reproduced or recaptured – "unique data".
- **Experimental**: Captured from lab equipment, often under controlled conditions. Usually reproducible but can be expensive.
- **Simulation**: generated from test models studying actual or theoretical systems.
- **Derived** or **Compiled**: Results of data analysis or aggregated from multiple sources.
- **Reference** or **Canonical**: Fixed or organic collection datasets, usually peer-reviewed, published, and curated.

Exercise:

- What kind of data do you work with?
- What organizational problems have you faced?
- What tools and techniques work for you?

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

- 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

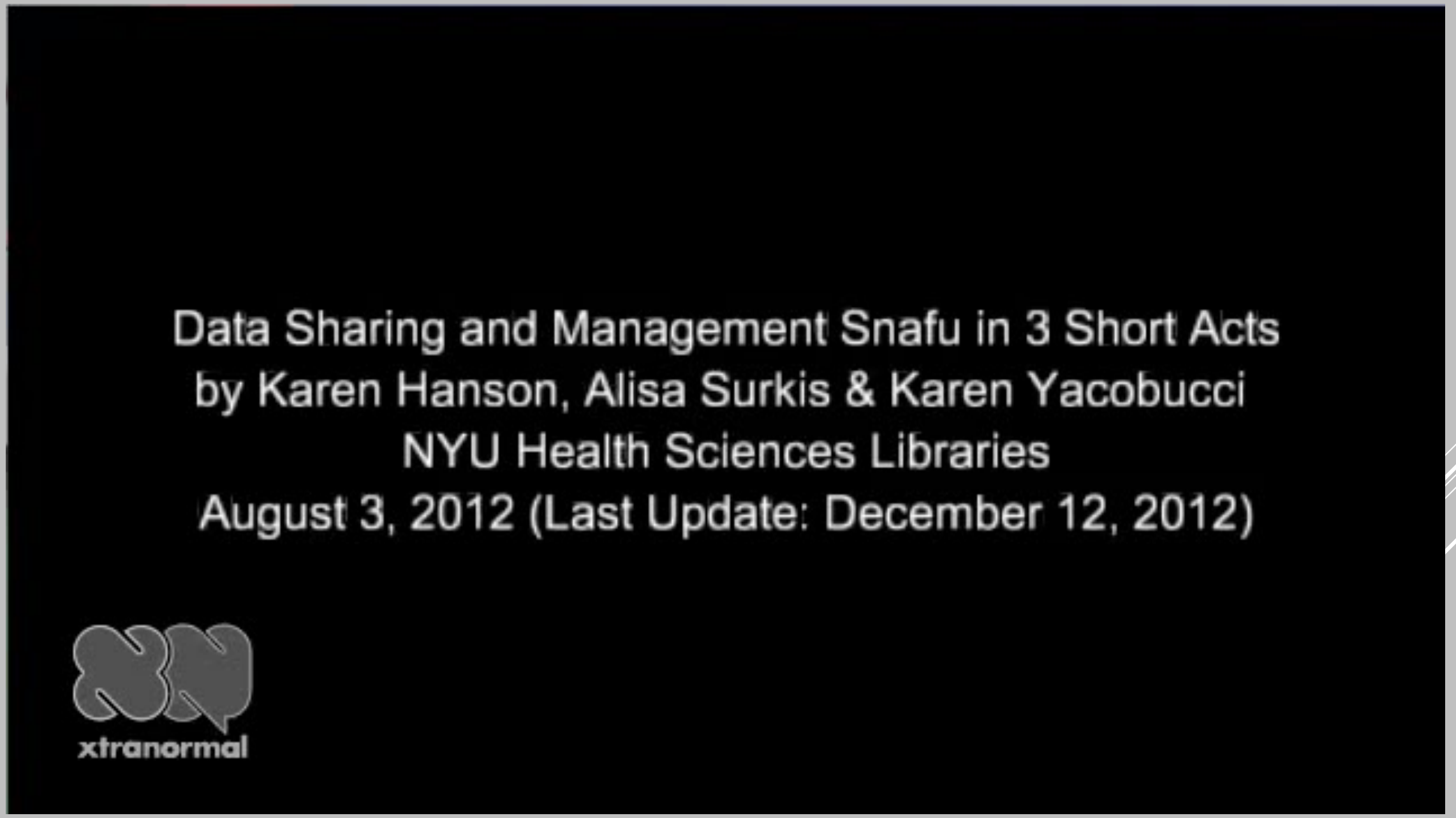
https://www.ukdataservice.ac.uk/manage-data/plan/why-share.aspx

Why Manage Data: Researcher Benefits

- Keep yourself organized – be able to find your files (data inputs, analytic scripts, outputs at various stages of the analytic process, etc.)
- Track your science processes for reproducibility – be able to match up your outputs with exact inputs and transformations that produced them
- Better control versions of data – easily identify versions that can be periodically purged
- Quality control your data more efficiently
- To avoid data loss (e.g. making backups)
- Format your data for re-use (by yourself or others)
- Be prepared: Document your data for your own recollection, accountability, and re-use (by yourself or others)
- Gain credibility and recognition for your science efforts through data sharing!

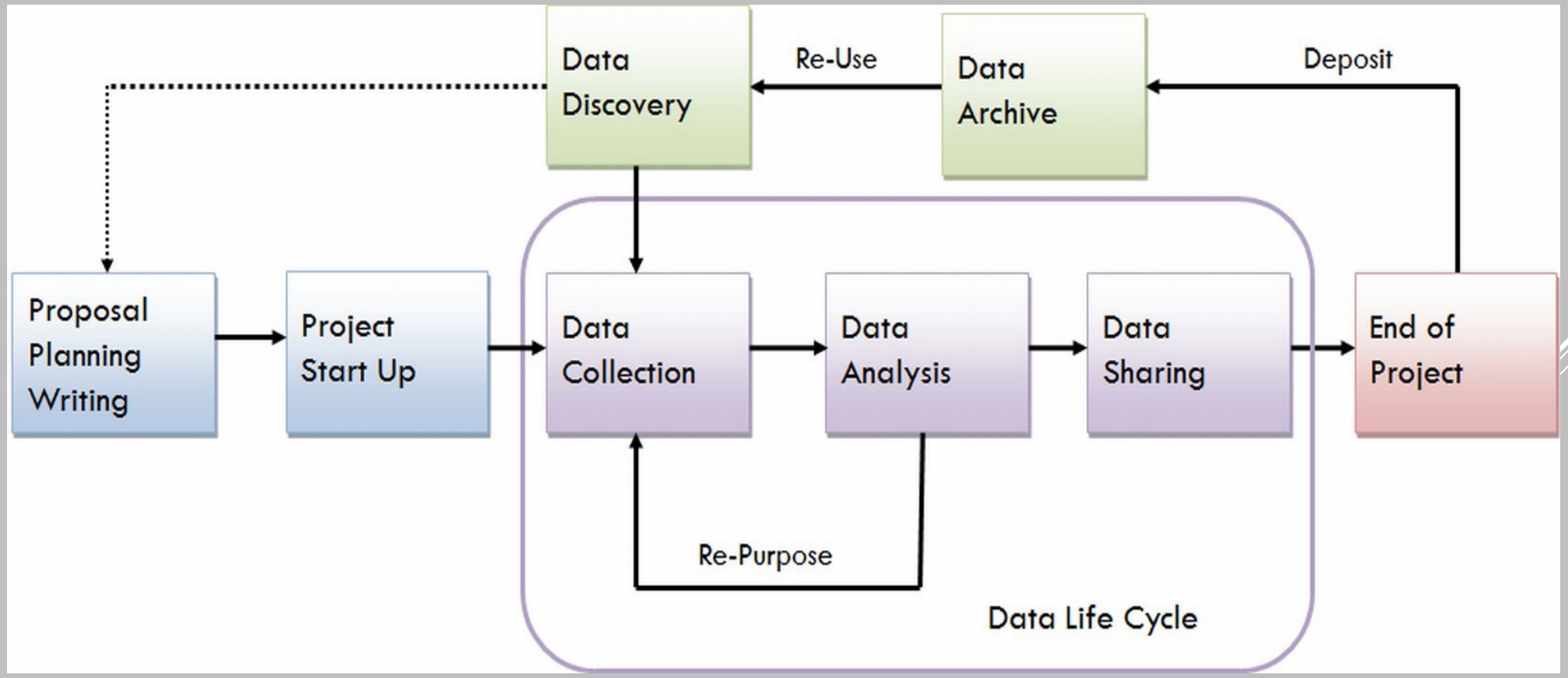
Data Loss

- Natural disasters
- Facilities infrastructure failures
- Storage failures
- Server hardware or software failures
- Human errors
- Malicious attacks
- Format obsolescence
- Loss of funding
- Loss of institutional commitment
- Loss of competencies


Data Sharing and Management Snafu in 3 Short Acts
by Karen Hanson, Alisa Surkis & Karen Yacobucci
NYU Health Sciences Libraries
August 3, 2012 (Last Update: December 12, 2012)
xtranormal

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.



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

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)

Data Management Plan

A DMP should describe how you will collect, organize, analyze, preserve, and share your data.

- **Identify your data**: type of data, software used to collect or analyze it, how you will collect it, quantity of data and file size, file types.
- **Organize your data**: file naming, organization, version control, QA.
- **Document your data**: metadata, data dictionaries, codebooks, ReadMe files, data paper.
- **Data Storage, Security, Backup**: storage methods and locations, backup schedule, privacy, ethics and legal concerns.
- **Data Preservation and Sharing**: archive or repository, formats.
- **Roles and Responsibilities**: who is responsible for managing the data now and later.

File Organization

Best practices:

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

File Naming

Why file naming is important:

- You think you'll remember but over time…
- Multiple formats and different versions
- Easier to share if everyone understands
- Time saving – set it up right at the beginning makes it easier to locate later

The 5 C's: Be Clear, Concise, Consistent, Correct, and Conformant.

There is no one right way to do it – find a balance you are comfortable with. Create a ReadMe that explains your naming conventions so you and others will know your methodology.

File Naming

Be Consistent! Remember the 5 C's:

Be Clear, Concise, Consistent, Correct, and Conformant.

Best practices:

- Descriptive names
- Unique identifier or project name/acronym
- Primary investigator (PI) or researcher name or initials
- Location and/or spatial coordinates
- Year of study, date, or date range - YYYYMMDD
- Data type
- Version number
- Sequential numbering – add leading zeros to allow for additional files

File Naming

Worst practices (things to avoid):

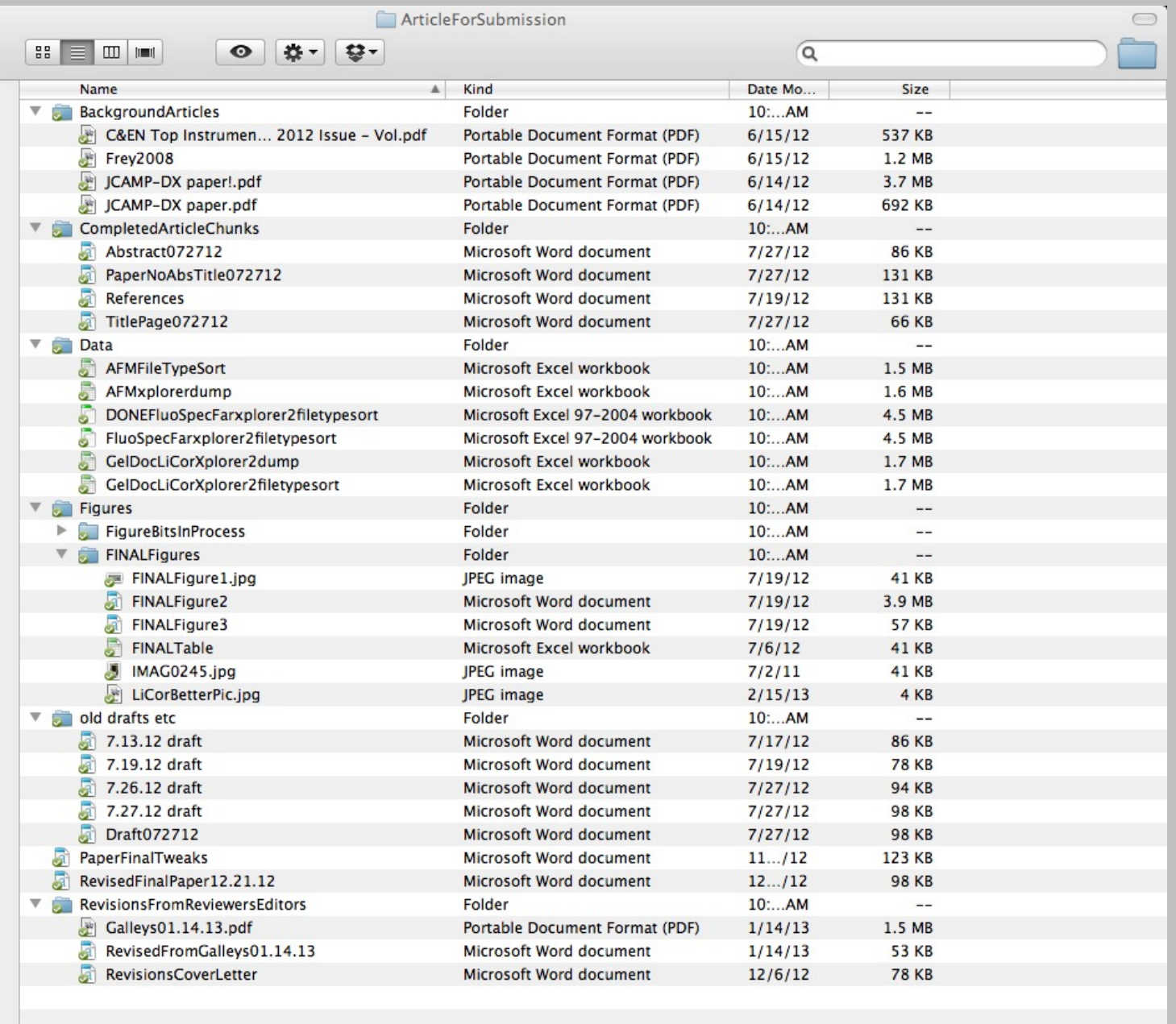
- Cryptic codes only you understand
- Using more than 32 characters
- Special characters – **do not use** & , * % # ; * ( ) ! @$ ^ ~ ' { } [ ] ? < > -
- Spaces – use dashes, underscores, Camel case instead
- Common terms – data, sample, final, document, resume
- Multiple dots or periods – only one before the file extension
- Inconsistent case

Directory Structure

Best practices:

- Mimic the way you work and keep it simple
- Think of folder names as keywords
- Make a template (so you don't start over for each project)
- Keep a flow chart (cheat sheet) or use a mind map
- Don't overlap folders or categories
- Keep the folders manageable – not too big
- Document your system – define data types, file formats, naming convention. Use the same rules as file naming.

Exercise: File Naming and Directory Structure



Name	Kind	Date Mo...	Size
BackgroundArticles	Folder	10:...AM	--
C&EN Top Instrumen... 2012 Issue - Vol.pdf	Portable Document Format (PDF)	6/15/12	537 KB
Frey2008	Portable Document Format (PDF)	6/15/12	1.2 MB
JCAMP-DX paper!.pdf	Portable Document Format (PDF)	6/14/12	3.7 MB
JCAMP-DX paper.pdf	Portable Document Format (PDF)	6/14/12	692 KB
CompletedArticleChunks	Folder	10:...AM	--
Abstract072712	Microsoft Word document	7/27/12	86 KB
PaperNoAbsTitle072712	Microsoft Word document	7/27/12	131 KB
References	Microsoft Word document	7/19/12	131 KB
TitlePage072712	Microsoft Word document	7/27/12	66 KB
Data	Folder	10:...AM	--
AFMFileTypeSort	Microsoft Excel workbook	10:...AM	1.5 MB
AFMxplorerdump	Microsoft Excel workbook	10:...AM	1.6 MB
DONEFluoSpecFarxplorer2filetypesort	Microsoft Excel 97-2004 workbook	10:...AM	4.5 MB
FluoSpecFarxplorer2filetypesort	Microsoft Excel 97-2004 workbook	10:...AM	4.5 MB
GelDocLiCorXplorer2dump	Microsoft Excel workbook	10:...AM	1.7 MB
GelDocLiCorXplorer2filetypesort	Microsoft Excel workbook	10:...AM	1.7 MB
Figures	Folder	10:...AM	--
FigureBitsInProcess	Folder	10:...AM	--
FINALFigures	Folder	10:...AM	--
FINALFigure1.jpg	JPEG image	7/19/12	41 KB
FINALFigure2	Microsoft Word document	7/19/12	3.9 MB
FINALFigure3	Microsoft Word document	7/19/12	57 KB
FINALTable	Microsoft Excel workbook	7/6/12	41 KB
IMAG0245.jpg	JPEG image	7/2/11	41 KB
LiCorBetterPic.jpg	JPEG image	2/15/13	4 KB
old drafts etc	Folder	10:...AM	--
7.13.12 draft	Microsoft Word document	7/17/12	86 KB
7.19.12 draft	Microsoft Word document	7/19/12	78 KB
7.26.12 draft	Microsoft Word document	7/27/12	94 KB
7.27.12 draft	Microsoft Word document	7/27/12	98 KB
Draft072712	Microsoft Word document	7/27/12	98 KB
PaperFinalTweaks	Microsoft Word document	11.../12	123 KB
RevisedFinalPaper12.21.12	Microsoft Word document	12.../12	98 KB
RevisionsFromReviewersEditors	Folder	10:...AM	--
Galleys01.14.13.pdf	Portable Document Format (PDF)	1/14/13	1.5 MB
RevisedFromGalleys01.14.13	Microsoft Word document	1/14/13	53 KB
RevisionsCoverLetter	Microsoft Word document	12/6/12	78 KB

Would you organize these files differently?

What do you think about the naming conventions used in this directory? Would you change anything?

Courtesy of the New England Collaborative Data Management Curriculum (NECDMC)
https://library.umassmed.edu/resources/necdmc/modules

Version Control

Best practices:

- Use a sequential numbered system for major changes with ordinal numbers – e.g. v01, v02, v03… Add decimals for minor changes - e.g. v01.1, v01.2, v01.3…
- Use precise labels
- Place older files in a separate folder (archive)
- Use dates to distinguish versions – e.g. 09222019, 09232019, 09242019
- Use version control software – Git, GNU RCS, Mercurial SCM, Tortoise SVN
- Keep the original version of the data file the same and create a copy to start the iterative version process

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

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 fails, 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

Finding data to reuse:

- Search discipline-specific repositories
- Search Community repositories
- Search NIH-approved repositories
- Search DataCite for datasets (by DOI)
- Search DataCite for researchers (by their Orcid ID)

Data Sharing Repository Search: re3data

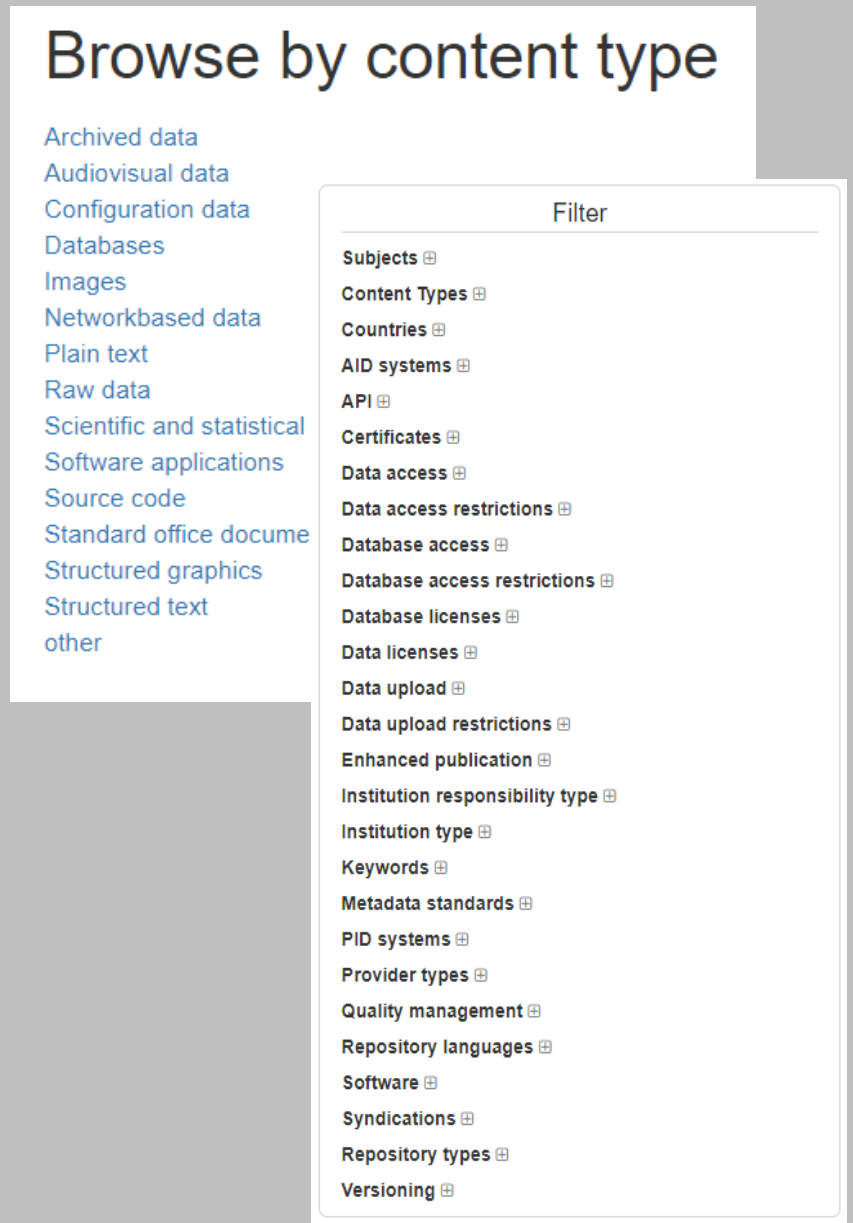
2399 repositories

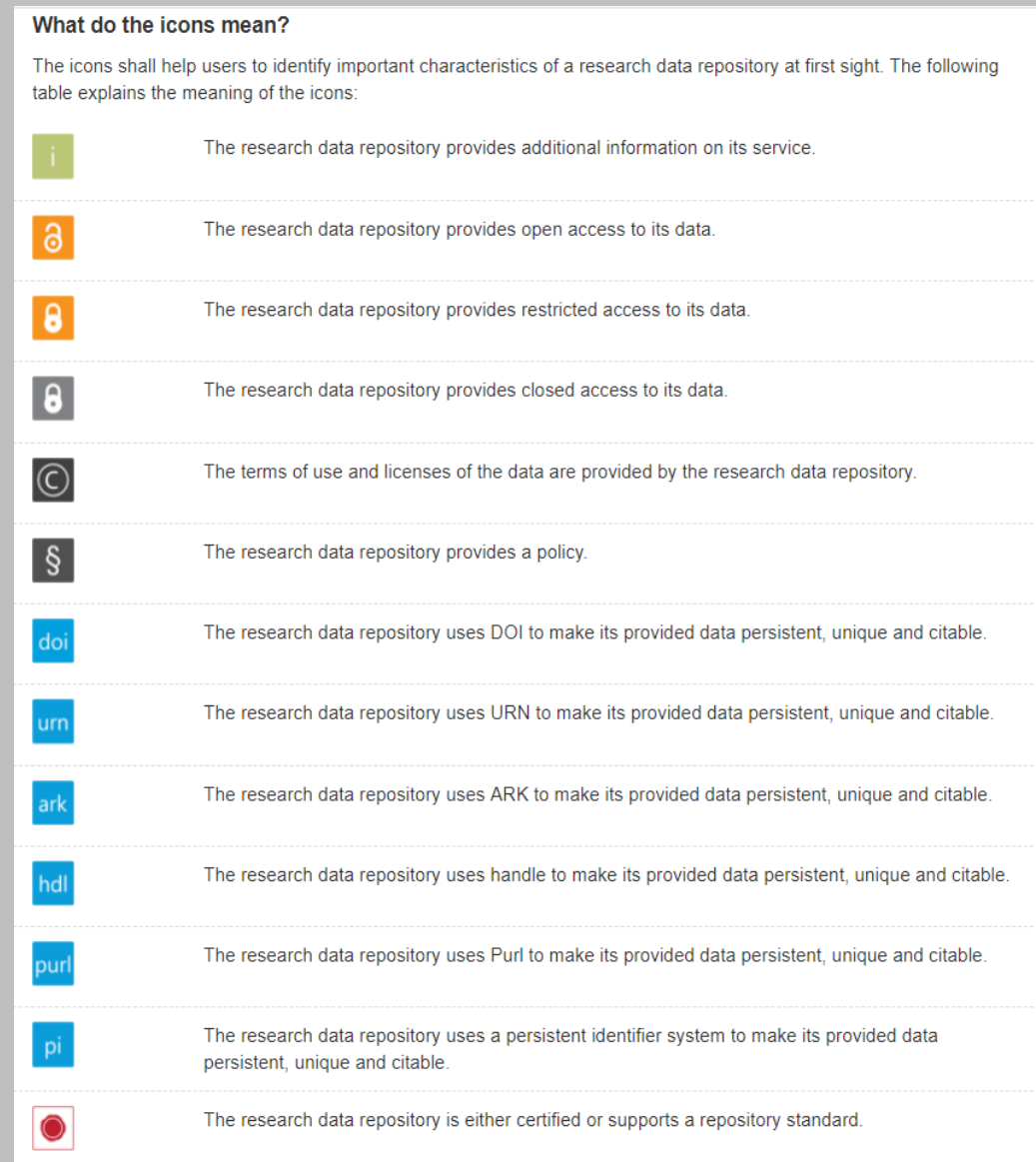
1044 in US

Browse by

- **Subject**
- **Content type**
- **country**

re3data.org





Exercise:

Data Sharing Repository Search: re3data

You can search in several ways:

- Primary Search box
- Click on Search to see the Filter
- Browse by subject
- Browse by content type
- Browse by country

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

Summary

If your data are:

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

Then the results are:

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

courtesy DataONE

Training

- RDS workshops: Fall 2019 and archived
- UK Data Service – Data management training resources
- Lamar Soutter Library (UMASS Medical School): New England Collaborative Data Management Curriculum
- Digital Preservation Coalition – Knowledge Base
- ESIP Data Management Training Clearinghouse: Learning Resources and Data Management Short Course for Scientists
- Open Data Handbook
- Data Scientist Training for Librarians
- Data Observation Network for Earth (DataONE): Education Modules
- Coursera: Research Data Management and Sharing MOOC

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