Welcome!

Please type into the chat your department and what you hope to get out of this workshop.

Research Data Management

Lisa Spiro & Catherine Barber

October 4, 2022

This workshop draws heavily on materials from the University of Minnesota Libraries, New England Collaborative Data Management Curriculum, and MIT Libraries

Poll: Have you ever…

- Forgotten what you called a file or where you put it
- Discovered unnecessary duplicates, then struggled over which to keep
- Been unsure about who has responsibility for managing files
- Lost data due to hardware failure, lost devices, etc.

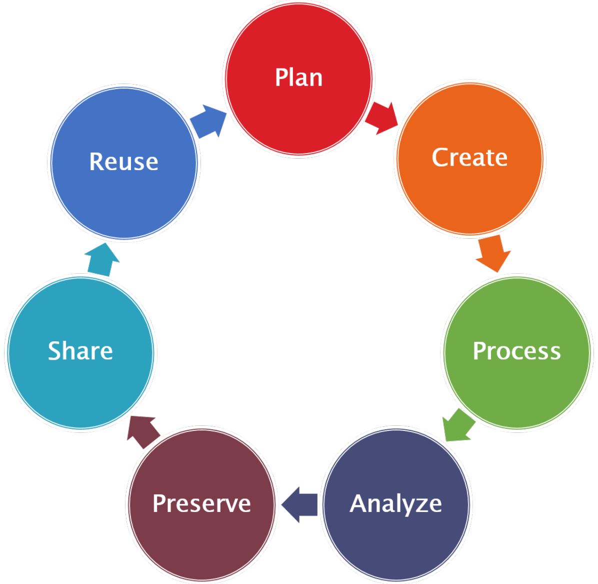
Objectives for This Session

1. Understand the importance of managing data.
2. Learn how to create a good data management plan.
3. Name and organize your files effectively.
4. Create tidy data.
5. Manage versions.
6. Document your data.
7. Know options for storing, backing up and archiving your data.

Why Managing Your Data Matters

What is data management?

The process of storing, organizing, describing, preserving, and sharing data so that research results can be validated, data can be understood, and future use is facilitated.



https://biblio.uottawa.ca/en/services/faculty/research-data-management/what-research-data-management

Why Is Managing Your Data Important?

- Keep track of your data, work more efficiently.
- Prevent data loss.
- Uphold standards of research integrity.
- Make it easier to share and re-use data.
- Meet funder, university, and increasingly journal requirements.
- Be kind to Future You and your collaborators.

Data Wisdom

If the data you need still exists;
If you found the data you need;
If you understand the data you found;
If you trust the data you understand;
If you can use the data you trust;
Someone did a good job of data management.

- *Rex Sanders, USGS*

Plan

Data Management Plan (NSF)

- types of data
- standards for data and metadata format and content
- policies for access and sharing
- policies and provisions for re-use, re-distribution, and the production of derivatives
- plans for archiving data, samples, and other research products, and for preserving access to them

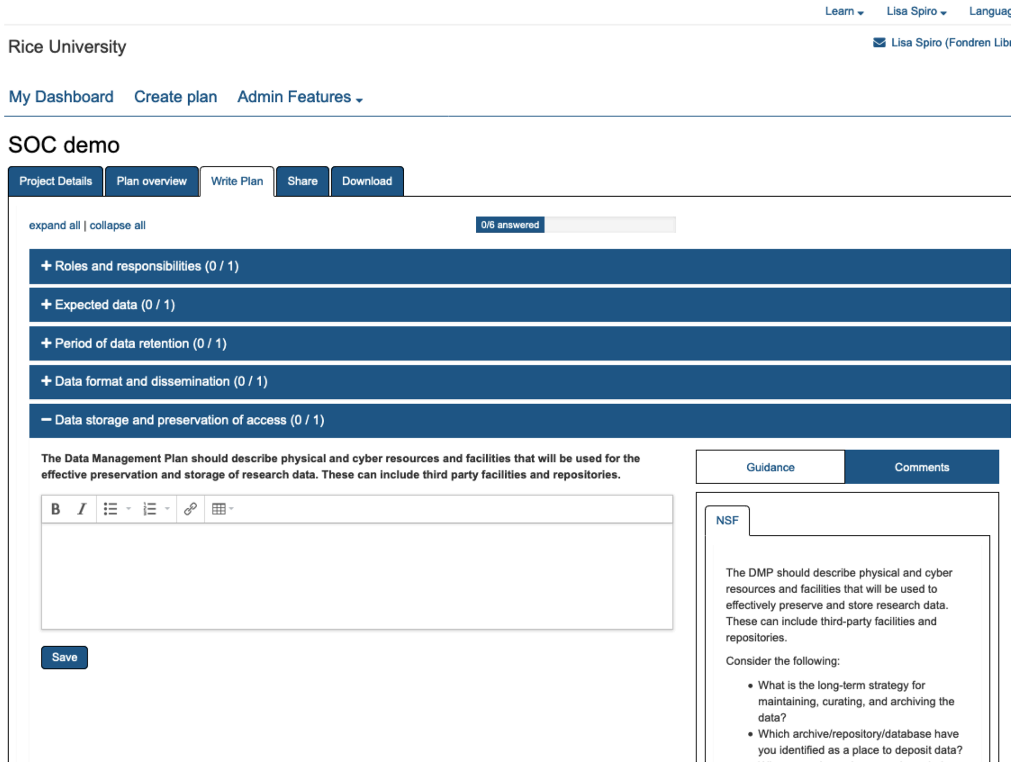
From NSF Proposal Preparation Instructions

Data Inventory

Plan for, monitor, and prepare to share your data by recording in a spreadsheet:

- what the dataset is (title, description, date)
- who is responsible for it (owner, creator, steward)
- where it is stored and preserved (location, duplicates, versions)
- how important it is
- who can access and edit it (rights, restrictions)
- how the data will be used

Create a Data Management Plan Using DMP Tool



The Data Management Plan should describe physical and cyber resources and facilities that will be used for the effective preservation and storage of research data. These can include third party facilities and repositories.

The DMP should describe physical and cyber resources and facilities that will be used to effectively preserve and store research data. These can include third-party facilities and repositories.

Consider the following:

- What is the long-term strategy for maintaining, curating, and archiving the data?
- Which archive/repository/database have you identified as a place to deposit data?

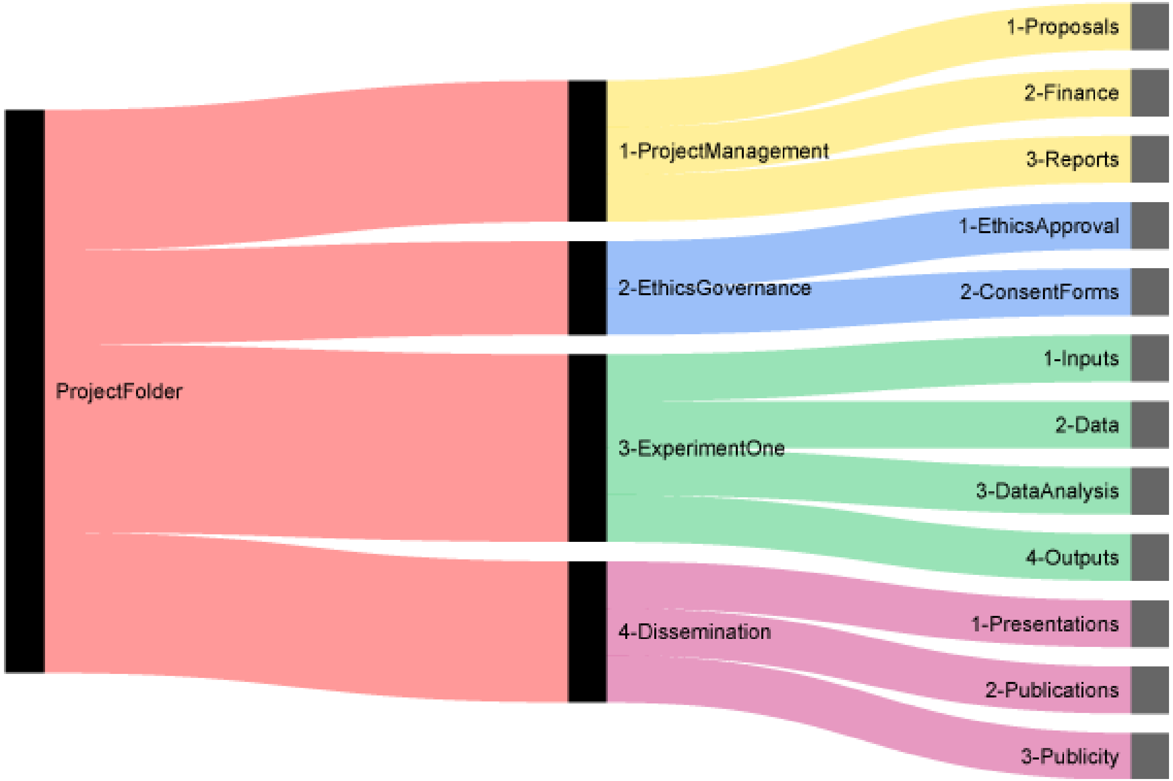
https://dmptool.org/

Key Principles for Data Management Planning

- Investing time in organizing your data now will save you time later.
- Be clear and consistent.
- Document your procedures.
- Work with your collaborators to define data management roles and responsibilities.
- Use what works for you and your collaborators.

Organize Your Data

Example of a Directory Structure

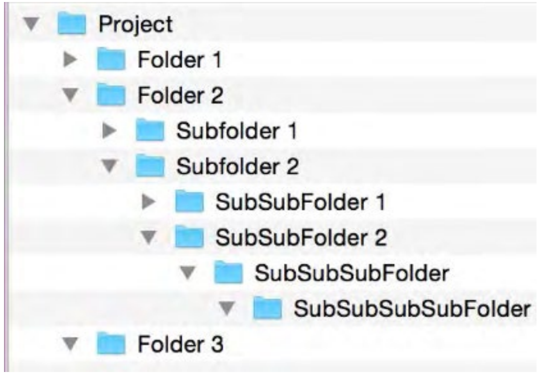


Nikola Vukovic

How to Create a Hierarchical File System

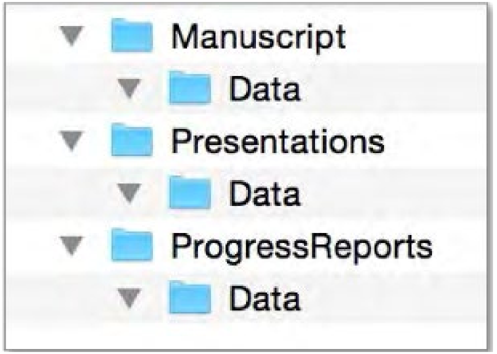
- Organize your files in a predictable, easy-to-sort way.
- Use relevant categories to organize folders, e.g.:
  - Activity (e.g., interviews, experiments)
  - Stage (raw, active, completed)
- Select a meaningful naming convention for folders.

What to Avoid in a File System…



- ▼ Project
  - ▶ Folder 1
  - ▼ Folder 2
    - ▶ Subfolder 1
    - ▼ Subfolder 2
      - ▶ SubSubFolder 1
      - ▼ SubSubFolder 2
        - ▼ SubSubSubFolder
          - ▼ SubSubSubSubFolder
  - ▼ Folder 3

Too much depth

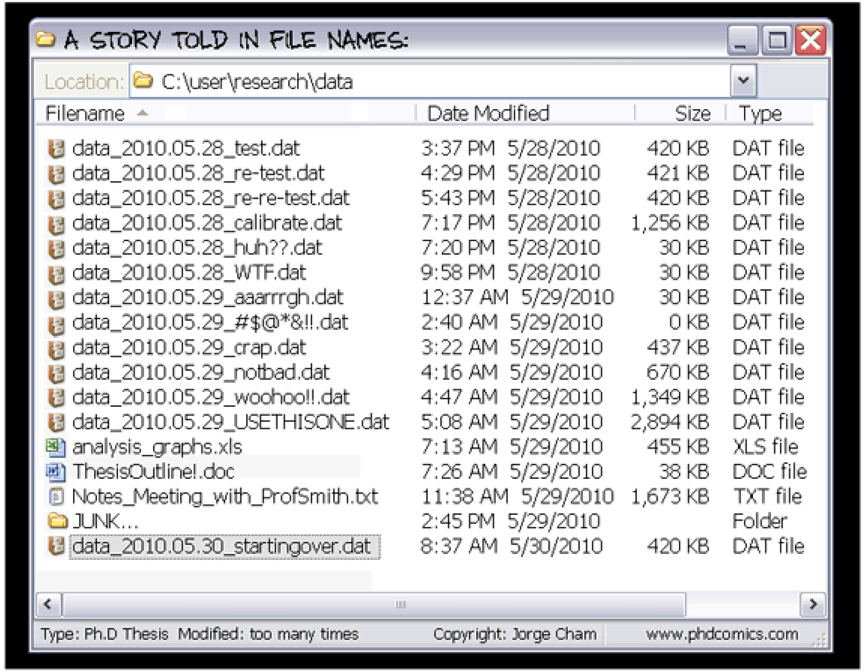


- ▼ Manuscript
  - ▼ Data
- ▼ Presentations
  - ▼ Data
- ▼ ProgressReports
  - ▼ Data

Overlapping categories

MIT

The Problem of File Names



A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

PhD Comics

Principles for Effective File Naming

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

(EDINA and Data Library, n.d.) | RDMRose

File Naming Best Practices

- Be descriptive.
    - Use shared, meaningful terminology.
    - Incorporate relevant terms.
    - Example:
        - AirQual_Lufkin_Sensor1_20170907
- Be consistent.
    - Use the same structure and terms across projects
    - Example:
        - AvSAT_Ric_2017
        - AvSAT_Ric_2016
        - AvSAT_UTx_2017

Guidelines for File Naming

Guideline	**Example**
Avoid special characters, like / , . # ?	Exp01a.xls, NOT Exp#1.a.xls
Don’t use blank spaces. Use CamelCharacters or _ to link keywords.	Site01_Sensor002, NOT Site 1 Sensor 2
Use yyyymmdd for **dates**.	20180617, NOT 061718
Use leading zeroes, e.g., 0001, 001, etc.	Experiment002.xls, NOT Experiment2.xls

Which file naming scheme works the best?

A. bridgedata1
bridgedata2
bridgedata3

B. bridge1_sensor2_02142013
bridge1_sensor2_02152013
bridge1_sensor2_02162013

C. madisonavebridge_sensor2_20130214
madisonavebridge_sensor2_20130215
madisonavebridge_sensor2_20130216

D. madisonavebridge_sensor2_feb142013
madisonavebridge_sensor2_02152013
madbridge_s2_feb162013

University of Minnesota Libraries

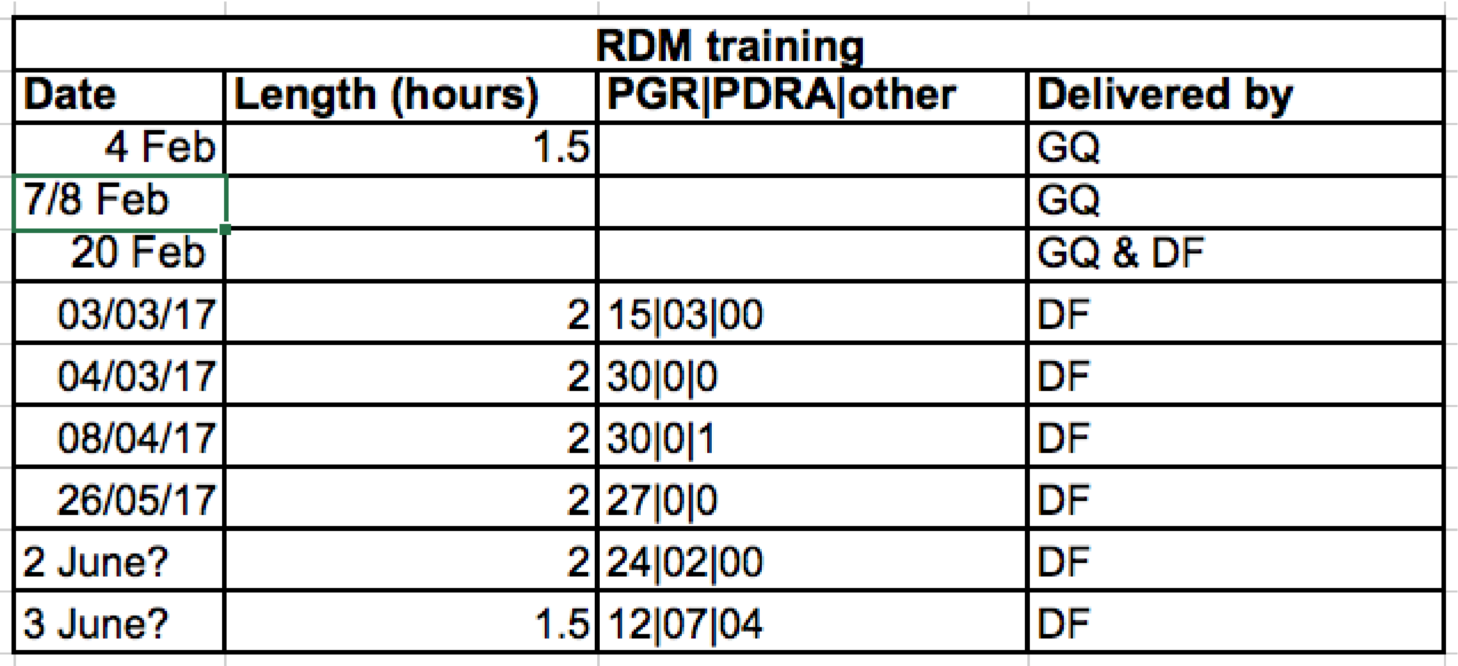
Exercise: File Naming Scheme

Look at the handout at **https://tinyurl.com/FIleNamingExercise**

What file naming scheme would you create to make it easy to find, sort, and understand files? Discuss in your breakout room. (approx. 5 minutes)

Create Tidy Data

Example of Messy Data



RDM training

Date	Length (hours)	PGR|PDRA|other	Delivered by
4 Feb	1.5		GQ
7/8 Feb			GQ
20 Feb			GQ & DF
03/03/17	2	15|03|00	DF
04/03/17	2	30|0|0	DF
08/04/17	2	30|0|1	DF
26/05/17	2	27|0|0	DF
2 June?	2	24|02|00	DF
3 June?	1.5	12|07|04	DF

The Problems with Messy Data

- Difficult to analyze
- Requires time to clean
- Confusing to other users and to Future You
- Raises questions about your credibility

Keep Your Data Tidy

- Make each variable a column and 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 in how you will handle empty values (e.g. NULL, leave blank).

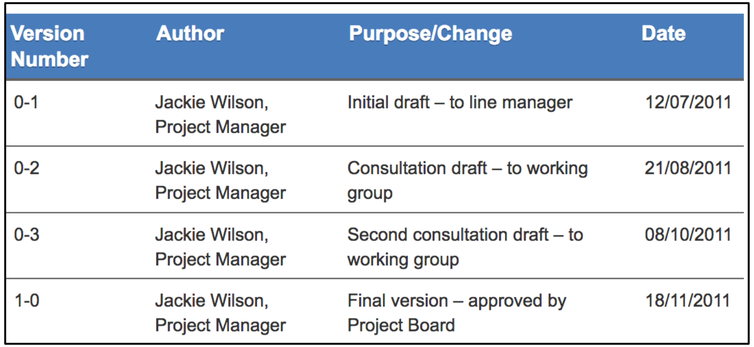
Manage Versions

Versioning: 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:
  - major changes: whole numbers (v01)
  - minor changes: additional number (v02_01)
- Put older versions in an archive folder.
- Create a version control table.

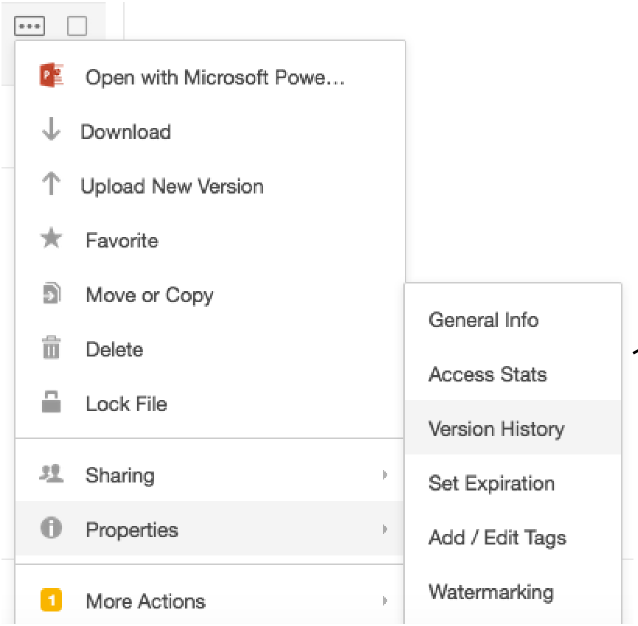


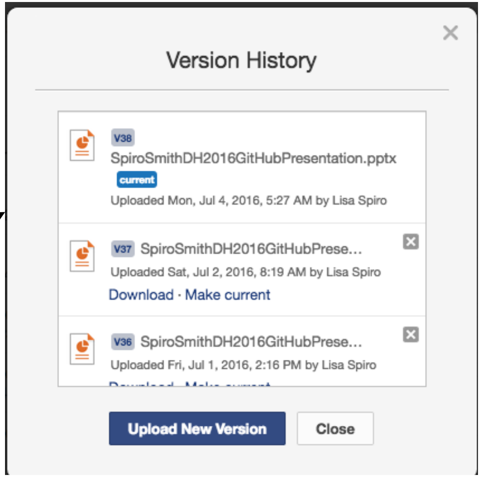
Version Number	Author	Purpose/Change	Date
0-1	Jackie Wilson, Project Manager	Initial draft – to line manager	12/07/2011
0-2	Jackie Wilson, Project Manager	Consultation draft – to working group	21/08/2011
0-3	Jackie Wilson, Project Manager	Second consultation draft – to working group	08/10/2011
1-0	Jackie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

Software for Managing Versions

- Accessing multiple versions:
  - Box, Google Drive, and other storage services
- Version control software:
  - GitHub: Researchers and educators can receive GitHub Team (unlimited repositories) for free.

Accessing Version History on Box.com





V38 SpiroSmithDH2016GitHubPresentation.pptx current

Uploaded Mon, Jul 4, 2016, 5:27 AM by Lisa Spiro

V37 SpiroSmithDH2016GitHubPrese...

Uploaded Sat, Jul 2, 2016, 8:19 AM by Lisa Spiro

Download · Make current

V36 SpiroSmithDH2016GitHubPrese...

Uploaded Fri, Jul 1, 2016, 2:16 PM by Lisa Spiro

Download · Make current

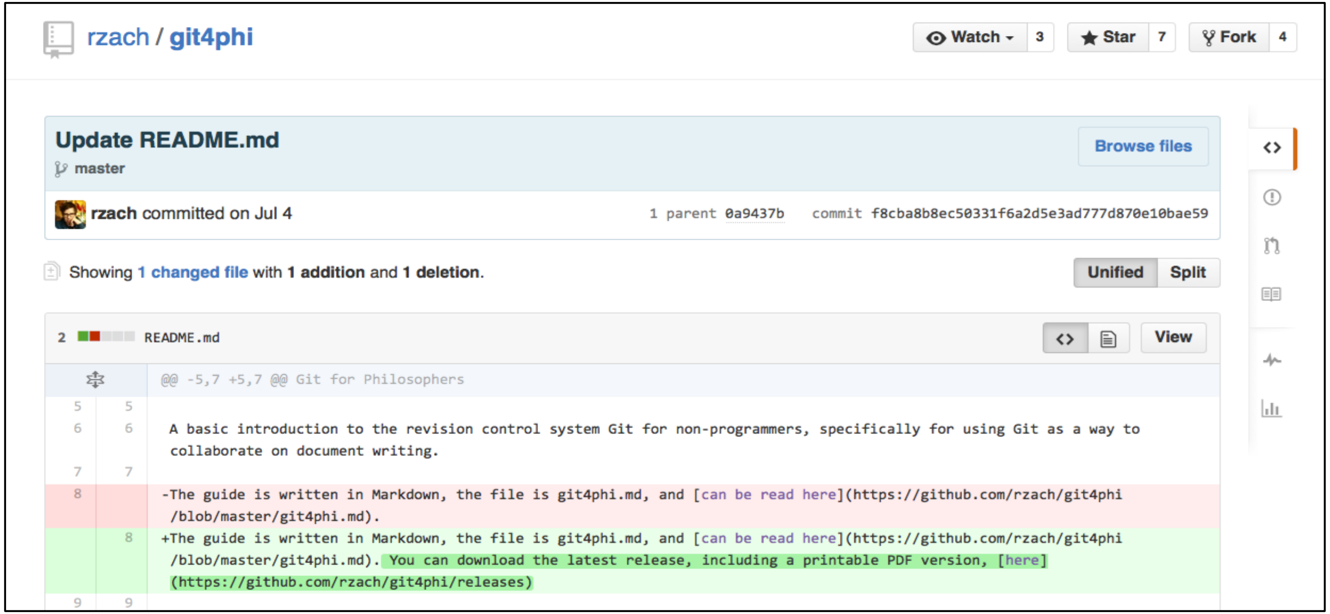
https://community.box.com/t5/Organizing-and-Tracking-Content/Accessing-Version-History/ta-p/50452

Version Control Software

“Version control is a system that records changes to a file or set of files over time so that you can recall specific versions later.” (Pro Git)

- See who does what.
- Access any version of file.
- Roll back changes.
- Enable new branches of project.

GitHub



Update README.md

master

rzach committed on Jul 4

1 parent 0a9437b commit f8cba8b8ec50331f6a2d5e3ad777d870e10bae59

Showing 1 changed file with **1 addition** and **1 deletion**.

2 README.md

```
@@ -5,7 +5,7 @@ Git for Philosophers

 A basic introduction to the revision control system Git for non-programmers, specifically for using Git as a way to
 collaborate on document writing.

-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).
+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)
```

Document your Data

What information would you want to know about this file?

ObscureFile.txt

Enter questions into the chat.
(For example, “who created the file?”)

Why Document Data?

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

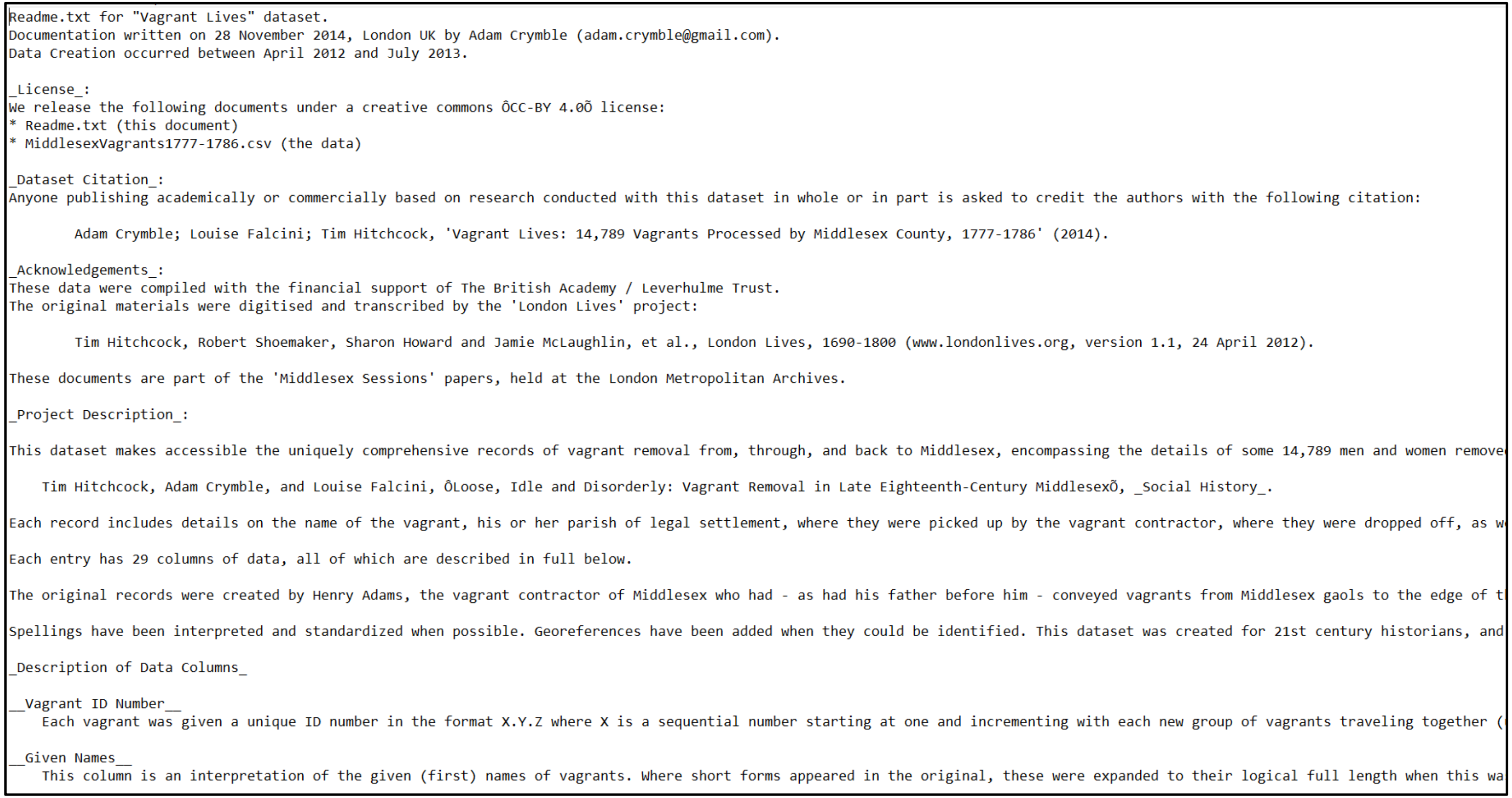
New England Collaborative Data Management Curriculum

Metadata and Readme Files

- Typical contents:
  - 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

https://data.research.cornell.edu/content/readme

Detailed ReadMe file from Zenodo



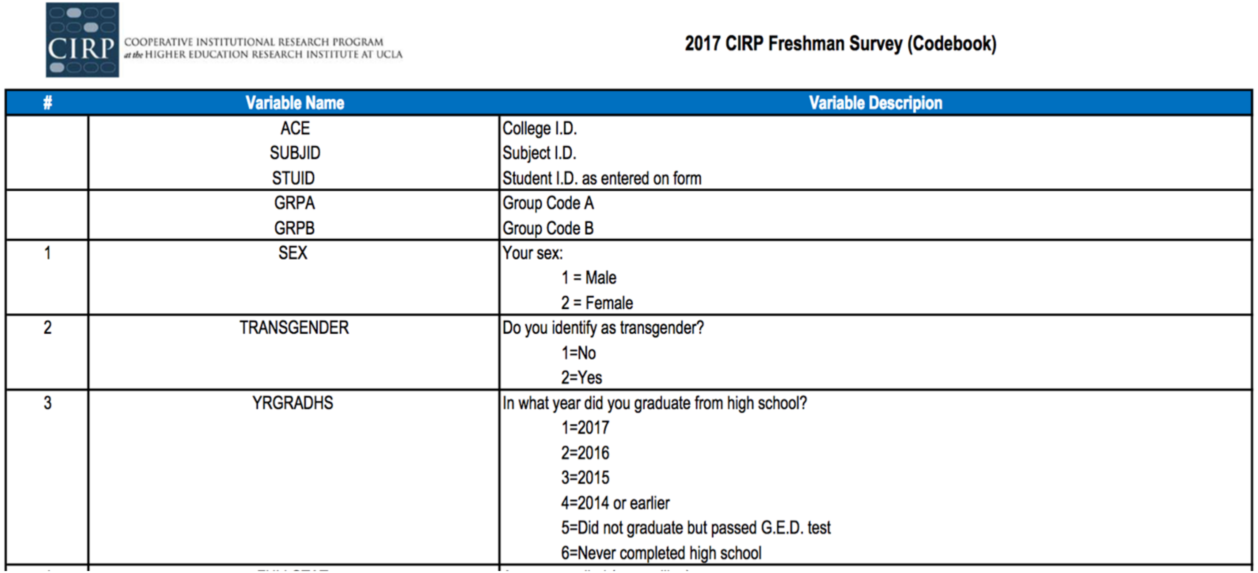
```
Readme.txt for "Vagrant Lives" dataset.
Documentation written on 28 November 2014, London UK by Adam Crymble (adam.crymble@gmail.com).
Data Creation occurred between April 2012 and July 2013.

_License_:
We release the following documents under a creative commons ÔCC-BY 4.0Õ license:
* Readme.txt (this document)
* MiddlesexVagrants1777-1786.csv (the data)

_Dataset Citation_:
Anyone publishing academically or commercially based on research conducted with this dataset in whole or in part is asked to credit the authors with the following citation:

        Adam Crymble; Louise Falcini; Tim Hitchcock, 'Vagrant Lives: 14,789 Vagrants Processed by Middlesex County, 1777-1786' (2014).

_Acknowledgements_:
These data were compiled with the financial support of The British Academy / Leverhulme Trust.
The original materials were digitised and transcribed by the 'London Lives' project:

        Tim Hitchcock, Robert Shoemaker, Sharon Howard and Jamie McLaughlin, et al., London Lives, 1690-1800 (www.londonlives.org, version 1.1, 24 April 2012).

These documents are part of the 'Middlesex Sessions' papers, held at the London Metropolitan Archives.

_Project Description_:

This dataset makes accessible the uniquely comprehensive records of vagrant removal from, through, and back to Middlesex, encompassing the details of some 14,789 men and women remove

    Tim Hitchcock, Adam Crymble, and Louise Falcini, ÔLoose, Idle and Disorderly: Vagrant Removal in Late Eighteenth-Century MiddlesexÕ, _Social History_.

Each record includes details on the name of the vagrant, his or her parish of legal settlement, where they were picked up by the vagrant contractor, where they were dropped off, as w

Each entry has 29 columns of data, all of which are described in full below.

The original records were created by Henry Adams, the vagrant contractor of Middlesex who had - as had his father before him - conveyed vagrants from Middlesex gaols to the edge of t

Spellings have been interpreted and standardized when possible. Georeferences have been added when they could be identified. This dataset was created for 21st century historians, and

_Description of Data Columns_

__Vagrant ID Number__
    Each vagrant was given a unique ID number in the format X.Y.Z where X is a sequential number starting at one and incrementing with each new group of vagrants traveling together (

__Given Names__
    This column is an interpretation of the given (first) names of vagrants. Where short forms appeared in the original, these were expanded to their logical full length when this wa
```

Data Files

“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.”

http://www.ddialliance.org/training/getting-started-new-content/create-a-codebook

Example



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

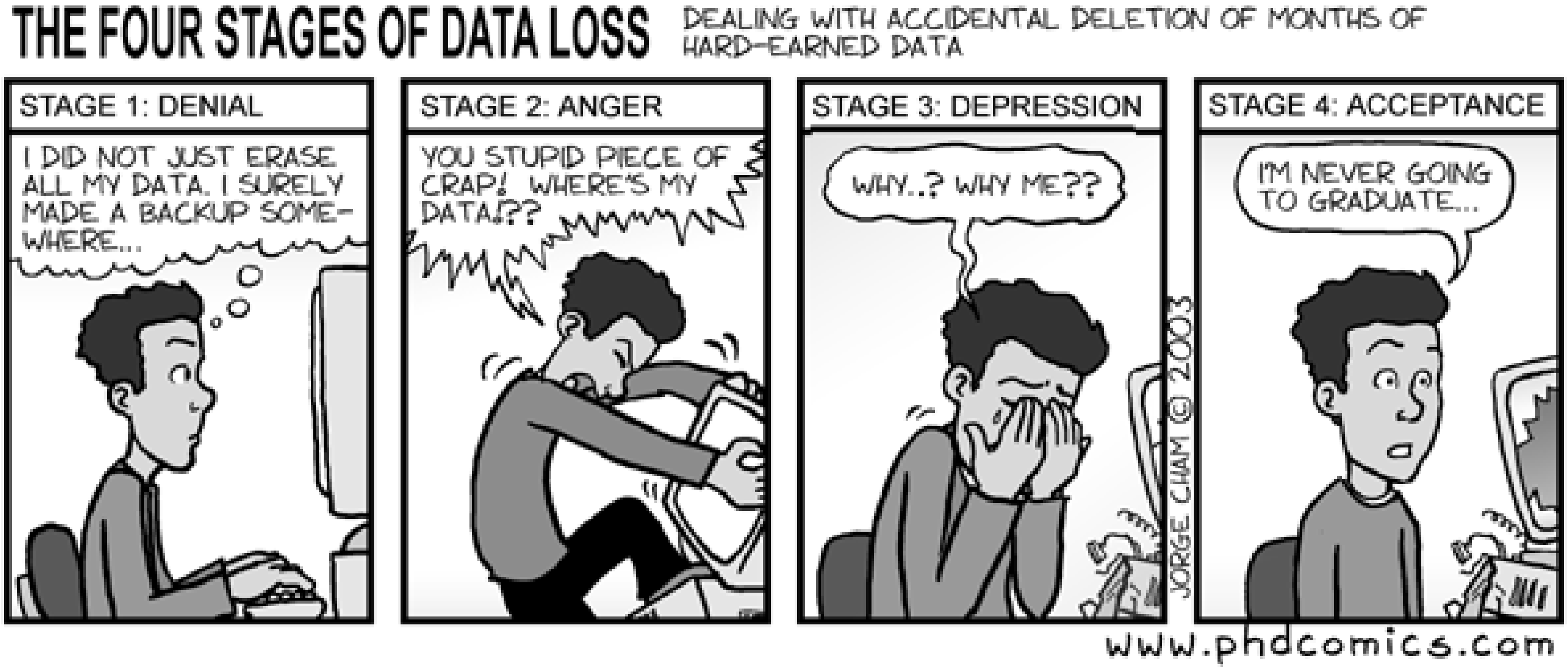
2017 CIRP Freshman Survey (Codebook)

#	Variable Name	Variable Descripion
	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

https://ucla.app.box.com/v/TFS-Codebook

Store, Share, and Archive Data

Perils of Poor Data Storage


THE FOUR STAGES OF DATA LOSS
DEALING WITH ACCIDENTAL DELETION OF MONTHS OF HARD-EARNED DATA
STAGE 1: DENIAL
I DID NOT JUST ERASE ALL MY DATA. I SURELY MADE A BACKUP SOMEWHERE...
STAGE 2: ANGER
YOU STUPID PIECE OF CRAP! WHERE'S MY DATA!??
STAGE 3: DEPRESSION
WHY..? WHY ME??
STAGE 4: ACCEPTANCE
I'M NEVER GOING TO GRADUATE...
JORGE CHAM © 2003
www.phdcomics.com

http://phdcomics.com/comics/archive.php?comicid=382

3-2-1 Backup Rule

3

Save 3 copies of your data.

2

Use 2 types of storage.

1

Keep 1 remote copy.

Data Storage, Backup, and Sharing: Rice Options

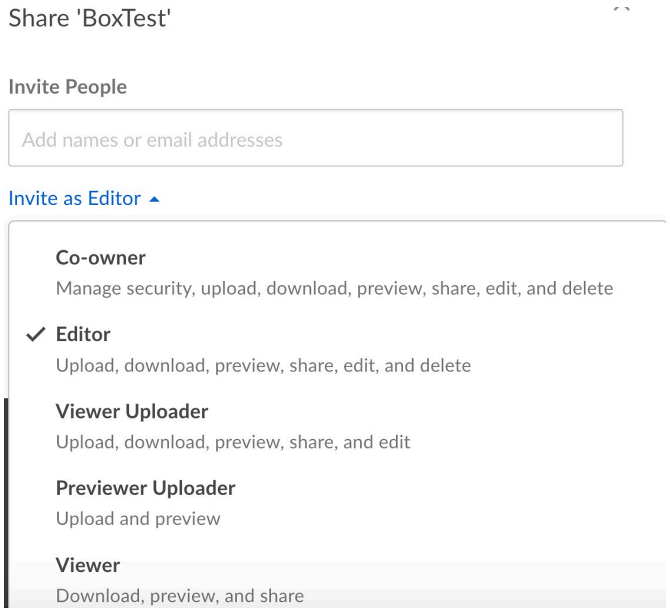
- Network/Cloud Storage
  - Rice Box
    - 2 TB limit (faculty, staff, research)
    - 1 TB limit (students)
  - Google Drive
    - 50 GB limit (faculty, staff, research)
    - 10 GB limit (students)
  - Research Data Facility (RDF): larger scale research storage

Data Storage, Backup, and Sharing (cont.)

- Backup options
  - Crashplan for Rice workstations
- Data sharing
  - Globus DTN
- Additional information:
  - Storage, file sharing, and backup: https://kb.rice.edu/70762
  - Storage options for students: https://kb.rice.edu/65636

Features of Rice Box

- Box is an “enterprise cloud-based storage and collaboration service.”
- Access prior versions (up to 100)
- Sync files and download for offline use
- Files automatically backed up at multiple data centers
- Control file/folder permissions



Share 'BoxTest'

Invite People

Add names or email addresses

Invite as Editor ▴

Co-owner
Manage security, upload, download, preview, share, edit, and delete

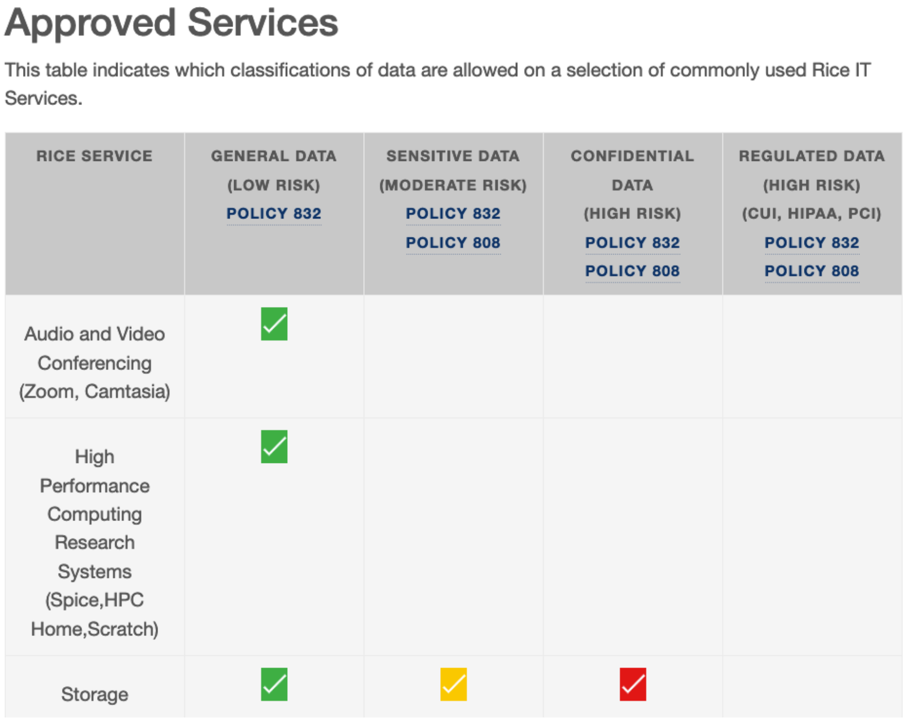
✓ **Editor**
Upload, download, preview, share, edit, and delete

Viewer Uploader
Upload, download, preview, share, and edit

Previewer Uploader
Upload and preview

Viewer
Download, preview, and share

Consult IT Regarding Data Security



Approved Services

This table indicates which classifications of data are allowed on a selection of commonly used Rice IT Services.

RICE SERVICE	GENERAL DATA (LOW RISK) POLICY 832	SENSITIVE DATA (MODERATE RISK) POLICY 832 POLICY 808	CONFIDENTIAL DATA (HIGH RISK) POLICY 832 POLICY 808	REGULATED DATA (HIGH RISK) (CUI, HIPAA, PCI) POLICY 832 POLICY 808
Audio and Video Conferencing (Zoom, Camtasia)	✓			
High Performance Computing Research Systems (Spice,HPC Home,Scratch)	✓			
Storage	✓	✓	✓	

https://iso.rice.edu/approved-services

Data Archiving Options

Public Repositories:

- Discipline based repository (e.g. GenBank or PANGEA)
- General data repository (e.g. FigShare or Dataverse)
- Institutional repository (e.g. Rice Digital Scholarship Archive)

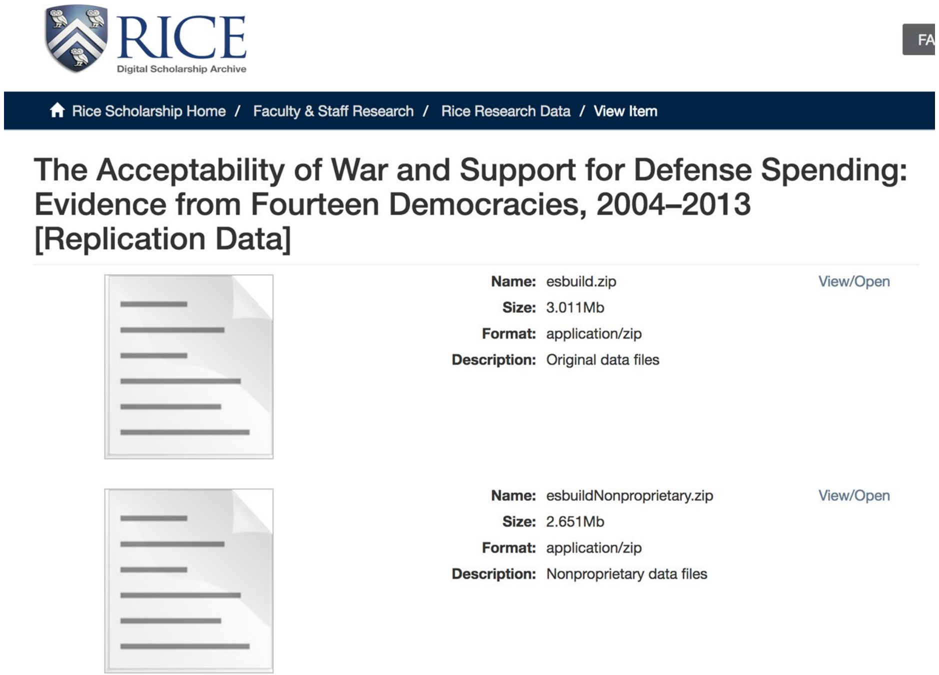
Private Approaches:

- Long-term storage

Repository?

- Conform to publisher or funder requirements.
- Get cited:
  - “studies that made [gene expression microarray] data available in a public repository received 9% … more citations than similar studies for which the data was not made available.” (Piowowar & Vision, 2013)
- Promote future research by making data available publicly for the long term.

Rice Data Sharing Option: Rice Digital Scholarship Archive



RICE
Digital Scholarship Archive
FA
Rice Scholarship Home / Faculty & Staff Research / Rice Research Data / View Item

The Acceptability of War and Support for Defense Spending: Evidence from Fourteen Democracies, 2004–2013 [Replication Data]

Name: esbuild.zip View/Open
Size: 3.011Mb
Format: application/zip
Description: Original data files

Name: esbuildNonproprietary.zip View/Open
Size: 2.651Mb
Format: application/zip
Description: Nonproprietary data files

RDSA

Data Archiving Caveats

- Do not share confidential data (unless it has been completely de-identified and approved through IRB).
- Consult with your collaborators before publishing data.
- It may be possible to embargo data so that it is not available until the related publication is released.

Fondren’s Research Data Services

- Consulting on finding, managing, analyzing, visualizing, and sharing data
- Publishing and preserving through the Rice Digital Scholarship Archive
- Providing DOIs
- Reviewing data management plans
- Workshops on R, Python, SQL, etc.

Thanks!

- Please contact researchdata@rice.edu with any questions.
- Visit us online at https://library.rice.edu/research-data-services.
- Help us shape future workshops! Please complete this evaluation:
    - https://tinyurl.com/FondrenEval

Resources

Borer, Elizabeth T., et al. "Some Simple Guidelines for Effective Data Management." *Bulletin of the Ecological Society of America* (2009): 205–14.

Cornell University Research Data Management Service Group. (n.d.) Readme template.

Dataverse, *Data Management Plans*, https://dataverse.org/best-practices/data-management/

ICPSR *Guide to Social Science Data Preparation and Archiving,* http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/

Juul, Svend et al. "Take good care of your data," http://www.epidata.dk/downloads/takecare.pdf

UK Data Archive, *Managing and Sharing Data: Best Practices for Researchers*, http://www.data-archive.ac.uk/media/2894/managingsharing.pdf