



Organizing and Sharing Data

Lisa Spiro

September 2017

This workshop draws on materials from the [University of Minnesota Libraries](#), [New England Collaborative Data Management Curriculum](#) and [DataOne](#).

Quick Poll: Raise Your Hand If You Have Ever...

- Forgotten what you called a file and/or where you put it
- Discovered unnecessary duplicates, then struggled over which to keep
- Not had access to needed data in someone else's possession
- Lost data due to hardware failure, lost devices, etc.

What We Will Explore

1. How to understand your data and workflow.
2. How to name & organize files & directories.
3. How to manage versions of data.
4. How to create tidy data.
5. How to document data.
6. How to be ready to share data.
7. How to use tools to manage your data.

1. How to understand your data and workflow



Why Is Organizing Your Data Important?

- Keep track of your data, working more efficiently.
- Prevent data loss.
- Uphold standards of research integrity and [reproducibility](#).
- Meet funder, [university](#) & increasingly journal [requirements](#).
- Make it easier to share and publish data.

>> Be kind to future you!

Use a Data Inventory to Understand, Track & Share Your Data

Plan for, monitor & prepare to share your data by recording:

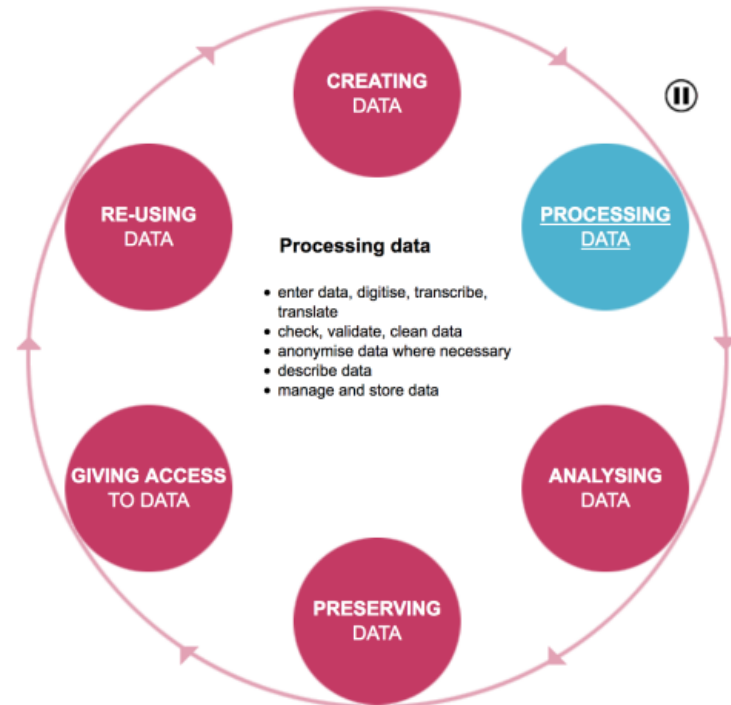
- what the dataset is
- who is responsible for it
- how data were created
- where it is
- how important it is
- who can access & edit it
- where it is stored and preserved

Exercise 1: Jot Down What Might Belong in Your Data Inventory

Data inventory

Develop an Effective Workflow

- Replicable
- Efficient
- Automated
- Something that you will follow, regularly

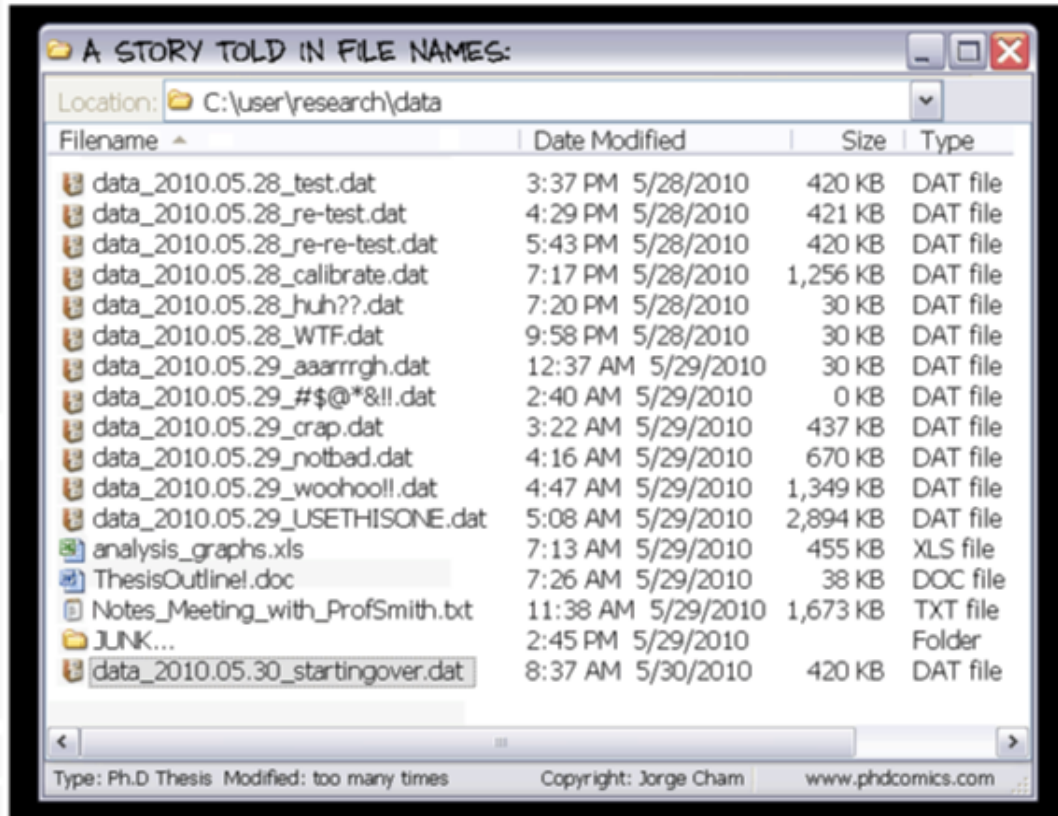


Key Principles

1. Investing some time in organizing your data now will save you time and headaches later.
2. Be clear and consistent.
3. Work out your data organization procedures with collaborators.
4. Document your procedures.
5. Understand that there is no one right way; it's what works for you.

2. How to name & organize files & directories





A Story Told in File Names (PhD Comics)

Principles for Effective Naming

- Data files are **distinguishable** from each other within their containing folder.
- Data files are easy to **locate, browse** and **sort**.
- If data files are moved to other storage platform, their names will retain **useful context**.

File Naming Best Practices

- **Be descriptive:** Use shared, meaningful terminology. Incorporate relevant terms such as project name, place, date, experiment, instrument, subject, etc.

Example: AirQual_Lufkin_Sensor1_201709007

- **Be consistent:** Use the same structure and terms across projects so that files fall into a useful *order* (for sorting) and you can easily identify them.

Example: AvSAT_Ric_2017
AvSAT_Ric_2016
AvSAT_UTx_2017

File Naming Best Practices, II

- **Be concise:** Software may have difficulty processing long file names.
- **Avoid special characters,** like / , . # ?
- **Don't use blank spaces.** Use CamelCharacters or _ to link together keywords.
- **Date/time:** Use yyymmdd rather than Dec09
- **Use leading zeros:**
009DataCollection rather than 9DataCollection (helps with sorting)

Which file naming scheme works the best?

A. bridgedata1
bridgedata2
bridgedata3

C. madisonavebridge_sensor2_20130214
madisonavebridge_sensor2_20130215
madisonavebridge_sensor2_20130216

A. bridge1_sensor2_02142013
bridge1_sensor2_02152013
bridge1_sensor2_02162013

D. madisonavebridge_sensor2_feb142013
madisonavebridge_sensor2_02152013
madbridge_s2_feb162013

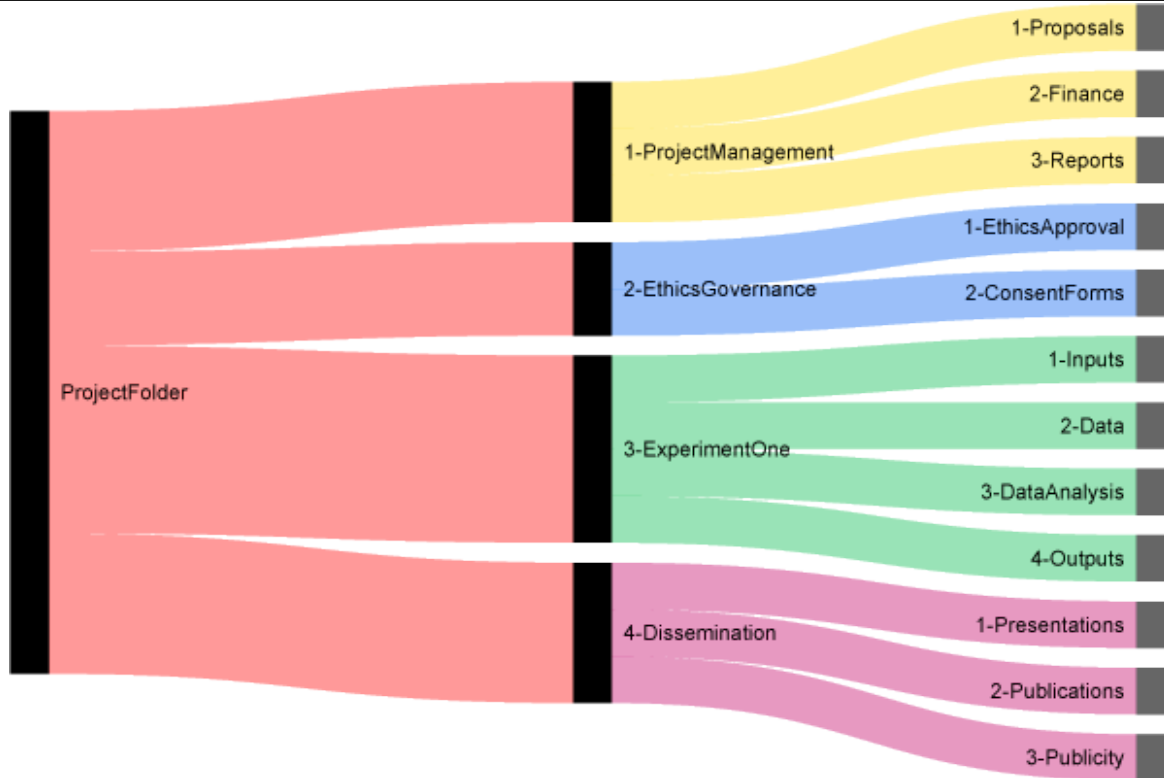
How to Manage Files

Principle	Approach
Data file naming prevents confusion when multiple people are working on shared files.	Establish common conventions for file naming and organization.
Data files can be retrieved not only by the creator but by other users.	Use networked/ cloud based platforms to support collaboration, e.g. Box .

How to Manage Files

Principle	Approach
Data files are not accidentally overwritten or deleted.	Use secure backup: 3 copies of your work 2 different kinds of storage 1 copy offsite Have protocols for handling files.
Different versions of data files can be identified.	Use version control (more later).

Example of Directory Structure



Exercise

Instructions: Review the handout, then partner with 2-3 people to decide on a file naming system in order to archive all files in one folder and sort by interviewee name.

3 minutes to discuss

3. How to manage versions of data.



Which one is authoritative?

DataAnalysis.xls

DataAnalysis2.xls

DataAnalysisSept2017.xls

DataAnalysisFinal.xls

DataAnalysisFinalFINAL.xls

Manual Options for Managing Versions

- Retain original, raw files and significant iterations.
- Use careful file naming: record major changes via whole numbers (v01), minor via an additional number (v02_01)
- Create a version control table:

Version Number	Author	Purpose/Change	Date
0-1	Jackie Wilson, Project Manager	Initial draft – to line manager	12/07/2011
0-2	Jackie Wilson, Project Manager	Consultation draft – to working group	21/08/2011
0-3	Jackie Wilson, Project Manager	Second consultation draft – to working group	08/10/2011
1-0	Jackie Wilson, Project Manager	Final version – approved by Project Board	18/11/2011

Why Use Version Control?

- Keep track of file versions.
- See who does what.
- Access any version of file.
- Synchronize and share, so that latest version is available to all collaborators.
- Roll back changes.
- Enable branches of project.

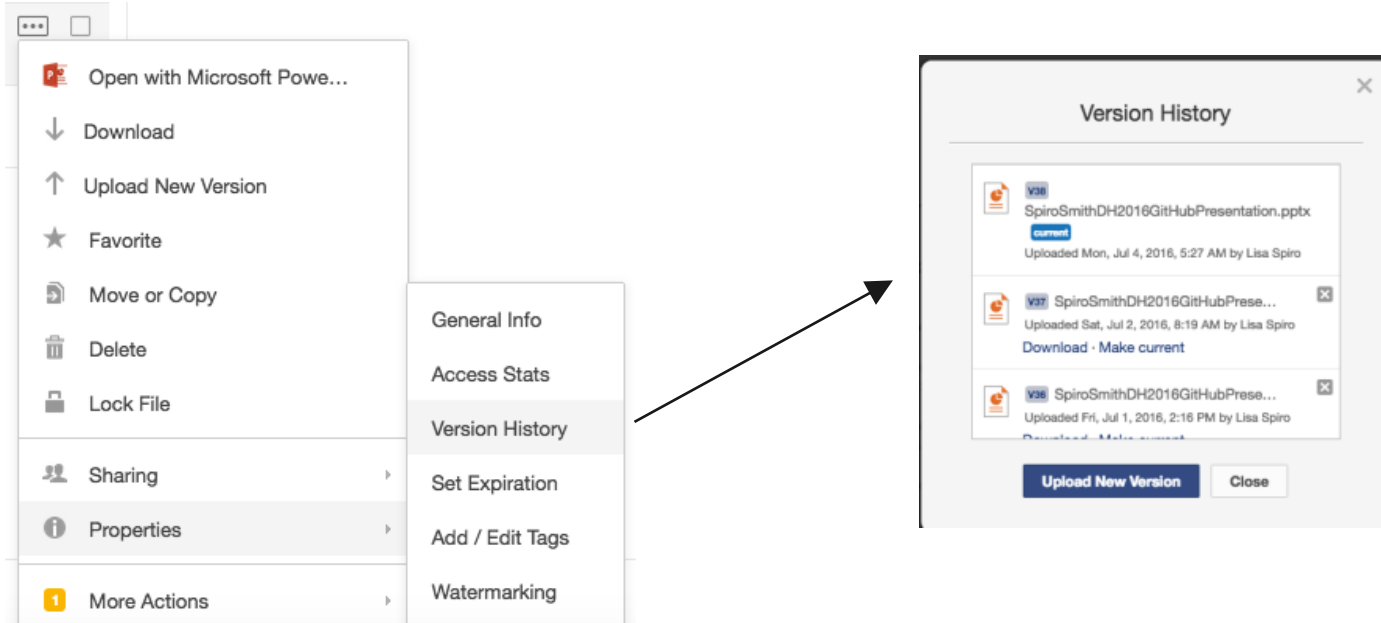
Software for Managing Versions

- Through [Box](#), [Google Drive](#) & other storage services

Version control software:

- [Subversion](#): supported by Rice OIT; free
- [GitHub](#): Public repositories are free. [Researchers](#) can receive to 5 free private repos, research groups up to 20

Accessing Version History on Box.com



Manage and Access Versions of Files with Git(Hub)

The screenshot shows a GitHub repository page for 'rzach / git4phi'. At the top, there are navigation buttons for 'Watch' (3), 'Star' (7), and 'Fork' (4). Below this, a commit titled 'Update README.md' is shown, committed by 'rzach' on Jul 4. The commit message is '1 parent 0a9437b commit f8cba8b8ec50331f6a2d5e3ad777d870e10bae59'. Below the commit message, it says 'Showing 1 changed file with 1 addition and 1 deletion.' The file 'README.md' is shown with a diff view. The diff shows a change in line 8, where a new sentence is added: '+The guide is written in Markdown, the file is git4phi.md, and [can be read here](https://github.com/rzach/git4phi/blob/master/git4phi.md). You can download the latest release, including a printable PDF version, [here](https://github.com/rzach/git4phi/releases)'. The previous line 8 content is shown in red, indicating it was deleted or replaced.

- Track changes to files
- Collaborate
- Roll back to earlier versions

4. How to create tidy data.



Keep Your Data Tidy

- Make each variable a column & each observation a row
- Make column headers variable names
- Atomize your data; put only a single piece of information in each cell (e.g. city, state, country)
- Be consistent how you will handle empty values (e.g. NULL, leave blank)

See Hadley Wickham, [“Tidy Data”](#) (2014)

Messy vs. Tidy Data

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

(a) Molten data

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

(b) Tidy data

Wickham

Table 10: Tidying the TB dataset requires first melting, and then splitting the `column` column into two variables: `sex` and `age`.

More on Tidiness

- Be explicit about measurement type (e.g. lb, kg)
- Document your variables
- Use standard (ideally non-proprietary) formats for data, e.g. CSV, .txt

The Problems with Messy Data

- Difficult to analyze
- Requires time to clean
- Confusing to other researchers– and to Future You
- Raises questions about your credibility as a researcher

	A	B	C	D	E
1	Date	ID	Plasmid	Primer	Results
2	970910	E1 5411	MDM970905E1	MSAF5411	unreadable
3	970911	J1 5411	MDM970905J1	MSAF5411	unreadable
4		E5411	MDM970905E	MSAF5411	T173A, HA tag present
5	970917	J5411	MDM970905J	MSAF5411	S191A, HA tag present
6	971104	A4	AH971022A4	MSAF8259	GST clone -- wrong, no GSTI
7		A6	AH971204A6	pUC19SP2	U.S.E. -- clone wrong
8	971216	C9	AH971216C9	pUC19SP2	U.S.E. -- clone wrong
9		A15	AH971230A15	pUC19SP2	R261A, L263A
10	980114	A5	AH971230A5	pUC19SP2	WT
11		D9	AH971230D8	MSAF1818	N-terminal HA tag present
12	980313	AH2	AH971118A7	MSAF1818	HA tag present
13	980330	A2	AH980325A2	MSAF1818	R261A, L263A, R269A, F271A
14		C1	AH980325C1	MSAF8259	R261A, L263A
15		C2	AH980325C2	MSAF8259	unreadable
16	980402	C3	AH980325C3	MSAF8259	R261A, L263A
17		C4	AH980325C4	MSAF8259	R261A, L263A
18		C5	AH980325C5	MSAF8259	no mutation
19	980424	E8	AH980325E8	MSAF8259	L263A only
20	980504	H1B	random mut. H1B	MSAF8259	221-284 no mutation
21		430A1	AH980430A1	MSAF8259	WT -- no R269A, F271A
22	980507	430A2	AH980430A2	MSAF8259	WT -- no R269A, F271A
23		325E20	AH980325E20	MSAF8259	L263A only
24		325E21	AH980325E21	MSAF8259	correct, R261A, L263A
25		325E22	AH980325E22	MSAF8259	L263A only
26	980511	325E26	AH980325E26	MSAF8259	WT
27		325E28	AH980325E28	MSAF8259	L263A only
28		325E30	AH980325E30	MSAF8259	WT
29	980716	B12REV	AH980707B12	reverse	215-284 3xHA correct
30		C1REV	AH980707C1	reverse	226-284 3xHA correct
31		A1REV	AH980717A1	reverse	not close enough to primer
32	980722	A3REV	AH980717A3	reverse	WT (incorrect)
33		A7REV	AH980717A7	reverse	unreadable
34	980902	A23REV	AH980707A23	reverse	221-284 3xHA correct
35		A11	AH981015A11	1818	R269A, F271P
36	981021	A4	AH981015A4	1818	R269A, F271A
37	A11	AH981015A11	1818	R269A, F271A

What errors do you see with this spreadsheet?

What problems might this pose to researchers?

5. How to document data.



Why Document Data?

- Makes it easier for you to interpret your own data
- Facilitates collaboration, sharing, and reuse
- Ensures successful long-term preservation of findings

Create a Readme File

- Simple way to describe & contextualize a dataset.
- Usually plaintext.
- Typically named “readme.”

Typical Contents of Readme File

- **What:**
 - Title
 - Description
- **When:** date of data collection
- **Who:** name & contact info of creator
- **Where:** location where data was captured
- **How:**
 - Method of data collection, creation or processing
 - Restrictions on accessing files

Files to replicate Sean Bolks and Richard J. Stoll, “[The Arms Acquisition Process](#): The Effect of Internal and External Constraints on Arms Race Dynamics,” *The Journal of Conflict Resolution* 44, no. 5 (October 1, 2000): 580–603.

File	Content
table1.dta	Stata data file with data for Table 1
table1.do	Stata .do file with commands to replicate Table 1
table2.dta	Stata data file with data for Table 2
table2.do	Stata .do file with commands to replicate Table

Simple Example of a ReadMe File

Create a Codebook

“A codebook is an essential document that informs the data user about the **study, data file(s), variables, categories**, etc., that make up a complete dataset. The codebook may include a dataset’s record layout, list of variable names and labels, concepts, categories, cases, missing value codes, frequency counts, notes, universe statements, and so on.”

Codebook Example



COOPERATIVE INSTITUTIONAL RESEARCH PROGRAM
at the HIGHER EDUCATION RESEARCH INSTITUTE AT UCLA

2017 CIRP Freshman Survey (Codebook)

#	Variable Name	Variable Description
	ACE SUBJID STUID	College I.D. Subject I.D. Student I.D. as entered on form
	GRPA GRPB	Group Code A Group Code B
1	SEX	Your sex: 1 = Male 2 = Female
2	TRANSGENDER	Do you identify as transgender? 1=No 2=Yes
3	YRGRADHS	In what year did you graduate from high school? 1=2017 2=2016 3=2015 4=2014 or earlier 5=Did not graduate but passed G.E.D. test 6=Never completed high school

Exercise

Think through creating a readme file for one of your datasets (real or imagined) or the “Dr. Psi” data using this template from [Cornell](#).

See “Guidelines for writing ‘readme’ style metadata”
http://data.research.cornell.edu/sites/default/files/SciMD_ReadMe_Guidelines_v4_1_0.pdf

6. How to prepare to share data.



Why Share Data?

- Meet reproducibility standards.
- Enable your data to be re-used– and cited.
“studies that made data available in a public repository received 9%... more citations than similar studies for which the data was not made available.”
(Piwovar & Vision 2013)
- Foster collaboration.
- Comply with journal or funder requirements.

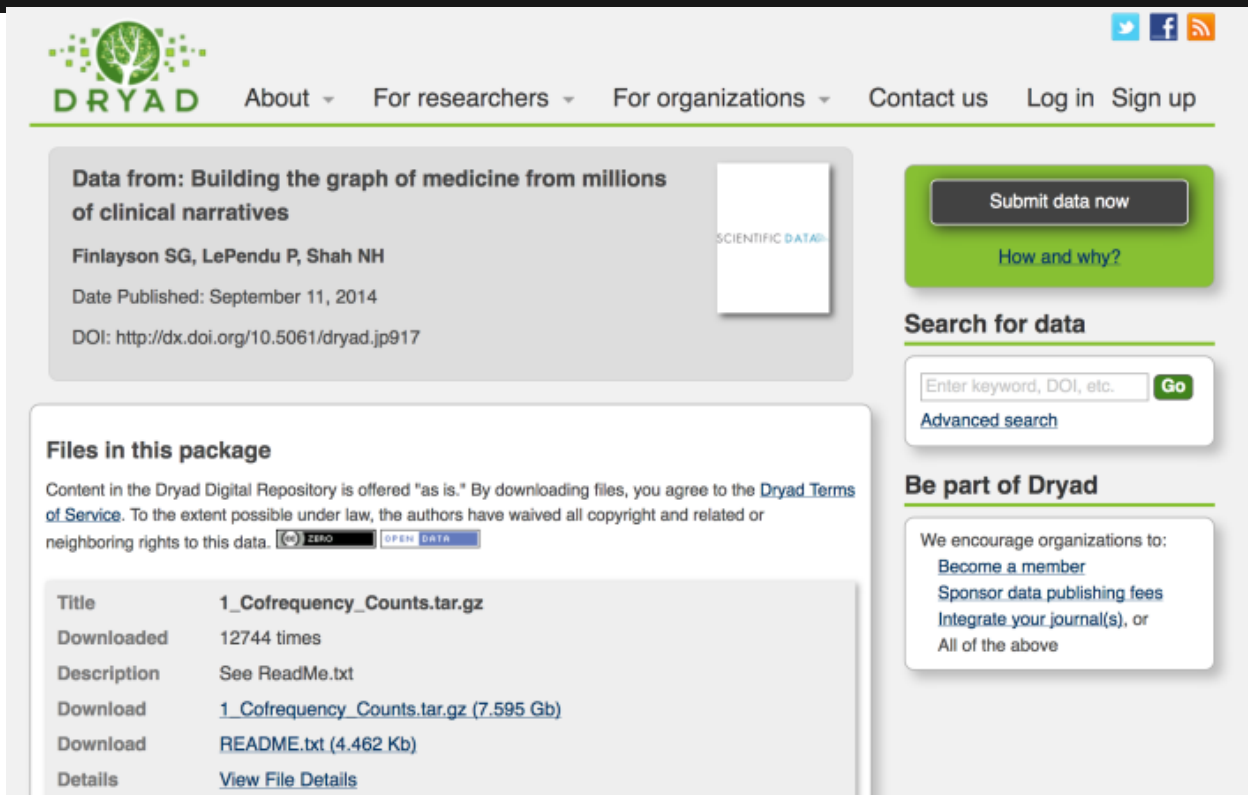
Caveats about Data Sharing

- Check with your adviser, PI, and collaborators about data sharing.
- Be aware of any restrictions on data sharing, e.g. confidentiality or intellectual property.
- Recognize the time required to get your data into shape for sharing.

Data Archiving & Sharing Options

- Deposit in an appropriate disciplinary repository
 - Nature, “Recommended Data Repositories”:
<https://www.nature.com/sdata/policies/repositories>
 - PLOS Guide: <http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories>
 - Re3data: <http://www.re3data.org/>
- Share small to medium datasets through the Rice Digital Scholarship Archive:
<https://scholarship.rice.edu/handle/1911/77660>

Example of Data Repository: Dryad



The screenshot shows the Dryad website interface. At the top, there is a navigation bar with the Dryad logo (a green tree icon) and the text "DRYAD". To the right of the logo are links for "About", "For researchers", "For organizations", "Contact us", "Log in", and "Sign up". Social media icons for Twitter, Facebook, and RSS are also present.

The main content area features a data package entry with the following details:

- Data from:** Building the graph of medicine from millions of clinical narratives
- Authors:** Finlayson SG, LePendu P, Shah NH
- Date Published:** September 11, 2014
- DOI:** <http://dx.doi.org/10.5061/dryad.jp917>

To the right of this entry is a placeholder for a thumbnail image labeled "SCIENTIFIC DATA".

Below the package information is a section titled "Files in this package". It contains a table with the following entries:

Title	1_Cofrequency_Counts.tar.gz
Downloaded	12744 times
Description	See ReadMe.txt
Download	1_Cofrequency_Counts.tar.gz (7.595 Gb)
Download	README.txt (4.462 Kb)
Details	View File Details

Below the table are two buttons: "CC0" and "OPEN DATA".

On the right side of the page, there is a green button labeled "Submit data now" with a link "How and why?" below it. Below that is a search box with the text "Enter keyword, DOI, etc." and a "Go" button. A link for "Advanced search" is also present.

At the bottom right, there is a section titled "Be part of Dryad" with the following text: "We encourage organizations to: [Become a member](#), [Sponsor data publishing fees](#), [Integrate your journal\(s\)](#), or All of the above".


How can I make my data submission as accessible and reusable as possible? [Close](#)

- Provide ALL files needed to replicate your results, including code. One way to help ensure completeness is to explicitly link your data files (through their titles/descriptions) to the figures and analyses in your publication.
- Submit your data files in non-proprietary formats from which data can be easily extracted (e.g., CSV rather than PDF).
- Keep your file names short, informative, unique, and free of special characters.
- Consider submitting your data files in multiple formats if you think that will enhance their ability to be reanalyzed. View [additional guidance](#) and a list preferred Dryad file formats.
- Provide descriptive information within your data files (e.g., column headers in a spreadsheet).
- Provide a [README file](#) that provides contextual information about the data file so that it can be interpreted correctly.
- Provide titles, descriptions and keywords for your datafiles, to make the data more discoverable and to assist in understanding the relationship of the datafile to the publication.

<http://datadryad.org/pages/faq#deposit>

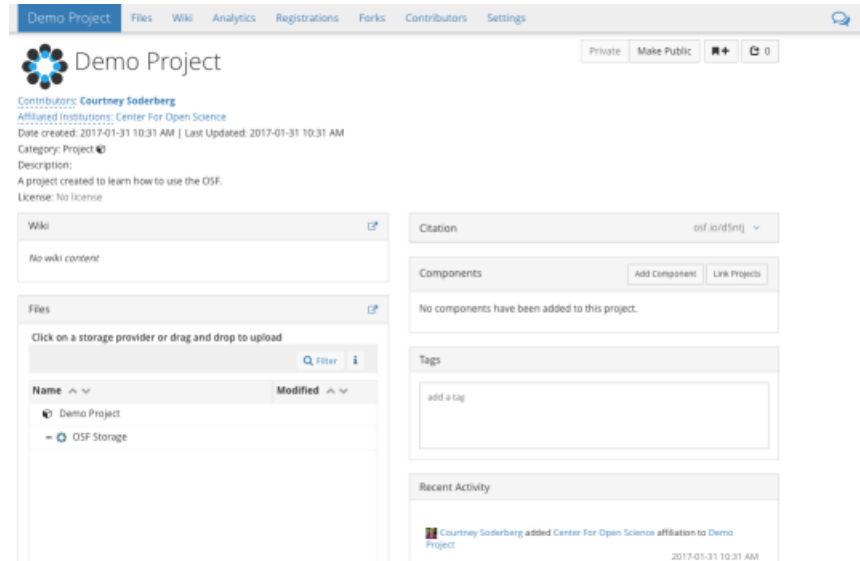
Preparing to Share Your Data

7. How to use tools to manage your data.



Consider Using Open Science Framework to Manage Your Research

- Organize files in one place
- Share with collaborators
- Control files access
- Integrate with tools like Box
- Track versions
- Make work citable
- Facilitate reproducibility
- Free & open source



The screenshot displays the OSF interface for a 'Demo Project'. The top navigation bar includes 'Demo Project', 'Files', 'Wiki', 'Analytics', 'Registrations', 'Forks', 'Contributors', and 'Settings'. The project page shows the following details:

- Contributors:** Courtney Soderberg
- Affiliated Institutions:** Center For Open Science
- Date created:** 2017-01-31 10:31 AM | **Last Updated:** 2017-01-31 10:31 AM
- Category:** Project
- Description:** A project created to learn how to use the OSF.
- License:** No license

The interface is divided into several sections:

- Wiki:** Currently empty, showing 'No wiki content'.
- Files:** A section for file management with a search filter and a list of files. The list includes 'Demo Project' and 'OSF Storage'.
- Citation:** A section for generating citations, with the URL 'osf.io/d5ntj'.
- Components:** A section for adding components, with buttons for 'Add Component' and 'Link Projects'.
- Tags:** A section for adding tags, with a text input field and a 'add a tag' button.
- Recent Activity:** A section showing recent activity, with a record: 'Courtney Soderberg added Center For Open Science affiliation to Demo Project' on 2017-01-31 10:31 AM.

<https://osf.io/>



Reproducibility Project: Cancer Biology

Contributors: [Tim Errington](#), [Fraser Elisabeth Tan](#), [Joelle Lomax](#), [Nicole Perfito](#), [Elizabeth Iorns](#), [William Gunn](#), [Brian A. Nosek](#), [Stuart Buck](#), [Erin Griner](#), [Nimet Maherali](#), [Mathew Veal](#), [Michael McCarthy](#), [Samuel LaBarge](#), [Hyun Yong Jin](#), [Christine Schaner Tooley](#), [Claudia-Gabriela Mitrofan](#), [Tim Smith](#), [Robert L Judson](#), [Matthew Cook](#), [Sarah Statt](#), [Nicole Vasilevsky](#), [Stefano Biressi](#), [Kevin Poindexter](#), [Kartoa Chow](#), [Heidi Hilton](#), [Hildegard Mack](#), [Teresa Krieger](#), [Minyoung Anna Lim](#), [Miguel A. S. Cavadas](#), [Michael V. Gormally](#),

Affiliated institutions: [Center For Open Science](#), [Laura and John Arnold Foundation](#)

Date created: 2013-10-08 06:31 PM | Last Updated: 2017-08-22 12:08 PM

Category: Project

Description: We are conducting a study to investigate the replicability of cancer biology studies. Selected results from a substantial number of high-profile papers in the field of cancer biology published between 2010-2012 are being replicated by the Science Exchange network.

Wiki

Biology is a collaboration between [Science Exchange](#) and the [Center for Open Science](#), and is independently replicating a subset of experimental results from a number of high-profile papers in the field of cancer biology published between 2010-2012 using the Science Exchange network of expert scientific labs

Citation

osf.io/e81xl

Components

Replication Studies

[Errington, Tan, Lomax & 82 more](#)

218 contributions

Identification Analysis of RP:CB

[Errington, Vasilevsky & Haendel](#)

45 contributions

<https://osf.io/ezcuj/>

Other Tools

See tool lists from:

- [Data ONE](#)
- [Digital Curation Centre](#)
- [UCLA Library](#)

[Home](#) » [Resources](#) » [Software Tools Catalog](#) » [Software Tools](#)

Resources

Tools

[Investigator Toolkit](#)
[Data Management Planning](#)
[Software Tools Catalog](#)

Materials

[Publications](#)
[Best Practices](#)
[Data Life Cycle](#)
[Librarian Outreach Kit](#)
[Developer Resources](#)
[Research Notebooks](#)

Featured Tool

[Archivematica](#)

Software Tools

3D World Studio



Tags: [GIS](#), [map](#), [visualization](#)

Adobe Illustrator



Illustrator CS5

Tags: [graphics](#), [image](#), [visualization](#)

Resources

- Borer, Elizabeth T., et al “[Some Simple Guidelines for Effective Data Management.](#)” *Bulletin of the Ecological Society of America* (2009): 205–14.
- DataOne Primer on Data Management, https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf
- Dataverse, *Data Management Plans*, <http://best-practices.dataverse.org/data-management/>
- ICPSR *Guide to Social Science Data Preparation and Archiving*, <http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/>
- Svend Juul et al, “Take good care of your data,” <http://www.epidata.dk/downloads/takecare.pdf>

More Resources

- Nosek, Brian. “Improving My Lab, My Science With the Open Science Framework,” <https://www.psychologicalscience.org/observer/improving-my-lab-my-science-with-the-open-science-framework>
- UK Data Archive, *Managing and Sharing Data: Best Practices for Researchers*, <http://www.data-archive.ac.uk/media/2894/managingsharing.pdf>

Thanks!

Please contact researchdata@rice.edu with any questions.

Visit us online at <http://researchdata.rice.edu/>.

Help us shape future workshops! Please complete this [evaluation](http://library.rice.edu/requests/course-evaluation-form): <http://library.rice.edu/requests/course-evaluation-form>

Course: Organizing & Sharing Data

Instructor: Lisa Spiro