The Research Data Management Workbook

Kristin Briney

2024-06-18


CALTECH
LIBRARY
THE
RESEARCH DATA
MANAGEMENT
WORKBOOK
KRISTIN BRINEY

Contents

About this Book

Description

The Research Data Management Workbook is made up of a collection of exercises for researchers to improve their data management. The Workbook contains exercises across the data lifecycle, though the range of activities is not comprehensive. Instead, exercises focus on discrete practices within data management that are structured and can be reproduced by any researcher.

The book is divided into chapters, loosely by phases of the data lifecycle, with one or more exercises in each chapter. Every exercise comes with a description of its value within data management, instructions on how to do the exercise, original source of the exercise (when applicable), and the exercise itself.

The Workbook is intended as a supplement to existing data management education. If you would like to learn more about the principles of data management, please see the article "Foundational Practices of Research Data Management" [Briney et al., 2020] or read the book "Data Management for Researchers" [Briney, 2015].

Edition

This is edition 1.0 of The Research Data Management Workbook.

ISBN (PDF): 978-1-60049-015-6

ISBN (EPUB): 978-1-60049-014-9

I'm happy to receive any feedback on the Workbook to improve it for the next edition; you can message me at briney@caltech.edu.

License

- Briney, K. (2023). *The Research Data Management Workbook.* Caltech Library. https://doi.org/10.7907/z6czh-7zx60

The Author

Figure 1: Headshot of author, Kristin Briney. Image is of a smiling white woman with chin-length brown hair.

Kristin Briney is the Biology & Biological Engineering Librarian at the California Institute of Technology and author of the books "Data Management for Researchers" [Briney, 2015] and, with Becky Yoose, "Managing Data for Patron Privacy" [Briney and Yoose, 2022]. She has a PhD in chemistry and an MLIS, both from the University of Wisconsin-Madison. Her research focuses on research data management, institutional data policy, and patron privacy vis-a-vis library data handling. Kristin is an advocate for the adoption of the international date standard ISO 8601 (YYYY-MM-DD) and likes to spend her free time making data visualizations out of yarn and fabric.

Chapter 1

Introduction

1.1 What is Research Data Management?

Research data management is a set of collective practices and decisions that make it easier for you, your collaborators, and your future self to find, understand, and use your research data. These practices cover the entire lifecycle of research data, from its collection and analysis through sharing and reuse. There is no one magical data management practice to rule them all. Rather, data management consists of a number of small activities that make dealing with your data a better experience. Research is hard enough as it is without having to fight with your files, so the goal of data management is for you to maximize your time doing research instead of spending extra time with file handling.

1.2 Why do Research Data Management?

Most researchers have spent time, at some point in their careers, digging through their computer to find a specific file that can't be located. It's incredibly frustrating and a waste of time and resources, especially if you end up recollecting missing data. The good news is that it is possible to avoid this situation entirely by strategically managing your data better.

Done well, research data management means:

- always understanding what your data is and how you collected it even if the data is a year old
- always finding the file you need quickly
- never losing your data even if your hard drive crashes
- knowing what rights and responsibilities you have over your data

- knowing how and where to share your data to comply with your funder's data sharing policy
- being able to pick up and easily reuse data from a past project

If all of those things sound like something you would like to implement in your research, you are reading the right book!

1.3 How to Use this Book to Manage Your Research Data Better

The Research Data Management Workbook is focused entirely on the "how" of data management. The Workbook consists of a series of worksheets, checklists, and procedures to set up new data management practices, check existing practices, and make good decisions about your data. You will find exercises covering all the ideals listed in the "Why Do Research Data Management?" section above, allowing you to streamline your use of research data.

The Workbook is not a complete set of exercises for everything under the umbrella of data management. Instead, the Workbook centers on activities that are structured, reproducible, and apply to many researchers. The strength and weakness of data management is that many of its practices are customizable to individual research workflows. Each exercise in the Workbook, therefore, is built on best principles while allowing for customization to suit local needs.

You may go through exercises in the Workbook collectively or individually as you chose. Do note that some exercises require completing one or more other exercises in the book, so it's best to have the whole workbook on hand just in case. Finally, for the exercises that have been formatted as worksheets, I recommend printing them out and writing your answers in the space provided.

1.4 Further Resources on Research Data Management

"The Research Data Management Workbook" does not comprehensively explain data management and therefore works best for those with some foundational data management knowledge or in tandem with other educational resources on research data management.

My best recommendation is to use the Workbook as the exercise book for my first book, "Data Management for Researchers" [Briney, 2015]. The following table lays out how chapters in the Workbook match with chapters in "Data Management for Researchers," allowing you to look up more information on any topic covered by the Workbook:

Research Data Management Workbook	Data Management for Researchers
Chapter 2	Chapter 4
Chapter 3	Chapter 5
Chapter 4	Chapter 8
Chapter 5	Chapter 3
Chapter 6	Chapter 10
Chapter 7	Chapter 9

Chapter 2

Documentation

Documentation has sometimes been called “a love letter to your future self” as it helps you remember important details about your research data. The great thing about research documentation is that it’s not limited to a laboratory or research notebook, though notebooks are still very important! This chapter introduces two types of useful documentation – a project-level README.txt and a data dictionary – and offers worksheets for writing both. The chapter also includes a worksheet to evaluate an older entry in your laboratory notebook to ensure your documentation is of sufficient quality.

2.1 Evaluate a Laboratory Notebook

Description: *The laboratory or research notebook is a fundamental documentation method for many researchers. But for how ubiquitous the lab notebook is, documentation can sometimes be lacking. The ideal laboratory notebook allows someone with similar training as you to be able to follow everything you did in your research. This exercise prompts you to review an old entry within your laboratory notebook to evaluate if your documentation is sufficient for reproducing your work.*

Instructions: *You will need a laboratory notebook entry from 6-12 months ago to do this exercise. Once you have the entry, read through it to try to understand what you did on that day. Answer the exercise questions to evaluate the entry and identify any note keeping improvements to make.*

Date of lab notebook entry being evaluated: _____

Read the entry and summarize the work you did on that date:

How easy was it to understand the work you did from your notes?

Could you reproduce your work based on the information in your notes? If not, what extra information do you need?

What worked well with your note keeping?

What should you improve about your note keeping?

List one change you will implement to take better research notes:

2.2 Write a Project-Level README.txt

Description: *Data files living on a computer often need extra documentation for someone to understand what research they correspond to. In particular, it is useful to record the most basic project information and store it in the top-level folder of each research project. This can be done with a README.txt. The name, "README", indicates that the file conveys important information and the file type, TXT, can be opened by many different software programs, making the content maximally accessible. This exercise walks you through the key information needed in a project-level README.txt file. The same information can also be recorded at the front of a physical laboratory notebook.*

Instructions: *Pick a research project and answer the following questions. Copy all of the text into a TXT file and save it with the name "README.txt". Store this file in the top-level of the project folder on your computer, alongside the project files.*

Source: *This exercise was adapted from the "Project Close Out Checklist" [Briney, 2020b].*

Project documented by this README.txt: _____

Write a brief project description:

Example: The Data Doubles project was a 4-year, IMLS-funded research project examining student perceptions of privacy in library learning analytics.

What is the time period the project was done over?

Example: The project was conducted between summer 2018 and summer 2022.

Who worked on the project?

Example: Eight researchers from eight different instutions worked on the project, including: KMLJ, AA, KB, AG, MP, MR, DS, and MS.

Where are the data, code, and other files are stored?

Example: Research files are stored on Google Drive, with the exception that participant data is stored in IU-hosted Box. Survey data is also in Qualtrics. Code is on GitHub. The shared literature library is in Zotero.

How and where is the project documented?

Example: Documentation for the project is in Google Drive. Notes on team decisions from meetings are in the DataDoubles/Meetings folder. Notes on data are in the DataDoubles/Research folder.

How are files organized? Are any naming conventions used and, if so, what are they (see Chapter 3)?

Example: All data is in the DataDoubles/Data folder, with subfolders labelled by interview theme code. Each site has its own folder within the project folder for individual site files. Interview data files are named with: interview theme, site, interview ID, interview date, and data type/analysis stage (e.g. "PRO_BL03_20180222_Audio.mp3" and "AWA_MK01_20180222_Notes.pdf"). Please see the living data management plan for complete set of codes and more details.

What else does someone need to know to understand these files?

Example: Additional documentation on the project and public research files are available on OSF.

2.3 Create a Data Dictionary

Description: *Ideally, a spreadsheet is formatted with a row of variable names at the top, followed by rows of data going down. This makes easy for data to be used in any data analysis software (interoperability is a good thing) but makes it impossible to document a spreadsheet within the file itself. For this reason, it's useful to create a data dictionary to describe the spreadsheet so that others can interpret the data. This exercise walks you through the major information you should record for each variable in the spreadsheet, adding up to a complete dictionary to accompany the spreadsheet file.*

Instructions: *Pick one spreadsheet variable and record its information in the corresponding rows of the table. Repeat this process for the remaining variables in the spreadsheet. Copy all information into a text document and save it next to the spreadsheet. It is useful to save the data dictionary with the same root name as its data file by appending "_dictionary" on the end of the file name; for example, the data dictionary for the file "myData.xlsx" would be "myData_dictionary.txt".*

Source: *This exercise was adapted from "Leveling Up Data Management" [Briney, 2023].*

Question	Example
Variable name	*site*
Variable description	*Two-letter abbreviation describing the name of the overall site where the sample was collected.*
Variable units	*N/A*
Relationship to other variables	*Partner to variable "sampleNum", which together define the sample ID (site name + sample number at that site). Related to variables "latitude" and "longitude", which record exact coordinate location and are more specific than the larger site code.*
Variable coding values and meanings	*Coding values and meanings: BL = Badlands NP; DV = Death Valley NP; GT = Grand Teton NP; JT = Joshua Tree NP; ZN = Zion NP*
Known issues with the data	*Some Badlands samples were collected outsideof the park boundaries; see latitude and longitude variables for specific locations.*

Question	Example
Anything else to know about the data?	*Older data (pre-2013) used one-letter abbreviations for site code but this was updated for clarity and ease of identification.*

Question	Variable
Variable name	
Variable description	
Variable units	
Relationship to other variables	
Variable coding values and meanings	
Known issues with the data	
Anything else to know about the data?	

Chapter 3

File Organization and Naming

Good file organization and naming are foundational data management practices, as they help you find files quickly when you need them. To set up file organization and naming conventions, this chapter offers two exercises: a card-sorting process for brainstorming a file organization system; and a worksheet for creating a file naming convention for a group of files.

3.1 Set Up a File Organization System

Description: *Implementing a file organization system is the first step toward creating order for your research data. Well-organized files make it easier to find the data you need without spending lots of time searching your computer. Every researcher organizes their files slightly differently, but the actual organizational system is less important that having a place where all of your files should logically go. This exercise prompts you to brainstorm organizational groupings and hierarchies to come up with an order for managing your research data.*

Instructions: *This is a card-sorting exercise, meaning you will need a stack of note cards or post-it notes to do this activity, ideally in three different colors. Follow the instructions to label cards and move them around until you develop your organizational system. There is no one correct way to do this so feel free to play around, add new cards, and move cards however you want! Once you put your new organizational system into place, be sure to always put your files where they're supposed to go.*

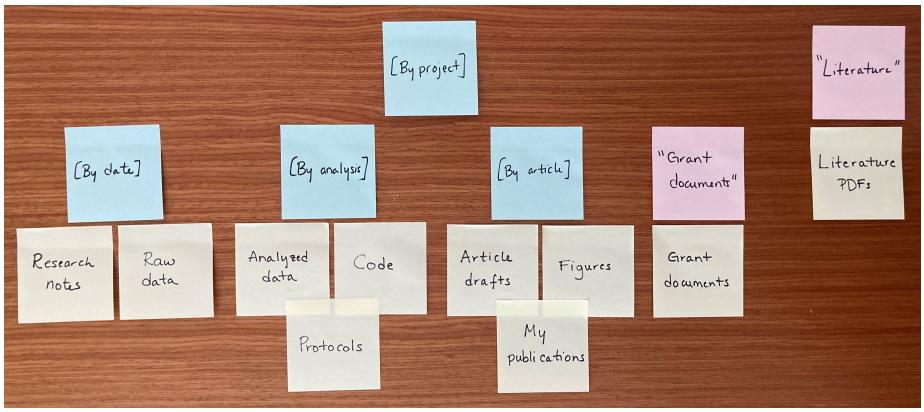

Figure 3.1: Example of organizational exercise. Image is of blue (group folder), pink (single folder), and yellow (data type) sticky notes, each with its own label, organized hierarchically on a wooden desktop.

---

1. Take a stack of note cards or post-it notes in the first color and write the following labels on one card each, omitting any file types that you do not use in your research:
    - Raw data
    - Analyzed data
    - Code
    - Protocols

- Article drafts
- Figures
- My publications
- Literature PDFs
- Grant documents
- Research notes

2. Move cards around and group together file-type cards that you want to store together. Files that will be stored near each other in a folder hierarchy, but not together, should be placed near each other while file types expected to be stored completely separately should be away from other cards.
3. Create hierarchies in file organization by adding new "folder" cards in one of two types; use different colored cards for each folder type:
    - Cards in the second color represent a single folder. These should be labeled with the folder name in quotations (e.g. "Literature" or "My publications").
    - Cards in the third color represent a group of folders, such as for folders organized by date or by project. Use only one card to represent the organizational pattern that will be repeated. These cards should be labeled with the organizational system in square brackets (e.g. [By date] or [By project]). Note: folders organized by date should, in real life, be labelled using the convention YYYYMMDD or YYYY-MM-DD to facilitate chronological sorting.
4. Move existing file-type cards/groups of file-type cards underneath the new folder cards to show the hierarchy of how a file type will be saved in a specific folder or group of folders. Organizational-group folders (cards in the third color) only need to be represented once in the card sorting, as they are assumed to represent multiple folders on a computer.
5. Make copies of any type of card and add folder levels, as needed. Adjust placement and hierarchies until you are happy with the organizational system you developed.
6. Record your organizational system in your lab notebook and/or a README.txt.

3.2 Create a File Naming Convention

Description: *File naming conventions are a simple way to add order to your files and help to find them later. Rich and descriptive file names make it easier to search for files, understand at a glance what they contain, and tell related files apart. This exercise guides researchers through the process of creating a file naming convention for a group of related files.*

Instructions: *Fill in each section for a group of related files following the instructions; an example for microscopy files is provided. This exercise may be redone as needed, as different groups of files require different naming conventions.*

Source: *This exercise is based on the "File Naming Convention Worksheet" [Briney, 2020a].*

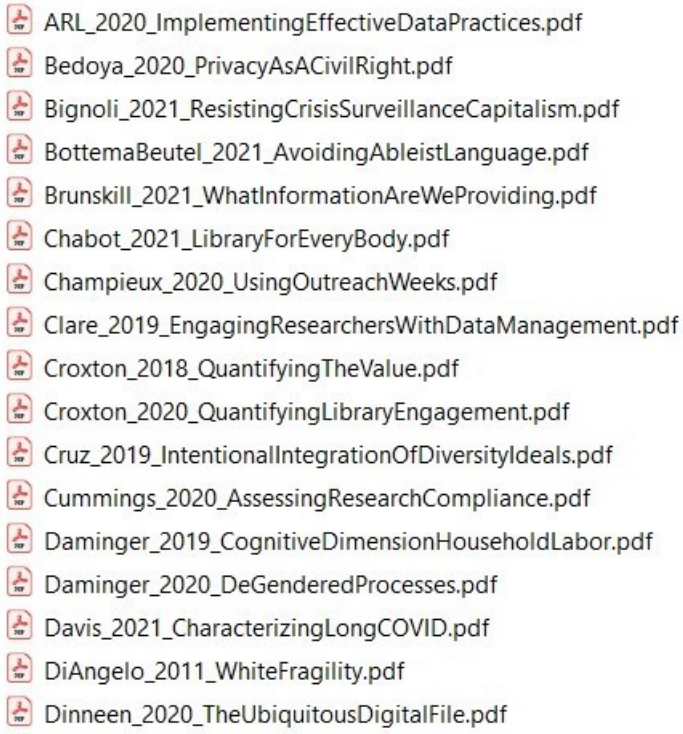

Figure 3.2: Screenshot of article pdf's in a file system with consistent file names. File names use the convention FirstAuthorLastName_YEAR_ShortTitle.pdf

---

1. What group of files will this naming convention cover?

You can use different conventions for different file sets.

Example: This convention will apply to all of my microscopy files, from raw image through processed image.

2. What information (metadata) is important about these files and makes each file distinct?

Ideally, pick three pieces of metadata; use no more than five. This metadata should be enough for you to visually scan the file names and easily understand what's in each one.

Example: For my images, I want to know date, sample ID, and image number for that sample on that date.

1.
2.
3.
4.
5.

3. Do you need to abbreviate any of the metadata or encode it?

If any of the metadata from step 2 is described by lots of text, decide what shortened information to keep. If any of the metadata from step 2 has regular categories, standardize the categories and/or replace them with 2- or 3-letter codes; be sure to document these codes.

Example: Sample ID will use a code made up of: a 2-letter project abbreviation (project 1 = P1, project 2 = P2); a 3-letter species abbreviation (mouse = "MUS", fruit fly = "DRS"); and 3-digit sample ID (assigned in my notebook).

4. What is the order for the metadata in the file name?

Think about how you want to sort and search for your files to decide what metadata should appear at the beginning of the file name. If date is important, use ISO 8601-formatted dates (YYYYMMDD or YYYY-MM-DD) at the beginning of the file names so dates sort chronologically.

Example: My sample ID is most important so I will list it first, followed by date, then image number.

1.
2.
3.
4.
5.

5. What characters will you use to separate each piece of metadata in the file name?

Many computer systems cannot handle spaces in file names. To make file names both computer- and human-readable, use dashes (-), underscores (_), and/or capitalize the first letter of each word in the file names. A good convention is to use underscores to separate unrelated pieces of metadata and dashes to separate related pieces of metadata for parsing and readability.

Example: I will use underscores to separate metadata and dashes between parts of my sample ID.

6. Will you need to track different versions of each file?

You can track versions of a file by appending version information to end of the file name. Consider using a version number (e.g. "v01") or the version date (use ISO 8601 format: YYYYMMDD or YYYY-MM-DD).

Example: As each image goes through my analysis workflow, I will append the version type to the end of the file name (e.g. "_raw", "_processed", and "_composite").

7. Write down your naming convention pattern.

Make sure the convention only uses alphanumeric characters, dashes, and underscores. Ideally, file names will be 32 characters or less.

Example: My file naming convention is "SA-MPL-EID_YYYYMMDD_###_status.tif" Examples are "P1-MUS-023_20200229_051_raw.tif" and "P2-DRS-285_20191031_062_composite.tif".

8. Document this convention in a README.txt (or save this worksheet) and keep it with your files.

Chapter 4

Data Storage

All research data needs to be stored and backed up, but it can be frustrating to pick these systems and ensure that they are working correctly. This chapter consists of two exercises: a worksheet to document available storage and backup options and decide between them; and a procedure for testing that a backup system is working.

4.1 Pick Storage and Backup Systems

Description: *Research data needs to be stored and backed up reliably so that important data is not lost. But storage is commonly a challenge, as institutions don't always offer uniform options for storage and backup. This exercise prompts you to examine the storage and backup systems available to you before determining which is the best set of options for your data.*

Instructions: *Answer the questions and then fill out the table of information about each possible storage and backup systems. Examine all of the options, evaluating them based on the criteria listed below. Then select primary storage and backup systems and, optionally, an alternate backup.*

What is the estimated total data storage you will need over the next five years?

Example: I estimate that I will generate 100 GB of data over the next five years of my project.

Does your data require meeting any specific security standards? If so, what level of security?

Example: My data will include some human subjects data, so my storage systems must have restrictions on access but it's not medical data so they don't have to be HIPAA compliant.

What storage and backup systems are available to you, such as through your institution, workplace, or elsewhere?

Example: I have the following systems available to me: my computer, a Time Machine backup, a departmental server, institution-licensed Box account, and Google Drive.

Fill out the information in the table for *each* storage and backup system you are considering:

Question	Example
System name	*Departmental server*
Is it storage or backup?	*Storage*
What is the cost?	*No cost for 10GB and under. Cost is $5 per 10 GB per year after that.*
What is the hardware type?	*Server, exact hardware type unknown.*
Is the system backed up?	*No backup.*
For backup systems, is backup automatic?	*N/A*
What level of security does the system provide?	*Storage is password protected.*
Is the system local or remote?	*System is local.*
Is there a limit to storage capacity?	*Storage limit is 500GB per research group.*
Who manages the system?	*Departmental IT manages the server.*
Is it easy or difficult to use?	*Very easy to use once set up.*

Question	System
System name	
Is it storage or backup?	
What is the cost?	
What is the hardware type?	
Is the system backed up?	
For backup systems, is backup automatic?	
What level of security does the system provide?	

Question	System
Is the system local or remote?	
Is there a limit to storage capacity?	
Who manages the system?	
Is it easy or difficult to use?	

Optimize your storage and backups on the following considerations:

1. You need a primary storage system that:
   - will hold all of your data files,
   - meets your needed level of security.
2. You need one backup that:
   - will hold all of your data files,
   - meets your needed level of security,
   - is reliable/managed by someone you trust,
   - is easy to use,
   - backs up automatically.
3. At least one backup should be in a different location than your main storage system for disaster resiliency. If your main backup is nearby your primary storage and/or if your primary storage system is not reliable, you need a second backup that:
   - will hold all of your files,
   - meets your needed level of security,
   - is reliable/managed by someone you trust.

Pick your storage and backup systems:

Example: My primary storage will be my computer with added security restrictions. I will use Time Machine as my first automatic backup and institutional Box, which is controlled access, as my second backup because it is remote.

4.2 Test Your Backup

Description: *Backups are super important for your data, so it's always good to test that your backups are still working. Nothing is worse than losing your data from your primary storage and then realizing that your backup isn't working either. Beyond checking that your backup is working, it's also good to know how to recover your files so that you don't have to learn this for the first time while panicking about lost data. This short exercise walks you through getting a file off your backup to test that it is working and to learn how the data-recovery process works.*

Instructions: *Pick a backup system and a file to recover and work through the steps. The hard part of this exercise is finding instructions for file recovery and recovering the file, which vary by backup system.*

---

1. Identify where your data is backed up.
2. Find instructions for recovering data from your backup system.
3. Pick a data file from your computer.
4. Follow the instructions from step 2 to get a copy of the data file from step 3 out of your backup system.
5. If this process didn't work, fix your backup system. If this process did work, congrats your backup is working and you know how to recover your files!

Chapter 5

Data Management

While this entire workbook covers data management activities, it's often useful to take a step back and document the data management decisions that have been made. This chapter provides exercises in documenting data management in two areas: a worksheet for writing a living data management plan (which builds on exercises from previous chapters); and a worksheet for discussing roles and responsibilities around data management with your research collaborators.

5.1 Write a Living Data Management Plan (DMP)

Description: *Many researchers are aware of the two-page data management plan (DMP) for a grant application, but you may not be aware of the more useful type of DMP: a living DMP. This document describes how data will be actively managed during a project and may be updated whenever necessary to reflect current data practices. A living DMP is a useful touchstone for understanding where data lives, how it's labelled, how it moves through the research process, and who will oversee the data management. This exercise guides you through the process of creating a living DMP for your research.*

Instructions: *Pick a project and answer the following questions to build your living DMP. This DMP may be changed at any time to improve practices. If you are doing collaborative research, work through this exercise with your collaborators to agree on shared conventions.*

Write a short summary of the project this DMP is for:

Example: This project uses mass spectrometry to identify isotopic composition of soil samples.

Where will data be stored? How will data be backed up? (See Exercise 4.1: Pick Storage and Backup Systems.)

Example: The data is generated on the mass spectrometer then copied to a shared lab server. The server is backed up by departmental IT.

How will you document your research? Where will your research notes be stored?

Example: Data collection and analysis is primarily documented in a laboratory notebook, organized by date. README.txt files add documentation to the digital files as needed.

How will your data be organized? (See Exercise 3.1: Set Up a File Organization System.)

Example: Each researcher has their own folder on the shared server. Data within my folder is organized in folders by sample site with subfolders labeled by sample ID. Sample ID consists of: two-letter sample site code, three-digit sample number, and date of sample collection formatted as YYYYMMDD (e.g. "MA006-20230901" and "CB012-20100512").

What naming convention(s) will you use for your data? (See Exercise 3.2: Create a File Naming Convention.)

Example: Files will be named with the sample ID, type of measurement, and stage in the analysis process; these pieces of information will be separated by underscores. Examples: "MA006-20230901_TIMS_raw" and "CB012-20100512_SIMS_analyzed".

Do you need to do any version control on your files? How will that be done?

Example: Version control will be very simple through file naming, appending analysis information onto the end of file names to keep track of which version of the file it is.

How will data move through the collection and analysis pipelines?

Example: Once data is collected on the mass spectrometer, I will copy it to the correct folder on the shared server for analysis. Data will stay in its sample ID-labeled folder as it gets analyzed, with different file names to annotate analysis stage. Data that will be published will be copied into separate folders, organized by article.

Record any project roles and responsibilities around data management:

Example: It is each researcher's responsibility to ensure that data moves through the analysis pipeline and is labeled correctly. The lab manager will ensure that the shared server stays organized and will periodically check that backups are working.

Record any other details on how data will be managed:

Example: Copies of this DMP will live in my top-level folder on the lab server so that others can find and use my data as needed.

5.2 Determine Data Stewardship

Description: *It is often helpful to be up front about requirements and permissions around research data. This exercise encourages you to discuss these issues with supervisors and peers to make sure that there are no misunderstandings about who has what rights to use, retain, and share data.*

Instructions: *Determine which research data should be discussed. Bring together the Principle Investigator, the researcher collecting the data, and anyone else who works with that data. As a group, answer the questions in the exercise, making sure that everyone agrees on the final decisions. Record the results of the discussion and save them with the project files.*

Source: *This exercise was adapted from the "Project Close Out Checklist" [Briney, 2020b].*

Who is participating in this discussion?

Example: This discussion includes the graduate student who collected the data, the project Principle Investigator (PI), and the laboratory manager.

What data is being discussed?

Example: This discussion covers all of the data collected by the graduate student during their time at the university.

Are there security or privacy restrictions on the data and, if so, what are they?

Example: Some of the research data includes human subjects data. This data must be held securely with limited data sharing, as outlined in the IRB protocols.

Are there intellectual property limitations on the data and, if so, what are they?

Example: There are no intellectual property concerns for the data.

Are there any requirements to publicly share the data and, if so, what are they?

Example: This research was funded by the NIH, which requires data sharing. The laboratory plans to share all data reproducing published results with the exception of the human subjects data.

Who will store the copy of record of the data and for how long?

Example: The project PI will retain the copy of record of the data for at least 3 years after the end of the grant award, with an ideal 10 year retention period.

Who is allowed to keep a copy of the data after the project ends? Which data?

Example: The graduate student may keep a copy of all data except the human subjects data after they leave the university.

Who is allowed to reuse the data after the project ends? Which data? Are there any requirements for reuse, such as co-authorship?

Example: The graduate student may reuse and publish with the data collected during their time at the university but must offer co-authorship of any papers using the data to the project PI and any relevant lab members.

Who keeps any physical research notebooks after the project ends?

Example: The PI will keep all physical laboratory notebooks but the graduate student may make copies to retain for their personal records.

Chapter 6

Data Sharing

Sharing data that underlies research has become a common expectation within scholarly research. However, the landscape of data repositories is still uneven and many researchers are still learning best practices for data sharing. To help in this area, this chapter offers of two exercises: a decision tree-inspired worksheet for picking the best data repository for your data; and checklist for working through the process of sharing data in a data repository.

6.1 Pick a Data Repository

Description: *It can be difficult to know where to share research data as so many sharing platforms are available. Current guidance is to deposit data in data repository that will give you a DOI or similar permanent identifier. This exercise guides you through the process of picking a data repository, starting with repositories for very specific types of data and defaulting to generalist data repositories. Note that some repositories charge fees for deposit, most often for large data (500 GB or larger).*

Instructions: *Identify the data that needs to be shared and work through repository selection from discipline-specific data repositories to more general data repositories. Once you have identified a repository for all of your data, deposit the data and record the corresponding permanent identifiers. Note that, depending on data types, you may need to deposit your data into multiple repositories (for example, a discipline-specific repository for one type of data and an institutional data repository for the rest of the data).*

---

1. Identify all of the data that needs to be shared.

Example: My data to be shared includes: 1) genetic data for Drosophila; and 2) microscope images of flies.

2. Is there a known disciplinary data repository for some or all of the data? For example, is there a data repository used by everyone in your research area or required for your data type by your funding agency?

If so, deposit some or all of your data there. Go to step 7 if the repository will accept all of your data or go to the next question if there is still some data left to deposit.

Example: The database FlyBase is used for Drosophila genes and genomes. My genetic data will be shared there.

3. Review the list of recommended data repositories from PLOS [PLOS ONE, 2023]. Is there a logical disciplinary data repository for some or all of your data?

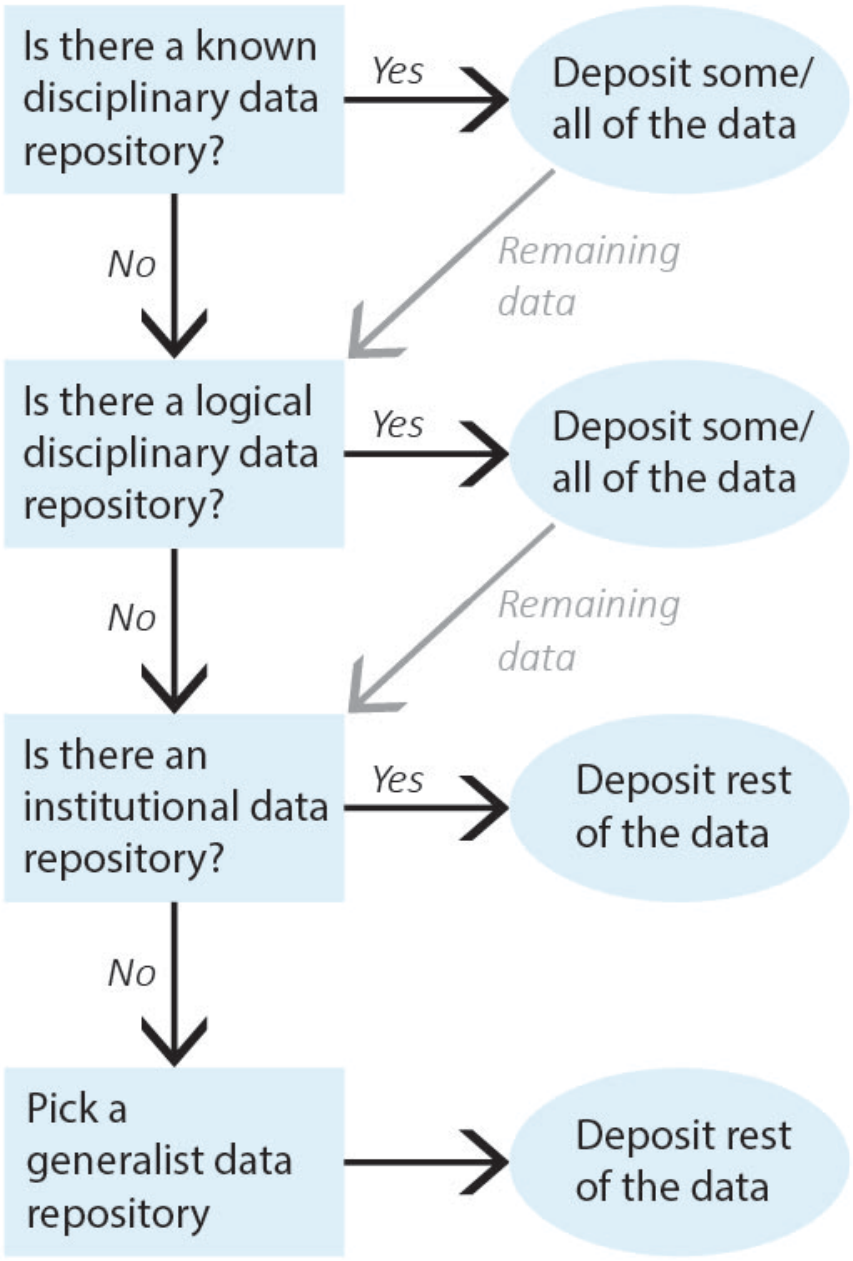

Figure 6.1: Workflow diagram upon which the exercise is based. Starting at the top, decide is there is a known disciplinary data repository (if so, deposit data), a logical disciplinary data repository (if so, deposit data), an institutional data repository (if so, deposit data), and if none of those work pick a generalist data repository.

If so, deposit some or all of your data there. Go to step 7 if you have shared all of your data or go to the next question if there is still some data left to deposit.

Example: There isn't a logical disciplinary data repository for microscope images of flies.

4. Does your institution have a data repository?

If so, deposit the remainder of your data there and go to step 7.

Example: The California Institute of Technology hosts the data repository CaltechDATA. I will deposit my microscope images in CaltechDATA.

5. Do you have a preferred generalist data repository [NIH, 2023]?

If so, deposit the remainder of your data there and go to step 7.

Example: [All data has been shared already.]

6. Pick a generalist data repository [NIH, 2023] and deposit the remainder of your data.

Deposit your data and go to step 7.

Example: [All data has been shared already.]

7. Record the permanent identifier, ideally a DOI, from each data deposit.

DOIs make data FAIR [Wilkinson et al., 2016] and aid with data sharing compliance. If you did not receive a permanent identifier (such as a DOI, permanent URL, etc.) during deposit, return to step 2 and pick a different data repository for your data.

Example: CaltechDATA provides DOIs for all deposits; my permanent identifier is doi.org/10.22002/XXXXXXXXXXX. FlyBase provides stable links to data reports using FlyBase ID numbers; my permanent identifier is flybase.org/reports/FBXXXXXXXXX.

6.2 Share Data

Description: *Data sharing is becoming common and expected by funding agencies and journals. While the process of depositing data into a data repository will vary between repositories, there are some common actions that should be taken to prepare data for sharing. This exercise walks you through these standard requirements for sharing data.*

Instructions: *This checklist enumerates the necessary steps and decisions to deposit data supporting a research article into a data repository. Identify the data to be shared and work through the list. Note that, if data will be shared as multiple deposits or in multiple repositories, the checklist should be worked separately for each group of data.*

Data Selection

___ Select the data files that reproduce published results.

___ Perform quality control on the data files.

___ Convert data in proprietary file types to open file types, as appropriate (see Exercise 7.2: Convert Data File Types).

___ Determine if data should be shared under one citation or as several citations. (Group data as makes most sense for citation and reuse. Options can include: sharing as one large group, grouping files by data type, giving large data files their own citations, etc. Each citation represents a unique deposit into a data repository.)

___ If there will be multiple deposits in one repository or data will be divided across more than one data repository, work through the remainder of the checklist separately for each citation/group of files.

Data Documentation

___ Document any spreadsheet data with a data dictionary (see Exercise 2.3: Create a Data Dictionary). The data dictionary should be shared with the other files.

___ Write a brief description of each data file, including any data dictionaries, and what it contains. Save this information as a README.txt file and share it with the other files.

Sharing Information (Metadata)

___ Give the dataset a title. Default title is "Data from: [name of the article]".

___ Write a brief description of the dataset to be used as the abstract/description.

___ Write down keywords/subject terms for the data.

___ Determine who will be listed as authors of the data and in what order; this may be different than the authors of the article.

___ Identify author ORCID numbers for submission (note: this is best practice but not all data repositories have integrated ORCID yet).

___ Record all funding information that applies to the dataset.

___ Chose a license for reuse rights. Default license is CC0 (for more information on CC0, see [Creative Commons Wiki, 2014]).

Deposit Data

___ Pick a data repository/data repositories for the shared data (see Exercise 6.1: Pick a Data Repository).

___ Deposit the data and documentation files into the data repository, and fill in metadata as determined above.

___ If you are depositing a large number of datasets, contact repository administrators about potentially using an Application Programming Interface (API) to skip manual entry of duplicate metadata.

Share Data

___ Share data with its DOI or, as applicable, other permanent identifier.

___ Link the publication to its data, either in a Data Availability Statement or as a citation.

6.3 Make Spreadsheets Accessible and Reusable

Description: *Making your data both accessible and reusable makes it easier for someone (including your future self) to use and understand your data. Slight tweaks to formatting can make a significant difference to a spreadsheet's reusability. This checklist provides guidance on making a spreadsheet reusable as well as more accessible to those with disabilities. Note that sometimes guidance for reusability and accessibility conflict, and this checklist is a best effort to balance the two considerations.*

Instructions: *For a given spreadsheet, work through the actions on this checklist to make that data more accessible and reusable. This is best done when the data is finalized and/or prior to sharing the data either publicly or with colleagues. It is recommended to also share data in its original form, but your data will be more FAIR [Wilkinson et al., 2016] when an accessible version is made available alongside the original.*

Further Resources:

- [Broman and Woo, 2018]
- [Oxford and Woodbrook, 2023]
- [Wickham, 2014]

___ Break data into several smaller rectangular tables instead of one large complex table, as necessary. Each sheet should contain only one table.

___ Arrange data so that the top row contains variable names, with data in all following rows. See Wickham's guidelines on tidy data for more information [Wickham, 2014].

___ Clean up the variable names in the first row of the spreadsheet to be both human and machine readable:

___ Use short but meaningful names;

___ Use full words or readable abbreviations (e.g. "number" or "num" instead of "n") in variable names;

___ Use only alphanumeric characters in variable names;

___ Remove spaces from variable names;

___ Capitalize the first letter of each word in the variable name, though the first word can be lower case depending on preference (e.g. myVariableName or MyVariableName).

___ Place the key, or most identifying, variable in the first column on the left, column A. (Spreadsheets should be readable from left to right then top to

bottom, and placement of the key variable in the first column will help with readability.)

___ Convert any dates to YYYY-MM-DD format. (To work around Excel's weird date formatting you can separate year, month, and day into three separate variables.)

___ Ensure that spreadsheet cells contain only one data point. If there is more than one data point per cell, divide columns into multiple variables as appropriate.

___ Remove formatting such as font, text alignment, highlighting, and merged cells. Any information represented by such formatting should be recorded as data under new variable.

___ Fill in empty cells:

___ Input any missing data values;

___ Use "NA" (or the preferred null value for your analysis software) for any cells that do not have recorded values.

___ Perform quality control on the data, removing:

___ Errors,

___ Inconsistencies,

___ Accidental spaces.

___ Enclose any cells containing commas inside of double quotes (e.g. "text, example").

___ Remove charts. Charts may be shared separately with corresponding alt text.

___ Remove underlying calculations so that the file only includes raw data. (You can do this in Excel by copying a column, using the special paste option to "paste as values", then deleting the original column.)

___ Use any built-in validation or accessibility checkers provided by your software.

___ Save data as a CSV file type (TSV is also an acceptable file format). Save individual spreadsheet tabs as separate CSV files.

___ Create an accompanying data dictionary (see Exercise 2.3: Create a Data Dictionary).

___ Share the accessible CSV file(s), the original dataset, and the data dictionary.

Chapter 7

Project Wrap Up

The end of a project is a key time to perform data management activities in order to set yourself up for future data reuse. This is because you still remember all of the important details about your data and can make good decisions about preparing it for the future. This chapter has three exercises to work through for project wrap up: a worksheet on converting data to more open file types; a checklist for populating a project Archive folder; and a checklist for preparing data for reuse, which leverages the previous two exercises.

This chapter also covers project wrap up in the form of separating from your institution. This checklist exercise for the departing researcher is important to work through so that critical data does not get lost in the transition. A fuller version of this checklist, intended for both the departing personnel and a project administrator to work through together, is also available [Goben and Briney, 2023].

7.1 Prepare Data for Future Use

Description: *The end of a project is a good time to prepare data for potential future reuse, as you still know the important details about the data to record and have access to any software used to create the data. This checklist exercise walks you through steps to gather your data into a central place and document the project. Working through the checklist results in project data being in one central location, well documented, and organized and formatted in a way to make future reuse easier.*

Instructions: *Gather all of the data from a project and work through the checklist to organize and document the data for future reuse. This exercise refers to several other exercises in the Workbook that should be completed during this process, if they have not been already*

Source: *This exercise was adapted from the "Project Close Out Checklist" [Briney, 2020b].*

Prepare Data

___ Move all data into one central project folder; this folder may have sub-folders and should be organized however makes sense for your data.

___ As necessary, work through Exercise 7.2: Convert Data File Types to copy data into more open/common file formats.

Back Up Your Research Notes

___ If your notes are electronic, save a copy in the project folder

___ If your notes are physical, scan them and save a copy in the project folder.

Create a Project Archive Folder

___ Work through Exercise 7.3: Create an Archive Folder.

___ Put the Archive folder in the project folder.

Create a Project-Level README File

___ If you haven't done so already, work through Exercise 2.2: Project-Level README.txt.

___ Store a copy of the README file with the data.

Save Files in a Stable Location

___ Save the project folder on a storage system that you will have access to for the next several years.

7.2 Convert Data File Types

Description: *Data is often stored in a file type that can only be opened by specific, costly software – this is referred to as a "proprietary file type." You can tell that you have data in a proprietary file type if you lose access to the data when you lose access to the software. When data is in a proprietary file type, it's always a good idea to copy the data into a more common, open file type as a backup; you may lose a bit of functionality, but it's better to have a backup than to not have your data at all! This exercise works through identifying possible alternative file types for the data's proprietary file type before instructing you to make a copy of the data in the new file type.*

Instructions: *For any data in a proprietary file type, identify the data and answer the following questions. Once you have picked a more open, common file type, make a copy of the data in that file type but do not delete the original data. (Keeping a copy in the original file format means that, while you access to the necessary software, your data has full functionality. If you lose access to the software, you'll still have your data in some format, which is better than not having your data at all.)*

Is your data stored in a proprietary file type? What file type and how does this limit future data reuse?

Example: Data is stored in a .CZI file format, which is a proprietary Zeiss microscope image format. These files do not open in other software.

Is it possible to convert your data to other file types? If so, list the possible types:

Example: I can use the Bio-Formats tool to convert .CZI files to: .AVI, .CH5, .DCM/.DICOM, .EPS/.EPSI/.PS, .ICS/.IDS, .JPG, .JP2/.J2K/.JPF, .MOV, .OME.TIFF/.OME.TIF, .OME/.OME.XML, .PNG, or .TIFF/.TIF.

Which of the possible file types are in common use? Which of the possible file types can be opened by multiple software programs?

Example: JPG, PNG, and TIFF are all image formats in common use. OME-TIFF is a common image format within microscopy; most software will read

the TIFF portion of the file but only some software will read the extra OME metadata. Common movie file types are AVI and MOV.

Of the possible options above, do you have a preference for a specific file type?

Example: I prefer an image file over a movie file. TIFF is best because it doesn't lose resolution due to compression and can store all of the 4-dimensional image layers. OME-TIFF gives all of the benefits of TIFF but with added metadata.

Pick one of the more open or common file types and copy your important data files into that file type. Do not delete the original files.

Example: I will convert my data to OME-TIFF files.

7.3 Create an Archive Folder

Description: *To save your future-self time spent digging through all of your research files, set aside the most important files into a separate "Archive" folder. Do this at the end of the project while you still remember which files are important and where they are located. The Archive folder should only contain a small subset of the most important documents that are likely to be reused; you may still need to go through all of your files but, in the majority of instances, you will save time by easily finding what you need in the Archive folder.*

Instructions: *This exercise consists of a checklist of the key documents that are likely to be most useful in a research project archive. Create a separate folder within the larger project folder (or in a highly visible place within the storage system) labelled "Archive". Copy – do not move – the files on this checklist into the Archive folder. Add copies other important research documents, as needed. Remember, the Archive folder does not need to be comprehensive, so focus on the subset of files that are most likely to be reused or referenced in the future.*

Source: *This exercise was adapted from the "Project Close Out Checklist" [Briney, 2020b].*

Project Documentation

___ README file of project information

Data Snapshots

___ Important raw data

___ Key data analyses

___ Final published data

Code

___ Analysis code

___ Record software version, as appropriate

Other Research Documents

___ Protocols

___ Survey instruments

Research Notes

___ Scans of research notebook

___ Digital notes

Images

___ Flat files of figures (e.g. .JPG or .TIFF)

___ Editable image files (e.g. .XLSX or .PSD)

Publications

___ Published articles in .PDF format

___ Accepted version of articles in editable document format (e.g. .DOCX)

___ Poster files

Administrative Documents

___ Grant proposals

___ Grant progress reports and final report

7.4 Separate from the Institution

Description: *Researchers regularly leave institutions in order to take new jobs. For how common this occurrence is, it represents a critical transition during which data may be lost. This checklist enumerates a number of important steps that researchers can take to ensure that they retain the appropriate data yet leave behind what belongs to the institution.*

Instructions: *The researcher leaving the institution should work through this checklist to ensure they keep the proper information while returning what does not belong to them. The researcher and project administrator may also jointly work through the extended version of this exercise, the Data Departure Checklist [Goben and Briney, 2023].*

Source: *This exercise was adapted from the "Data Departure Checklist" [Goben and Briney, 2023].*

Retain Copies of Data that You Have Permission to Keep

___ If you have not done so already, work through Exercise 5.2: Determine Data Stewardship to determine what data you may retain

___ Identify and keep pertinent research data from personal devices

___ Identify and keep pertinent research data from storage systems (e.g. AWS/Azure, Box, campus HPC, Dropbox, Electronic Lab Notebook, Globus, Google Drive, lab/department/college servers, Microsoft OneDrive, Microsoft Sharepoint, or shared collaborator drives)

___ When appropriate, make a copy of research notes

Delete Personal Information and Remove Personal Devices

___ Remove personal information from lab devices

___ Remove personal devices from lab

___ Remove personal access to shared accounts (e.g. lab Github, lab repository page, lab website, mailing lists, or social media)

Return Lab Hardware

___ Individual computer / workstation

___ Tablet(s)

___ Peripherals (e.g. keyboard, mouse, monitor)

___ External drives

___ Other lab equipment (e.g. cameras, recording devices)

Update Research Administration Documents, As Necessary

___ Update/transfer Institutional Review Board

___ Update/transfer IACUC

___ Update/transfer Data Use Agreements (DUA)

___ Update/transfer Material Transfer Agreements (MTA)

___ Update/transfer research grants

Handle Email

___ Set out of office, providing forwarding information

___ Forward/backup important emails

___ Check with University Archivist or Records Manager for retention policies (depends on rank)

Acknowledgements

This book came to me at a moment when I was looking for a new big project. I was washing my hands and BAM! the idea for this book erupted in my head, fully formed like Athena from the head of Zeus. Prior to writing the Workbook, I had created several data management handouts and worksheets to use in lectures and workshops. My big idea was to expand my worksheet offerings to cover the entire data management lifecycle. Basically, I wanted a workbook of activities to pull from for any data management lesson I taught. The result is this Workbook.

Books are never truly written by only one person. While I did the majority of the labor for this Workbook, it would not have been as helpful, accurate, or complete without the assistance of several people.

First, thank you to several people at Caltech Library for assistance with this book. Thank you Donna Wrublewski for being on board with this big idea and supporting it. Big thanks to Tom Morrell for critical feedback on the beta edition of this book, particularly on the storage exercise which was really terrible before his feedback. Thank you Tommy Keswick for helping move my bookdown repo to the Caltech Library GitHub account and for showing me the magic that is GitHub Pages. Thank you to George Porter for helping to get the Workbook online into CaltechAUTHORS, even though we were not quite finished migrating that giant repository to its new platform.

Thank you to Abigail Goben and Dorothea Salo for always being encouraging when I come to them with the idea for a new book, no matter the idea. Abigail was one of the first people I discussed the idea for this book with, which helped cement the big idea in my head.

Thank you to Rachel Woodbrook for feedback on the accessible spreadsheets exercise.

Finally, thank you to Andy for being calm when I had the idea for yet another book. It helped that this book was much shorter than my last two. And thank you to Hank and Willa for cuddles, questions, and all of our adventures that keep me grounded.

Bibliography

Kristin Briney and Becky Yoose. *Managing Data for Patron Privacy: Comprehensive Strategies for Libraries*. ALA Editions, Chicago, IL, 2022. URL https://www.alastore.ala.org/mdpp.

Kristin Briney, Heather Coates, and Abigail Goben. Foundational Practices of Research Data Management. *Research Ideas and Outcomes*, 6:e56508, July 2020. ISSN 2367-7163. doi: 10.3897/rio.6.e56508. URL https://riojournal.com/article/56508/.

Kristin A. Briney. *Data Management for Researchers : Organize, Maintain and Share Your Data for Research Success*. Pelagic Publishing, 2015. URL https://pelagicpublishing.com/products/data-management-for-researchers-briney.

Kristin A. Briney. File Naming Convention Worksheet, June 2020a. URL https://doi.org/10.7907/894q-zr22.

Kristin A. Briney. Project Close-Out Checklist for Research Data, May 2020b. URL https://doi.org/10.7907/yjph-sa32.

Kristin A. Briney. Leveling Up Data Management, June 2023. URL https://doi.org/10.7907/syk7-3z92.

Karl W. Broman and Kara H. Woo. Data Organization in Spreadsheets. *The American Statistician*, 72(1):2–10, January 2018. ISSN 0003-1305. doi: 10.1080/00031305.2017.1375989. URL https://doi.org/10.1080/00031305.2017.1375989. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2017.1375989.

Creative Commons Wiki. CC0 use for data, 2014. URL https://wiki.creativecommons.org/wiki/CC0_use_for_data.

Abigail Goben and Kristin A. Briney. Data Departure Checklist, August 2023. URL https://doi.org/10.7907/h314-4x51.

NIH. Generalist Repositories, 2023. URL https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/generalist-repositories.

Emily Oxford and Rachel Woodbrook. Accessibility Data Primer, March 2023. URL https://github.com/DataCurationNetwork/data-primers/blob/main/Accessibility%20Data%20Curation%20Primer/accessibility-data-curation-primer.md.

PLOS ONE. Recommended Repositories, 2023. URL https://journals.plos.org/plosone/s/recommended-repositories.

Hadley Wickham. Tidy Data. *Journal of Statistical Software*, 59:1–23, September 2014. ISSN 1548-7660. doi: 10.18637/jss.v059.i10. URL https://doi.org/10.18637/jss.v059.i10.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, March 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18. URL https://www.nature.com/articles/sdata201618. Number: 1 Publisher: Nature Publishing Group.