Data Management in Large-Scale Education Research

Crystal Lewis

2023-03-01

# 1 Preamble

This is the in-progress version of *Data Management in Large-Scale Education Research*. To see a previous version of this material, please visit this [website](https://cghlewis.github.io/mpsi-data-training/).

*The results of educational research studies are only as accurate as the data used to produce them.* *- Aleata Hubbard*[[1]](#footnote-21)

## 1.1 Introduction

In 2013, without knowing that the term research data management existed, I accepted a job as a research associate with a research center. My job was to coordinate the collection and management of data for federally funded randomized controlled trial efficacy studies taking place in K-12 schools, along with a team of PIs, other full-time staff, part-time data collectors, and graduate students. While I had some experience analyzing and working with education data, i.e. ECLS-K, I had no experience running research grants, collecting original data, or managing research data, but I was excited to learn.

In my time in that position I learned to plan, schedule, and track data collection activities, create data collection tools, organize and document data inputs, and produce usable data outputs; but I didn’t learn to do these things through any formal training. There were no books, courses, or workshops that I learned from. I learned from colleagues and a large amount of trial and error. Since then, as I have met more PIs, data managers, and project coordinators in education research, I realize that this is a common method for learning data management (mentoring and “winging it”). And while learning data management through these informal methods helps us get by, the ramifications of this unstandardized system are felt by both the project team and future data users.

## 1.2 Why this book

Research data management is becoming more complicated. We are collecting more data, in sometimes very novel ways, and using more complex technologies, all while increasing the visibility of our work with the push for data sharing and open science practices.[[2]](#footnote-24) Ad hoc data management practices may have worked for us in the past, but now others need to understand our processes as well, requiring researchers to be more thoughtful in planning their data management routines.

### 1.2.1 Lack of training, resources, and standards

In order to implement thoughtful and standardized data management practices, researchers need training. Yet there is a clear lack of data management training in higher education. In a survey of 274 psychology researchers, Borghi and Van Gulick[[3]](#footnote-26) found that only 33% of respondents learned data management from college level coursework, while 64% learned from collaborators, and 52% learned from self-education. In their survey of 202 education researchers (PIs and Co-PIs), Ceviren and Logan[[4]](#footnote-28) found that over 60% of respondents reported having no formal training in data management, yet across eight different data management practices, respondents were responsible for data management activities anywhere from 25-50% of the time.

Without training, resources and formal support systems are the next best option for learning best practices. During my data management journey I have discovered an excellent support system of professionals in university systems, i.e. research data librarians, who can consult with research teams in their data management planning, and I have also come across some solid existing research data management books and manuals which I will link to in this book. However, while education researchers are starting to put out some excellent resources,[[5]](#footnote-30) I still find there is a dearth of tangible guides for researchers to refer to when building a data management workflow in the field of education, especially those working on large-scale longitudinal research grants where there are many moving pieces. Researchers are often collecting data in real-world environments, such as school systems, and keeping that data secure and reliable in a deliberate and orderly way can be overwhelming.

Last, unfortunately, while other fields of research, such as psychology, appear to be banding together to develop standards around how to structure and document data,[[6]](#footnote-33) the field of education has yet to develop agreed upon rules for things such as data documentation or data formats. This lack of standards leads to inconsistencies in the quality of data products across the field.[[7]](#footnote-35)

### 1.2.2 Consequences

A lack of training in data management practices and an absence of agreed upon standards in the field of education leads to consequences. Implementing subpar and inconsistent data management practices, while typically only resulting in frustration and time lost, also has the potential to be devastating, resulting in analyzing erroneous data or even unusable or lost data. In a review of 1,082 retracted publications from the journal PubMed from 2013-2016, authors found that 32% of retractions were due to data management errors.[[8]](#footnote-38) In a 2013 study surveying 360 graduate students about their data management practices, 14% of students indicated they had to recollect data that had been previously collected because they could not find a file or the file had been corrupted, while 17% of students said they had lost a file and been unable to recollect it.[[9]](#footnote-40) In their 2021 study of 488 researchers who had published in a psychology journal between 2010 and 2018, Kovacs et al.[[10]](#footnote-42) asked respondents about their data management mistakes and found that the most serious data management mistakes reported led to a range of consequences including time loss, frustration, and even erroneous conclusions.

Poor data management can even prevent researchers from implementing other good open science practices. In waves 1 and 2 of the Open Scholarship Survey being collected by the Center for Open Science, the team has found that of the education researchers surveyed who are currently not publicly sharing their research data, about 10% mentioned “being nervous about mistakes” as a reason for not sharing.[[11]](#footnote-44) The well known replication crisis is another reason to be concerned with data management. Failure to implement practices such as quality documentation or standardization of practices (among many other reasons), resulted in one study finding that across 1,500 researchers surveyed, more than 70% had tried and failed to reproduce another researcher’s study.[[12]](#footnote-46)

## 1.3 About this book

While the field as a whole may not have agreed upon guidelines for data management, there are still practices that are proven to result in more secure, reproducible, and reliable data. My hope is that this book can be a foundation to help researchers think through how to build a quality, standardized data management workflow that works for their team and their projects. As suggested in the title of this book, this content is designed to specifically help teams navigate the complicated workflows associated with large-scale research studies, such as randomized controlled trial studies, but ultimately these practices are applicable to any research project, no matter the scale.

This book should be viewed as a handbook to be referenced regularly and is not necessarily meant to be read in its entirety in one sitting. While perusing through the entire book to better understand the entire research data life cycle is very helpful, this book is also intended to have chapters referenced as needed when you are ready to start planning a specific phase of your project.

### 1.3.1 What this book will cover

This book begins, like many other books in this subject area, by describing the research life cycle and how data management fits within the larger picture. The remaining chapters are then organized by each phase of the life cycle, with examples of best practices provided for each phase. Considerations on whether you should implement, and how to integrate those practices into your workflow will be discussed.

### 1.3.2 What this book will not cover

It is important to also point out what this book will not cover. This book is intended to be tool agnostic and provide suggestions that anyone can use, no matter what tools you work with, especially when it comes to data cleaning. Therefore, while I might mention options of tools you can use for different tasks, I will not advocate for any specific tools.

There are also no specific coding practices or actual syntax included in this book. To be honest, in many ways I feel that the actual “data cleaning” phase of data management is the *easiest* phase to implement, as long as you implement good practices up until that point. Because of that, this book introduces practices in all phases leading up to data cleaning that will prepare your data for minimal cleaning. With that said, I do provide examples of what I would expect to see in a data cleaning process, I just do not provide steps for any specific software system. That is beyond the scope of this book.

This book will also not talk about analysis or preparing data for analysis through means such as data imputation, removal of legitimate outliers, or calculating analysis specific variables. This book is written from the perspective of a data manager, and that perspective is to build datasets for general data sharing. This means we will cover practices that keep data in its most complete and true, but usable form, for any future researcher to analyze in a way that works best for them.

## 1.4 Who this book is for

This book is for anyone involved in a research study involving original data collection. This book in particular focuses on quantitative, observational data collection, while I do think that many of the practices covered can also apply to qualitative data as well. This book also applies to any team member, ranging from PIs, to data managers, to project staff, to students, to contractual data collectors. The contents of this book are useful for anyone who may have a part in planning, collecting, or organizing research study data.

## 1.5 Final note

Planning and implementing new data management practices on top of planning the implementation of your entire research grant can feel overwhelming. However, the idea of this book is to find the practices that work for you and your team and implement them consistently. For some teams that may look like implementing just a few of the suggestions mentioned and for others it may involve implementing all of the suggestions. Improving your data management workflow is a process and it becomes easier over time as those practices become part of your normal routine. At some point you may even find that you enjoy working on data management processes as you start to see the benefits of their implementation!

## 1.6 Acknowledgements

This book is a compilation of lessons I have learned in my personal experiences as a data manager, knowledge collected from existing books and papers (many written by librarians or those involved in the open science movement), as well as advice and stories collected through interviews with other researchers who work with data. I want to be clear that I did not formally study research data management, unlike research data librarians who are experts in this content. Much of this book will be based off of lessons learned from firsthand experience and this book is my attempt to hopefully save others from making the same mistakes I have personally made or seen others make. I can not emphasize enough that if you work for a university and you have the opportunity to consult with a librarian for your project, you absolutely should!

With that said, there is a long list of people I would like to acknowledge for their contributions to this book and for supporting me in this process.

Interviewees:

Others:

# 2 Research Data Management

## 2.1 What is research data management?

Research data management (RDM) involves the organization, storage, preservation, and dissemination of research study data.[[13]](#footnote-57) Research study data includes materials generated or collected throughout a research process.[[14]](#footnote-59) As you can imagine, this broad definition includes much more than just the management of digital datasets. It also includes physical files, documentation, artifacts, recordings, and more. RDM is a substantial undertaking that begins long before data is ever collected, during the planning phase, and continues well after a research project ends during the archiving phase.

## 2.2 Standards

Data management standards refer to rules for how data should be stored, organized, and described.[[15]](#footnote-62) Some fields have adopted standards across the research life cycle, such as CDISC standards used by clinical researchers,[[16]](#footnote-63) other fields have adopted standards specifically around metadata, such as the TEI standards used in digital humanities,[[17]](#footnote-65) and through grassroots efforts, other fields such as psychology are developing their own standards for things such as data structure and documentation based on the FAIR principles.[[18]](#footnote-67) Yet, it is common knowledge that there are currently no agreed-upon norms for how to structure and share data in the field of education.[[19]](#footnote-68) The rules for what data should be produced and how it should be documented is often left up to each individual team, as long as external compliance requirements are met.[[20]](#footnote-70) However, with a growing interest in open science practices and expanding requirements for federally funded research to make data publicly available,[[21]](#footnote-72) data repositories will most likely begin to play a stronger role in promoting standards around data formats and documentation.[[22]](#footnote-73)

While field standards for the structure and format of publicly shared products that aid in the preservation and re-use of data are very much needed, there are actually good reasons to not impose standardization on all data management activities across the field. Granting some flexibility in the process of managing data during active data collection allows teams to implement the best practices that work for their projects, as long as those projects implement practices consistently during their project and produce similar quality outputs across the field.

## 2.3 Why care about research data management?

Without current agreed-upon standards in the field, it is important for research teams to develop their own data management standards that apply within and across all of their projects. Developing internal standards, implemented in a reproducible data management workflow, allows practices to be implemented consistently and with fidelity. There are both external pressures and personal reasons to care about developing research data management standards for your projects.

### 2.3.1 External Reasons

1. **Funder compliance**: Any researcher applying for federal funding will be required to submit a data management plan (DMP) along with their grant proposal[[23]](#footnote-75). The contents of these plans may vary slightly across agencies but the shared purpose of these documents is to facilitate good data management practices and to mandate open sharing of data to maximize scientific outputs and benefits to society. Along with this mandatory data sharing policy, comes the incentive to manage your data for the purposes of data sharing.[[24]](#footnote-76)
2. **Journal compliance**: Depending on what journal you publish with, providing open access to the data associated with your publication may be a requirement (see PLOS ONE[[25]](#footnote-77) and AMPPS[[26]](#footnote-79) as examples). Again, along with data sharing, comes the incentive to manage your data in a thoughtful, responsible, and organized way.
3. **Compliance with legal and ethical mandates**: If you are required to submit your research project to the Institutional Review Board (IRB), they will monitor how you manage your data. The IRB is concerned with the welfare, rights, and privacy of research participants and will have rules for how data is managed and stored securely. Additionally your organization may have their own institutional data policies that mandate how data must be cared for and secured.[[27]](#footnote-81)
4. **Open science practices**: With a growing interest in open science practices, sharing well managed and documented data helps to build trust in the research process.[[28]](#footnote-83) Sharing data that is curated in a reproducible way is “a strong indicator to fellow researchers of rigor, trustworthiness, and transparency in scientific research” (Alston & Rick, 2021, p.2).[[29]](#footnote-85) It also allows others to replicate and learn from your work, validate your results to strengthen evidence, as well as potentially catch errors in your work, preventing decisions being made based on incorrect data.[[30]](#footnote-87) Well-managed data with sufficient documentation can also lead to more collaboration and greater impact as collaborators are able to access and understand your data with ease.[[31]](#footnote-88)

### 2.3.2 Personal reasons

Even if you never plan to share your data outside of your research group, there are still many compelling reasons to manage your data in a reproducible and standardized way.

1. **Reduces data curation debt**: Taking the time to plan and implement quality data management through the entire research study reduces data curation debt caused by suboptimal data management practices.[[32]](#footnote-92) Having poorly collected, managed, or documented data may make your data unusable, either permanently or until errors are corrected. Decreasing or removing this debt reduces the time, energy, and resources spent possibly recollecting data or scrambling at the end of your study to get your data up to acceptable standards.
2. **Facilitates use of your data**: Every member of your research team being able to find and understand your project data and documentation is a huge benefit. It allows for the easy use and re-use of your data, and hastens efforts like the publication process.[[33]](#footnote-94) Not having to search around for numbers of consented participants or asking which version of the data they should use allows your team to spend more time analyzing and less time playing detective.
3. **Encourages validation**: Implementing reproducible data management practices encourages and allows your team to internally replicate and validate your processes to ensure your outputs are accurate.
4. **Improves continuity**: Data management practices such as documentation ensures project continuity through staff turnover. Having developed thorough protocols allows new staff to pick up right where the former staff member left off and implement the project with fidelity.[[34]](#footnote-96) Furthermore, good data management enables continuity when handing off projects to collaborators or when picking up your own projects after a long hiatus.[[35]](#footnote-97)
5. **Increases efficiency**: Documenting and automating tasks reduces duplication of efforts for repeating tasks, especially in longitudinal studies.
6. **Upholds research integrity**: Errors come in many forms, from both humans and technology[[36]](#footnote-98). We’ve seen evidence of this in the papers cited as being retracted for “unreliable data” in the blog Retraction Watch.[[37]](#footnote-100) Implementing quality control procedures reduces the chances of errors occurring and allows you to have confidence in your data. Without implementing these practices, your research findings could include extra noise, missing data, or erroneous or misleading results.
7. **Improves data security**: Quality data management practices reduce the risk of lost or stolen data, the risk of data becoming corrupted or inaccessible, and the risk of breaking confidentiality agreements.

## 2.4 Existing Frameworks

Data management does not live in a space all alone. It co-exists with other frameworks that impact how and why data is managed and it is important to be familiar with them as they will provide a foundation for you as you build your data management structures.

### 2.4.1 FAIR

In 2016, the FAIR Principles[[38]](#footnote-104) were published in Scientific Data, outlining four guiding principles for scientific data management and stewardship. These principles were created to improve and support the reuse of scholarly data, specifically the ability of machines to access and read data, and are the foundation for how all digital data should be publicly shared.[[39]](#footnote-106) The principles are:

F: Findable

All data should be findable through a persistent identifier and have thorough, searchable metadata. As we move towards automation in our work and life, the need for machine-readable data and metadata becomes more prevalent for automatic discovery of information.

A: Accessible

Users should be able to access your data. This can mean your data is available in a repository or through a request system. At minimum, a user should be able to access the metadata, even if the actual data are not available.

I: Interoperable

Your data and metadata use standardized vocabularies as well as formats. Both humans and machines should be able to read and interpret your data. Software licenses should not pose a barrier to usage. Data should be available in open formats that can be accessed by any software (ex: .csv, .txt, .dat).

R: Reusable

In order to provide context for the reuse of your data, your metadata should give insight into data provenance, providing a project description, an overview of the data workflow, as well what authors to cite for appropriate attribution. You should also have clear licensing for data use.

### 2.4.2 SEER

In addition to the FAIR principles, the SEER principles, developed in 2018 by Institute of Education Sciences (IES), provide Standards for Excellence in Education Research.[[40]](#footnote-109) While the principles broadly cover the entire life cycle of a research study, they provide context for good data management within an education research study. The SEER principles include:

* Pre-register studies
* Make findings, methods, and data open
* Identify interventions’ core components
* Document treatment implementation and contrast
* Analyze interventions’ costs
* Focus on meaningful outcomes
* Facilitate generalization of study findings
* Support scaling of promising results

### 2.4.3 Open Science

The concept of Open Science has pushed quality data management to the forefront, bringing visibility to its cause, as well as advances in practices and urgency to implement them. Open Science aims to make scientific research and dissemination accessible for all, making the need for good data management practices absolutely necessary. Open science advocates for transparent and reproducible practices through means such as open data, open analysis, open materials, preregistration, and open access.[[41]](#footnote-112) Organizations such as the Center for Open Science,[[42]](#footnote-114) have become a well-known proponents of open science, offering the open science framework (OSF)[[43]](#footnote-116) as a tool to promote open science through the entire research life cycle. Furthermore, many education funders have aligned their fundee requirements with these open science practices, such as openly sharing study data and pre-registration of study methods.[[44]](#footnote-118)

## 2.5 Terminology

Before diving into the content of this training, I think it is helpful to cover terminology that will be used in data management. Many concepts in education research have multiple terms and can be used interchangeably. Across different institutions, researchers may use all or some of these terms.

## 2.6 The Research Life Cycle

The remainder of this book will be organized into chapters that dive into phases of the research data life cycle. It is imperative to understand this research life cycle in order to see the flow of data through a project, as well as to see how everything in a project is connected. If phases are skipped, the whole project will suffer.

You can see in Figure 2.1, how throughout the project, data management roles and project coordination roles work in parallel and collaboratively. These teams may be made up of the same people or different members, but either way, both workflows must happen and they must work together.

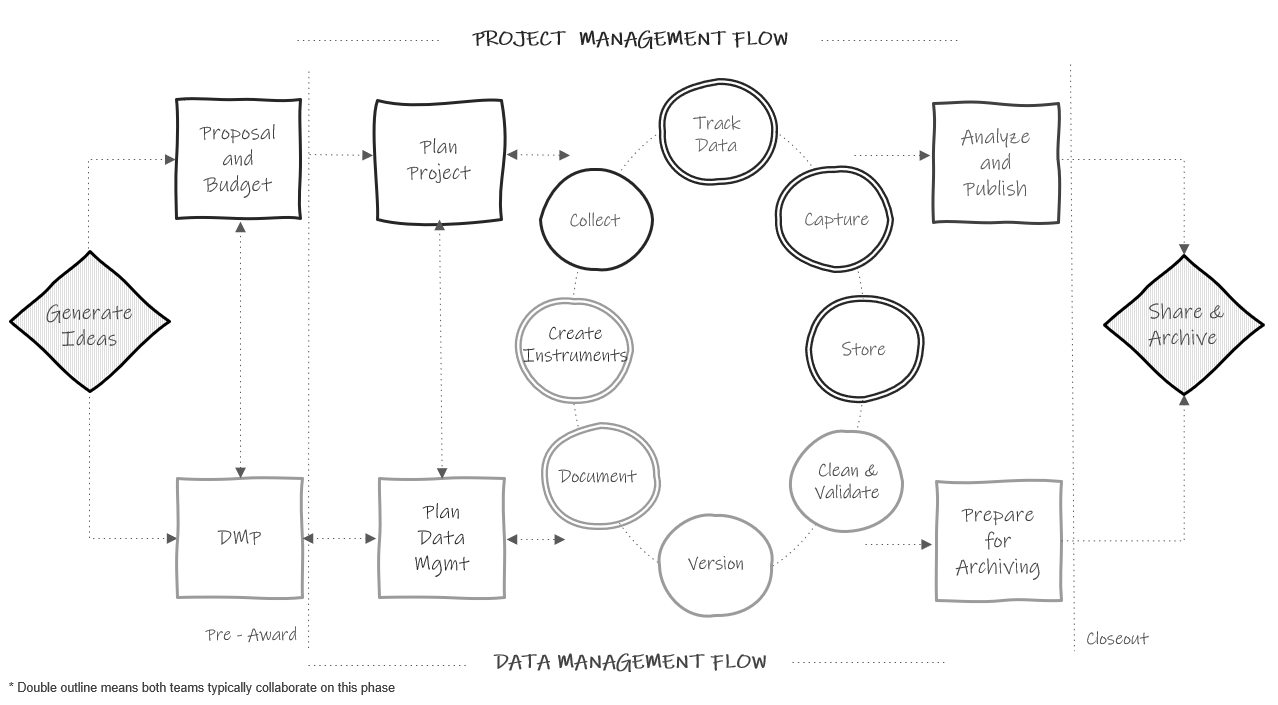


Figure 2.1: The research project life cycle

Let’s walk through this chart.

1. In a typical study we first begin by **generating ideas**, deciding what we want to study.
2. Then, most likely, we will look for grant funding to implement that study. This is where the two paths begin to diverge. If the team is applying for federal funding, the proposal and budget are created in the project management track, while the supplemental required [**data management plan**](#dmp) is created in the data track. Again, it may be the same people working on both of these pieces.
3. Next, if the grant is awarded, the project team will begin planning things such as hiring, recruitment, data collection, and how to implement the intervention. At the same time, those working on the data team will begin to **plan** out how to specifically implement the 2-5 page data management plan submitted to their funder and start putting any necessary structures into place.
4. Once planning is complete, the team moves into the cycle of data collection. It is called a cycle because if your study is longitudinal, every step here will occur cyclically. Once one phase of data collection wraps up, the team re-enters the cycle again for the next phase of data collection, until all data collection is complete for the entire project.
   * The data management and project management team begin the cycle by starting **documentation**. You can see that this phase occurs collaboratively because it is denoted with a double outline. Both teams begin developing documentation such as data dictionaries and standard operating procedures.
   * Once documentation is started, both teams collaboratively begin to create any necessary **data collection instruments**. These instruments will be created with input from the documentation. During this phase the teams may also develop their participant tracking database.
   * Next, the project management team moves into the **data collection** phase. In addition to actual data collection, this may also involve preliminary activities such as recruitment and consenting of participants, as well as hiring and training of data collectors. At this point, the data management team just provides support as needed.
   * As data is collected, the project team will **track data** as it is collected in the participant tracking database. The data management team will collaborate with the project management team to help troubleshoot anything related to the actual tracking database or any issues discovered with the data during tracking.
   * Next, once data is collected, the teams move into the **data capture** phase. This is where teams are actively retrieving or converting data. For electronic data this may look like downloading data from a platform or having data sent to the team via a secure transfer. For physical data, this may look like teams entering paper data into a database. Oftentimes, this again is a collaborative effort between the project management team and the data team.
   * Once the data is captured, it needs to be **stored**. While the data team may be in charge of setting up and monitoring the storage efforts, the project team may be the ones actively retrieving and storing the data.
   * Next the teams move into the **cleaning and validation** phase. At this time the data team is reviewing data cleaning plans, writing data cleaning scripts, and actively cleaning data from the most recent data collection round.
   * And last, the data team will **version** data as it is updated or errors are found.
5. The teams then only move out of the active data collection phase when all data collection for the project is complete. At this time the project team begins analyzing study data and working on publications as well as any final grant reports. They are able to do this because of the organized processes implemented during the data collection cycle. Since data was managed and cleaned throughout, data is ready for analysis as soon as data collection is complete. Then, while the project team is analyzing data, the data team is doing any additional **preparation to archive** data for public sharing.
6. Last, as the grant is closing out, the team submits data for **public sharing**.

As you work through the remaining chapters of this book, this chart will be a guide to navigating where each phase of practices fits into the larger picture.

# 3 Data Structure

Because data management is made up of just that, data, we need to have a basic understanding of what data looks like. Understanding the basic structure of data helps us write our Data Management Plan, organize our data management process, create our data dictionaries, build our data collection tools, and clean our data, all in ways that allow us to have analyzable data.

## 3.1 Basics of a dataset

In education research, data is often collected internally by your team using an instrument such as a questionnaire, an observation, an interview, or an assessment. However, data may also be collected from external entities, such as districts, states, or other agencies.

Those data come in many forms (ex: video, transcripts, documents, files), represented as text, numbers, or multimedia.[[45]](#footnote-128) In the world of quantitative education research, we are often working with digital data in the form of a dataset, a structured collection of data. These datasets are organized in a rectangular format which allow the data to be machine-readable. Even in qualitative research, we are often wrangling data to be in a format that is analyzable and allows categorization.

These rectangular (also called tabular) datasets are made up of columns and rows.



Figure 3.1: Basic format of a dataset

### 3.1.1 Columns

The columns in your dataset will consist of one or both of the following types of variables:

* Variables you collect (from an instrument or from an external source)
* Variables you create/add (ex: cohort, intervention, time, derivations)

Unless your data is collected anonymously, every dataset **must** also have the following:

* One or more variables that are **unique identifiers**, sometimes called primary keys. These are variables that uniquely define rows in your dataset (i.e. help you identify duplicate rows), and they also allow you to link data that contain the same identifiers (for example link all student data).
* If you plan to link datasets across entities (ex: link teachers to schools or students to teachers) then you will also need secondary unique identifiers in your dataset (also called foreign keys) that allow you to link across datasets.

We will talk more about creating these identification variables in our [Data Tracking](#ids) chapter.

**Column attributes**

It is important to know that variables have the following attributes:

1. Unique names (no variable name in a dataset can repeat). We will talk more about variable naming when we discuss [style guides](#style).
2. A measurement type (ex: numeric, character, date) which can also be more narrowly defined as needed (ex: continuous, categorical)
3. Acceptable values (ex: yes/no) or expected ranges (ex: 1-25 or 2021-08-01 to 2021-12-15). Anything outside of those acceptable values or ranges is considered an error.
4. Labels, descriptions of what the variable represents. This may be a label that you as the variable creator assigns (ex: “Treatment condition”) or they may be the actual wording of an item (ex: “Do you enjoy pizza?”).

### 3.1.2 Rows

The rows in your dataset are aligned with participants or cases in your data. Participants in your data may be students, teachers, schools, locations, and so forth. The unique identifier variable mentioned above will denote which row belongs to which participant.

### 3.1.3 Cells

The cells are the observations associated with each participant. Cells are made up of key/value pairs, created at the intersection of a column and a row. Consider an example where we collect a survey from students. In this dataset, each row is made up of a unique student in our study, each column is an item from the survey, and each cell contains a value/observation that corresponds to that row/column pair (that participant and that question).



Figure 3.2: Representation of a cell value

## 3.2 Dataset organization rules

In order for your dataset to be machine-readable and analyzable, it should adhere to a set of structural rules.[[46]](#footnote-140)

1. The first rule is that your data should make a rectangle. The first row of your data should be your variable names (only use one row for this). The remaining data should be made up of values in cells.



Figure 3.3: A comparison of non-rectangular and rectangular data

1. Your columns should adhere to your variable type.
   * For example, if you have a numeric variable, such as age, but you add a cell value that is text, your variable no longer adheres to your variable type. Machines will now read this variable as text.

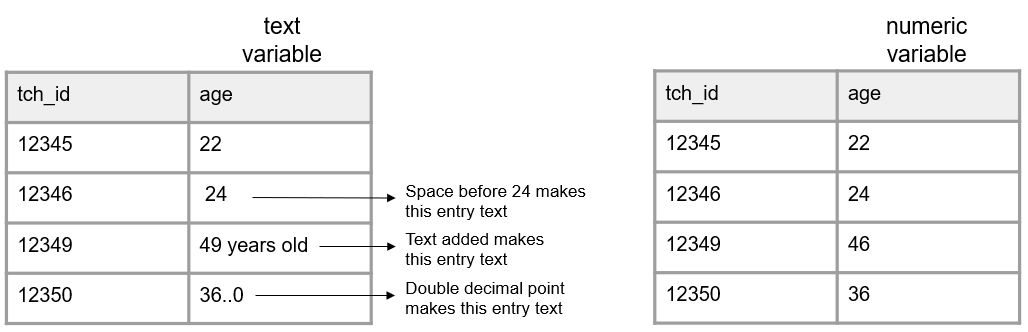


Figure 3.4: A comparison of variables adhering and not adhering to a data type

1. A variable should only collect one piece of information. If a variable contains more than one piece of information you may have the following issues:
   * You lose the granularity of the information (ex: location = Los Angeles, CA is less granular than having a city variable and a state variable separately)
   * Your variable may become unanalyzable (ex: a variable with a value 220/335 is not analyzable as a numeric variable). If you are interested in a rate, you can calculate a rate variable with a value of .657.
   * You may lose the variable type (ex: if you want an incident\_rate variable to be numeric, and you assign a value of 220/335, that variable is no longer numeric)



Figure 3.5: A comparison of two things being measured in one variable and two things being measured across two variables

1. All cell values should be explicit. This means all cells should be filled in with a physical value.
   * No cells should be empty
     + If a value is actually missing, make sure it contains a value to denote the missing data (ex: NA) to show that the cell was not left blank unintentionally
     + If a cell is left empty because it is “implied” to be the same value as above, the cells should be filled with the actual data
     + If the value for the cell is “implied” to be 0, fill the cells with 0



Figure 3.6: A comparison of of variables with empty cells and variables with not empty cells

* No values should be implied using color coding
  + If you want to indicate information, add an indicator variable to do this rather than cell coloring



Figure 3.7: A comparison of variables with implicit values and variables with explicit values

1. Your data should not contain duplicate rows. You do not want duplicate rows of a measurement collected **on the same participant**, **at the same time period**. Different types of duplicate rows can occur:
   * A true duplicate row where an entire row is duplicated (the row values are the same for every variable). This may happen if someone enters the same form twice.
   * A unique identifier is duplicated but the row values may or may not be the same across all of the variables. This could happen because one of three reasons:
     1. An instrument is accidentally collected more than once on the same participant in a collection period. This type of duplicate would need to be remedied.
     2. A unique identifier was entered incorrectly. In this case you don’t actually have a duplicate, you just have an incorrect unique identifier. This error would need to be remedied.
     3. More than one variable is used to identify unique participants and the row is not actually a duplicate.
        + Take for example a student id and a class id. Multiple unique identifiers may be used if data is collected on participants in multiple locations and treated as unique data. In this case, the data is not truly duplicate because the combined identifiers are unique.
        + Another example of this is if your data is organized in long format ([discussed below](#structure)). In this case unique study identifiers may repeat in the data but they should not repeat for the same form and same time period in your data.



Figure 3.8: A comparison of data with duplicate cases and data with no duplicate cases

## 3.3 Linking data

Up until now we have been talking about one, standalone dataset. However, it is more likely that your research project will be made up of multiple datasets, collected from different participants, from a variety of instruments, and possibly across different time points. And at some point you will most likely need to link those datasets together.

In order to think about how to link data, we need to discuss two things: data structure and database design.

### 3.3.1 Database design

A database is “an organized collection of data stored as multiple datasets.”[[47]](#footnote-162) Sometimes this database is actually housed in a database software system (such as SQLite or FileMaker), and other times we are loosely using the term database to simply define how we are linking disparate datasets together that are stored individually in some file system. No matter the storage system, the general concepts here will be applicable.

In database terminology, each dataset we have is considered a “table”. Each table has a primary key that identifies unique entries within a table and each table can be connected through both primary and foreign keys. This linking of tables creates a relational database and we will talk more about this structure when we discuss [participant data tracking](#track).

Let’s take the simplest example, where we only have primary keys in our data. Here we collected two pieces of data from students (a survey and an assessment) in one time period. Figure 3.9 shows what variables were collected from each instrument and how each table can be linked together through a primary key (circled in yellow).



Figure 3.9: Linking data through primary keys

However, we are often not only collecting data across different forms, but we are also collecting nested data across different participants (ex: students, nested in classrooms, nested in schools, and so on). Let’s take another example where we collected data from three instruments, a student assessment, a teacher survey, and a school intake form. Figure 3.10 shows what variables exist in each dataset (with primary keys still being circled in yellow) and how each table can be linked together through a foreign key (circled in blue).



Figure 3.10: Linking data through foreign keys

And as you can imagine, as we add more forms, or begin to collect data across time, the database structure begins to become even more complex. Figure 3.11 is another example where we collected two forms from students (a survey and an assessment), two forms from teachers (a survey and an observation), and one form from schools (an intake form). While the linking structure begins to look more complex, we see that we can still link all of our data through primary and foreign keys. Forms within participants can be linked by primary keys, and forms across participants can be linked by foreign keys.



Figure 3.11: Linking data through primary and foreign keys

### 3.3.2 Data structure

When it comes time to link our data, there are two ways we often think about linking or structuring our data, wide or long.

**Wide format**

When we structure our data in a wide format, all data collected on a unique participant will be in one row. Participants should **not** be duplicated in your data in this format.

This type of format can be used for the following situations:

* To link forms within time
* To link forms across time
* To link forms across participants

The easiest scenario to think about this format is with repeated measure data. If we collect a survey on participants in both wave 1 and 2, those waves of data will all be in the same row (joined together on a unique ID) and each wave of data collection will be appended to a variable name to create unique variable names. We will dive deeper into different types of joins in our [data cleaning](#clean) section.

Limitations: It is important to note here, that if your data do not have unique identifiers (primary and/or foreign keys), you will be unable to merge data in a wide format.



Figure 3.12: Data structured in wide format

**Long format**

In education research, long data is mostly used as a specific way to structure data that is collected over time. In long data a participant can and will repeat in your dataset.

Again, the most straight forward way to think about this is with repeated measure data, where each row will be a new time point for a participant. Here instead of merging forms on a unique id, we stack forms on top of each other, often called appending data. Rows are stacked on top of one another and variables are aligned by variable name. Now instead of linking data by an id, data is now “linked” by variable names. It is important here that variable names and types stay identical over time in order for this structure to work.

In this scenario, we no longer add the data collection wave to variable names. However, we would need to add a time period variable to denote the wave associated with each row of data.



Figure 3.13: Data structured in long format

**Choosing wide vs long**

There are different reasons for constructing your data one way or another. And it may be that you store or share your data in one format, and then restructure data into another format when it comes time for analysis.

Storing data in long format is usually considered to be more efficient, potentially requiring less memory. However, when it comes time for analysis, specific data structures may be required. For example, repeated measure procedures typically require data to be in wide format, where the unit of analysis is the subject. While mixed model procedures typically required data to be in long format, where the unit of analysis is each measurement for the subject.[[48]](#footnote-179) We will further review decision making around data structure in our [Data Cleaning](#clean) chapter.

## 3.4 File types

These rectangular datasets can be saved in a variety of file types. Some common file types in education research include interoperable formats such as .csv, .txt, .dat, or .tsv, or proprietary formats such as .xlsx, .sav, or .dta.

When you save your files, they will have a file size. Both the number of columns as well as the number of rows in your dataset will contribute to your file size. Just to get a feel for what size your files might be, small datasets (for example 5 columns and <100 rows) may be less than 100 KB. Datasets with several hundred variables and several thousand cases may start to be in the 1,000-5,000 KB range. The type of file you use also changes the size of your data. Saving data in a format that contains embedded metadata (such as variable and value labels), such as a .sav file, will greatly increase your file size. We will talk about the pros and cons to different file formats in the chapter on [Data Sharing](#share).

# 4 Data Management Plan

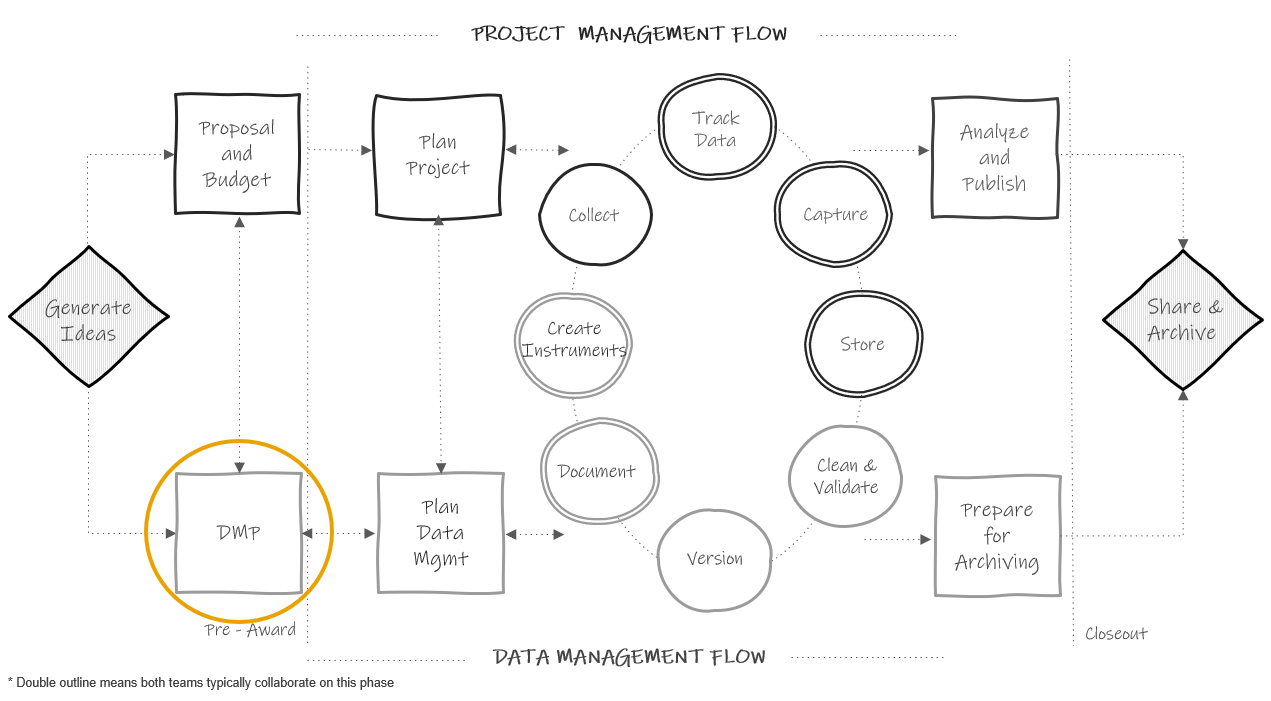


Figure 4.1: Data management plan in the research project life cycle

## 4.1 History and purpose

Since 2013, even earlier for the National Science Foundation, most federal agencies that education researchers work with have required a data management plan (DMP) as part of their funding application. While the focus of these plans is mostly on the future outcome of data sharing, the data management plan is a means of ensuring that researchers will thoughtfully plan for a research study that will result in data that can be shared with confidence, and free from errors, uncertainty, or violations of confidentiality. President Obama’s May 2013 Executive Order declared that “the default state of new and modernized government information resources shall be open and machine readable.”[[49]](#footnote-188) In August of 2022, the Office of Science and Technology Policy (OSTP) doubled down on their data sharing policy and issued a memorandum stating that all federal agencies must update their public access policies no later than December 31, 2025, to make federally funded publications and their supporting data accessible to the public with no embargo on their release.[[50]](#footnote-190) Even sooner than this, organizations like the National Institutes of Health have mandated that grant applicants, beginning January 2023, must submit a plan for both managing and sharing project data[[51]](#footnote-191).

### 4.1.1 Why are DMPs important?

Funding agencies see DMPs as important in maximizing scientific outputs from investments and increasing transparency. Mandating data sharing for federally funded projects leads to many benefits including accelerating discovery, greater collaboration, and building trust among data creators and users. In addition to the benefits viewed by funders, there are intrinsic benefits that come from having to write a data management plan. Having to thoughtfully plan and having transparency in that plan leads to better data management. Knowing that you will eventually be sharing your data and documentation with others outside of your team can motivate researchers to think hard about how to organize their data management practices in a way that will produce data that they trust to share with the outside world[[52]](#footnote-193).

## 4.2 What is it?

Generally, a data management plan is a supplemental 2-5 page document, submitted with your grant application, that contains details about how you plan to store, manage, and share your research data products. For most funders these DMPs are not part of the scoring process, but they are reviewed by a panel or program officer. Some funders may provide feedback or ask for revisions if they believe your plan and/or your budget and associated costs are not adequate.

### 4.2.1 What to include?

What to include in a DMP varies some across funding agencies. While you should check each funding agency’s site for their specific DMP requirements, there are typically 10 common categories covered in a data management plan[[53]](#footnote-197) . Those categories are:

1. Roles and responsibilities
   * What are the staff roles in management and long-term preservation of data?
   * Who ensures accessibility, reliability, and quality of data?
   * Is there a plan if a core team member leaves the project or institution?
2. Types of data
   * How is data captured? (Ex: surveys, assessments, observations)
   * Will data be item-level and summary scores?
   * Will you share raw data and clean data?
   * What are the expected number of files? Expected number of rows/cases in each file?
3. Format of data
   * Will data be in an electronic format?
   * Will it be provided in a non-proprietary format? (Ex: .csv)
   * Will more than one format be provided? (Ex: .sav and .csv)
   * Are there any tools needed to manipulate shared data?
4. Documentation
   * What [documentation](#document) will you share? (Consider project level, dataset level, and variable level documentation)
   * What [metadata](#metadata) will you create?
   * What format will your documentation be in? (Ex: .xml, .csv, .pdf)
   * What [supplemental documents](#supplement) do you plan to include when sharing data? (Ex: consort diagrams, data collection instruments, consent forms)
5. Standards
   * Do you plan to use any metadata [standards](#metastandards)?
6. Method of data sharing
   * How will you share your data? (Ex: Institutional archive, data repository, PI website)
   * Will data be restricted and is a data enclave required?
   * Is a data use agreement required?
   * How will you license your data?
   * Will your data have persistent unique identifiers?
7. Circumstances preventing data sharing
   * Do you have any data covered by FERPA/HIPAA that doesn’t allow data sharing?
   * Do you work with any partners that do not allow you to share data? (Ex: School districts, tribal regulations)
   * Are you working with proprietary data?
8. Privacy and rights of participants
   * How will you prevent disclosure of personally identifiable information when you share data? How will you anonymize data (if applicable)?
   * Do participants sign informed consent agreements? Does the consent communicate how participant data are expected to be used and shared?
9. Data security
   * How will you maintain participant privacy and confidentiality during your project?
   * How will you prevent unauthorized access of data?
   * Consider IRB requirements here.
10. Schedule for data sharing
    * When will you share your study data and for how long?
11. Pre-registration (less commonly required)
    * Where and when will you pre-register your study?

Again, the specifics of what should be included in each category will vary by funder. Here are sites to visit to learn more about the four most common federal education research funder DMP requirements.

* Institute of Education Sciences[[54]](#footnote-200)
* National Institutes of Health[[55]](#footnote-202)
* National Institute of Justice[[56]](#footnote-203)
* National Science Foundation[[57]](#footnote-205)

## 4.3 Getting help

Since DMPs are written before a project is funded, and therefore before additional staff members may be hired, oftentimes the investigators developing the grant proposal are the ones who write the DMP. However, when constructing your DMP it is well worth your time to enlist help. If you have an existing data manager or data team, you will most certainly want to consult with them when writing your plan to ensure your decisions are feasible. If you work for a university system, your research data librarians are also excellent resources with a wealth of knowledge about writing comprehensive data management plans. And last, if you plan to share your final data with a repository or institutional archive you will want to contact your repository when writing your plan as well. The repository may have its own requirements for how and when data must be shared and it is helpful to outline those guidelines in your data management plan at the time of submission. You can also specifically write the name of your repository into your data management plan as well. Last, you may want to obtain the help of your colleagues. Your colleagues have likely written DMPs before and many people are willing to share their plans as a way to help others better understand what to include.

Your DMP is a living document and you can always update your plan during or after your project completion. It may be helpful to keep in contact with your program officer regarding any potential changes throughout your project.

If you are looking for guidance in writing a DMP, a variety of generic DMP templates for different federal agencies are available, as well as actual copies of submitted DMPs that some researchers graciously make publicly available for example purposes.

| Document | Description |
| --- | --- |
| DMPTool Templates[[58]](#footnote-209) | Templates organized by funding agencies |
| Sara Hart DMP Example[[59]](#footnote-211) | A submitted DMP that is publicly available for example purposes |
| UMN Libraries Examples[[60]](#footnote-213) | Submitted DMP examples from University of Minnesota researchers |
| ICPSR NIH Template[[61]](#footnote-215) | NIH Data Management and Sharing Plan template with specific recommendations for depositing data with ICPSR |
| Figshare DMP Example Prompts[[62]](#footnote-217) | DMP prompts specific to depositing data with Figshare |

## 4.4 Budgeting

As briefly mention above, funding agencies acknowledge that there are costs associated with implementing your data management plan and allow you to explain these costs in your budget narrative. Costs associated with the entire data life cycle should be considered and may include data management personnel costs, fees, infrastructure, or tools needed to organize, document, store, and share study data.[[63]](#footnote-220) Make sure to review your funder’s documentation for information about allowable costs[[64]](#footnote-222). Examples of potential allowable costs include:[[65]](#footnote-224)

* Costs associated with curating and de-identifying data
* Costs associated with developing data documentation
* Fees associated with depositing data for long-term sharing in a repository

It can be difficult to estimate the costs of everything that is associated with the vast landscape of managing data. Luckily a few organizations have developed resources to aid in estimating those costs. The UK Data Service[[66]](#footnote-226), the University of Twente[[67]](#footnote-228), Utrecht University[[68]](#footnote-230), and DataOne[[69]](#footnote-232) have put together checklists to help you think through your various potential data management costs.

You can view slides from this talk [here](https://cghlewis.github.io/mpsi-training1/).

# 5 Planning Data Management

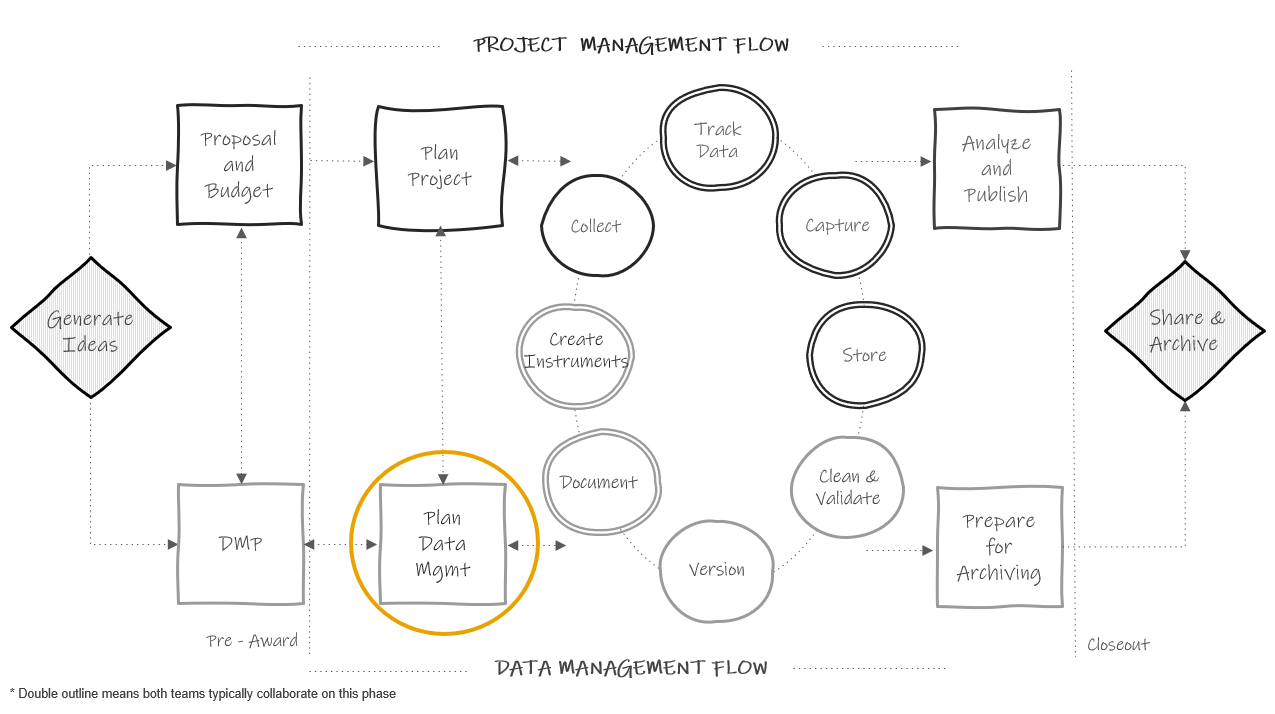


Figure 5.1: Planning in the research project life cycle

Planning data management is distinct from the 2-5 page data management plan (DMP) discussed in the previous chapter. Here we are spending a few weeks, maybe months, meeting regularly with our team and gathering information to develop detailed instructions for how we plan to manage data according to our DMP. This data management planning happens at the same time that the project team is planning for project implementation (things like how to collect data, how to hire staff, what supplies are needed, how to recruit participants, how to communicate with sites, etc). Team members such as PIs, project coordinators, and data managers, may be assisting in both planning processes.

## 5.1 Why spend time on planning?

Funder required data management plans are hopeful outlines for future practices. However, the broad theory behind our DMPs do not actually prepare us for the complex implementation of those plans in practice[[70]](#footnote-240). Therefore, it is important to spend time, before your project begins, planning and preparing for data management. It is an upfront time investment but this sort of slow science leads to better data outcomes. Reproducibility begins in the planning phase. Taking time to create, document, and train staff on data management standards before your project begins helps to ensure that your processes are implemented with fidelity and can be replicated consistently throughout the entire study.

Planning the day to day management of your project data has many other benefits as well. It allows you to anticipate and overcome barriers to managing your data, such as communication issues, training needs, or potential tool issues. This type of planning also saves you time in the long run, removing the last minute scrambling that can occur when trying to organize your data at the end of a project. Last, this type of planning can mitigate errors. Viewing errors as problems created by poorly planned workflows, rather than individual failures, helps us to see how data management planning can lead to better data[[71]](#footnote-242). While data management planning can not remove all chances of errors creeping into your data[[72]](#footnote-243), it can most certainly reduce those errors and prevent them from “compounding over time”(Alston & Rick, 2021, p.4).[[73]](#footnote-244)

## 5.2 Goals of planning

This planning phase should include a series of regular meetings with core decision makers. While there are many aspects of a research study to plan that should also occur during this time (ex: planning an intervention, planning analyses), here we are focusing on planning for data management (including data collection). Most likely you will want the PI/Co-PIs in attendance along with core project staff and your data management team. During this planning time, there are several goals to keep in mind.

1. To finalize project goals laid out in a grant proposal (i.e. what data will be collected)
2. To finalize a timeline for goals (i.e. when is data collected)
3. To lay out specific tasks needed to accomplish goals (i.e. how will data be collected, stored, managed)
4. To assign [roles and responsibilities](#roles) (i.e. who will be responsible for tasks)
5. To make decisions around task management and communication (i.e. how will tasks be monitored and communication tracked)

Make sure to come to every meeting with an agenda to stay on track and to take detailed notes. These notes will be the basis for creating all of your documentation in the next phase. All meeting notes can be stored in a central location such as a planning folder with notes ordered by date or in a running document.

At the end of the planning period, the team should have a clear plan for what the project goals are, when goals should be accomplished, how goals will be accomplished, who is in charge of completing tasks associated with goals, and what additional resources are needed to accomplish goals.

## 5.3 Planning checklists

Along with your existing data management plan, checklists are great tools to help guide your discussions as you work through this planning process with your team. Below are some sample checklists, one for each phase of the research cycle. These checklists can be added to or amended and brought to your planning meetings to help your team think through the various data management decisions that need to be made at each phase of your research project.

* Roles and Responsibilities[[74]](#footnote-247)
* Task Management[[75]](#footnote-249)
* Documentation[[76]](#footnote-251)
* Data Collection[[77]](#footnote-253)
* Data Tracking[[78]](#footnote-255)
* Data Capture[[79]](#footnote-257)
* Data Storage and Security[[80]](#footnote-259)
* Data Cleaning[[81]](#footnote-261)
* Data Sharing[[82]](#footnote-263)

### 5.3.1 Decision-making process

As you move through the remaining chapters of this book, you will begin to learn recommended practices for each phase of the research cycle. Going through each checklist above, you can start to fill in the practices that work for your project for each phase of the study.

This decision-making process is personalized. Borghi and Van Gulick[[83]](#footnote-265) view this process as a series of steps that a research team chooses, out of a the many possibilities not chosen. Maybe you won’t always be able to implement the “best practices” but you can decide what is good enough for your team based on motivations, incentives, needs, resources, skill set, and rules and regulations.

For example, one team may collect survey data on paper because their participants are young children, hand enter it into Excel because that is the only tool they have access to, and double enter 20% because they don’t have the capacity to enter more than that. Another team may collect paper data because they are collecting data in the field, hand enter the data into FileMaker because that is the tool their team is familiar with, and double enter 100% because they have the budget and capacity to do that.

Figure 5.2 is a very simplified example of the decision making process, based on the Borghi and Van Gulick[[84]](#footnote-266) flow chart. Of course in real life we are often choosing between many more than just two options!

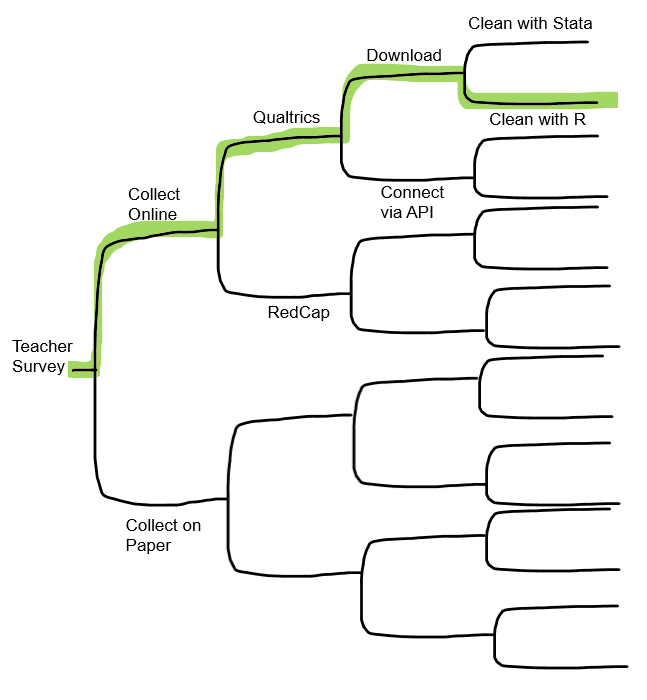


Figure 5.2: A simplified decision-making process

### 5.3.2 Checklist considerations

It’s important to consider how each team and project are unique as you work through these planning checklists. A technique that might work well for one team, may not work out so well for another. Make sure to consider the following:

1. All external requirements
   * Do your practices align with the plan laid out in your DMP? If no, you may need to revise your DMP to match your new decisions - remember your DMP is a living document.
   * Do your practices meet all other external compliance requirements such as those from your Institutional Review Board, your institutional policies, project partner requirements, or government mandates?
2. The skill set of your team
   * How does the skill set of your team align with the practices you plan to implement? Will additional training be required?
3. Your available tools
   * What tools are available to your team?
   * Does your organization only allow you to use certain platforms for data storage?
   * What is the complexity of your tools? Will additional training be needed?
4. Your budget
   * Do you have the budget to implement all of the practices you want to implement or will you need to plan something more feasible?
5. Complexity of your project
   * The size of your project, the amount and types of data you are collecting, the number of participants or the populations you are collecting data from, the sensitivity level of the data you are collecting, the number of sites you are collecting data at, and the number of partners and decision makers you are working with, all factor into your data management planning
6. Shared investment
   * Is your entire team invested in quality data management?
   * Is the entire team motivated to adhere to the standards and instructions laid out in your data management planning? If no, what safeguards can you implement to help prevent errors from creeping into your data?

## 5.4 Data management workflow

The last step of this planning phase is to build your workflows. Workflows allow data management to be seamlessly integrated into your data collection process. Often illustrated with a flow diagram, a workflow is a series of **repeatable** tasks that help you move through the stages of the research life cycle in an “organized and efficient manner”[[85]](#footnote-273). As you walk through your checklists, you can begin to enter your decisions into a workflow diagram that show actionable steps in your data management process. The order of your steps should follow the general order of the data management life cycle (specifically the data collection cycle). You will want to have a workflow diagram for every piece of data that you collect. So for example, if you collect the following three items below, you will have three workflow diagrams.

* Student online survey
* Student paper assessment
* Student district level administrative data

Your diagrams should include the who, what, where, and when of each task in the process. Adding these details are what make the process actionable.[[86]](#footnote-275) Your diagram can be displayed in any format that works for you and it can be as simple or as detailed as you want it to be. A template like the one in Figure 5.3 works very well for thinking through high level workflows. Remember, this is a repeatable process. So while this diagram is linear (steps laid out in the chronological order in which we expect them to happen), this process will be repeated every time we collect this same piece of data.

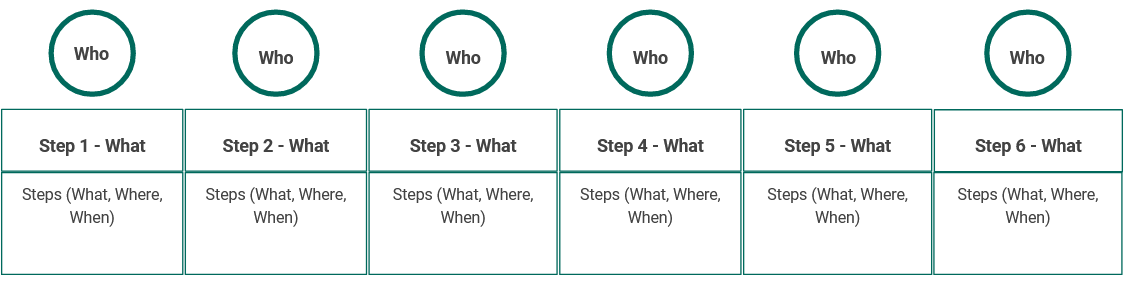


Figure 5.3: A simple workflow template

Here is how we might complete this diagram for a student survey.

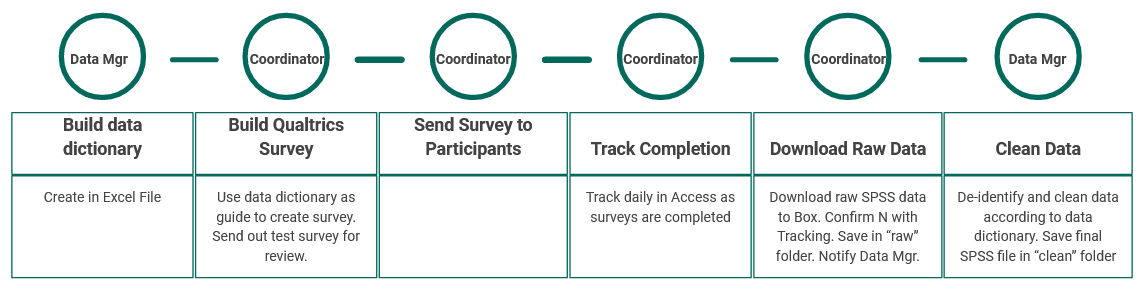


Figure 5.4: Example student survey workflow

But the format truly does not matter. Figure 5.5 is a diagram of the same student survey workflow as above, with more detailed added, and this time using a swimlane template instead, where each lane displays the tasks associated with that individual and the iterative processes that occur within and across lanes.

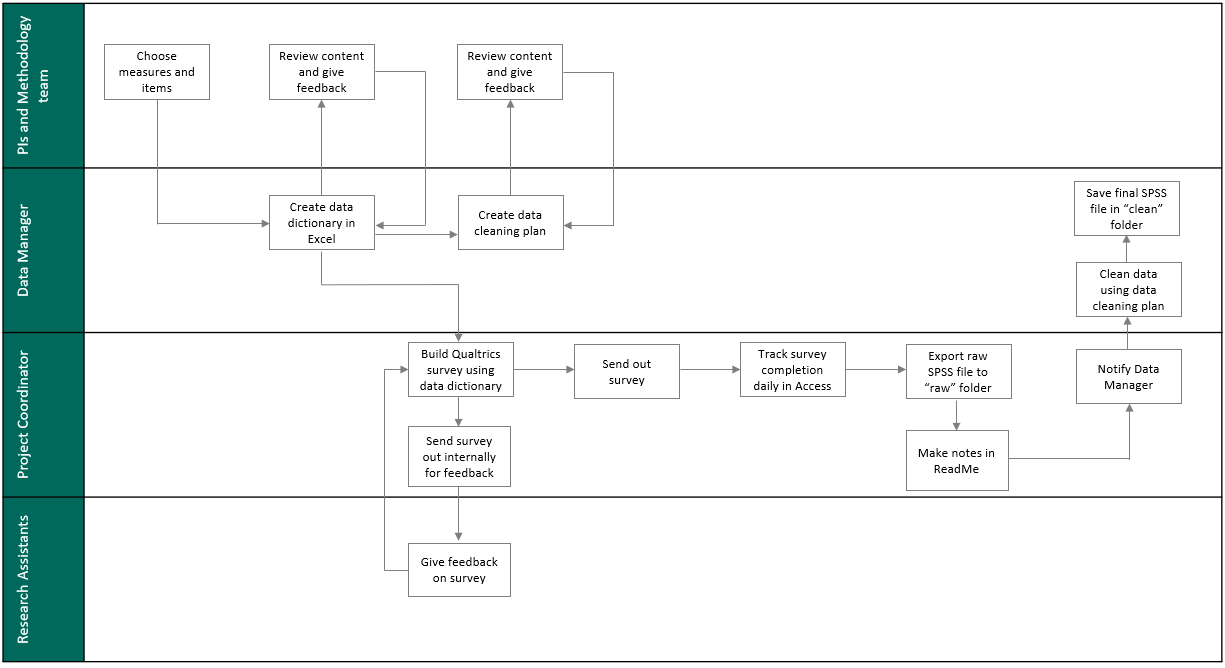


Figure 5.5: Example student survey workflow using a swimlane template

If you have a working [data collection timeline](#supplement) already created, you can even build time into your workflow. Figure 5.6 is another example of the same survey workflow again, this time displayed using a Gantt chart[[87]](#footnote-285) in order to better capture the expected timeline.

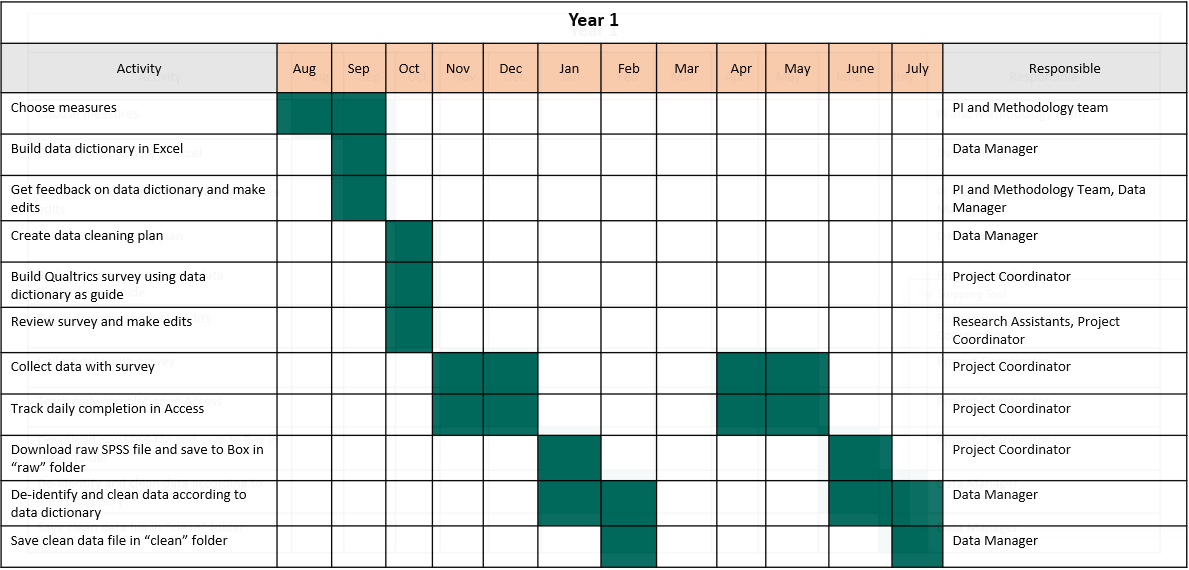


Figure 5.6: Example student survey workflow using a Gantt chart

While these workflow diagrams are excellent for high level views of what the process will be, we can see that we are unable to put fine details into this visual. So the last step of creating a workflow is to put all steps into a standard operating procedure (SOP). In your SOP you will add all necessary details of the process. You can also attach your diagram as an addendum or link your SOPs and diagrams in other ways for reference. We will talk more about creating SOPsin our chapter on [Documentation](#sop).

### 5.4.1 Benefits to visualizing a workflow

Visualizing your decisions in diagram format has many benefits. First, it allows your team to conceptualize their specific tasks in the process, the timing at which their tasks occur, and any dependencies associated with those tasks. It also allows your team to see how their roles and responsibilities fit into the larger research process[[88]](#footnote-290). Showing how data management is integrated into the larger research workflow can help team members view data management as part of their daily routine, rather than “extra work”[[89]](#footnote-292). And last, reviewing workflows as a team and allowing members to provide feedback may help create buy-in for data management processes, potentially leading to better adherence to practices.

### 5.4.2 Workflow considerations

Similar to the questions you need to consider when reviewing your planning checklists, you also need to evaluate the following things when developing your personalized workflow.

* Does your flow preserve the integrity of your data? Is there any point where you might lose or comprise data?[[90]](#footnote-294)
* Is there any point in the flow where data is not being handled securely? Someone gains access to identifiable information that should not have access?
* Is your flow in accordance with all of your compliance requirements (IRB, FERPA, HIPAA, Institutional Data Policies, etc.)?
* Is your flow feasible for your team (based on size, skill level, motivation, etc.)?
* Is your flow feasible for your budget and available resources?
* Is your flow feasible for the amount and types of data you are collecting?
* Are there any bottlenecks in the workflow? Areas where resources or training are needed? Any areas where tasks should be re-directed?

## 5.5 Task management systems

While tools such as our checklists, workflow diagrams, and SOPs allow us to document and share our processes, it can be tricky to manage the day to day implementation of those processes. The planning phase is a great time to choose a task management system[[91]](#footnote-298). Keeping track of various deadlines and communications across scattered sources can be overwhelming and using a task management system may help remove ambiguity about the status of task progress. Rather than having to regularly check in via email for status updates or reading through various meeting notes to learn about decisions made, a task management system allows you to assign tasks to responsible parties, set deadlines based on timelines, track progress, and capture communication and decisions all in one location.

There are many existing tools that allow teams to assign and track tasks, schedule meetings, track project timelines, and document communication. Without endorsing any particular product, some project/task management tools that I know education research teams have used include:

* Trello
* Smartsheet
* Todoist
* Microsoft Planner
* Notion
* Basecamp
* Confluence
* Asana

Of course, as with all processes we’ve discussed so far, a task management system is only useful if your team is trained to use it, is invested in using it, and actually uses it as part of their daily routine. So make sure to consider this as you choose what tool, if any, is right for you.

# 6 Project Roles and Responsibilities

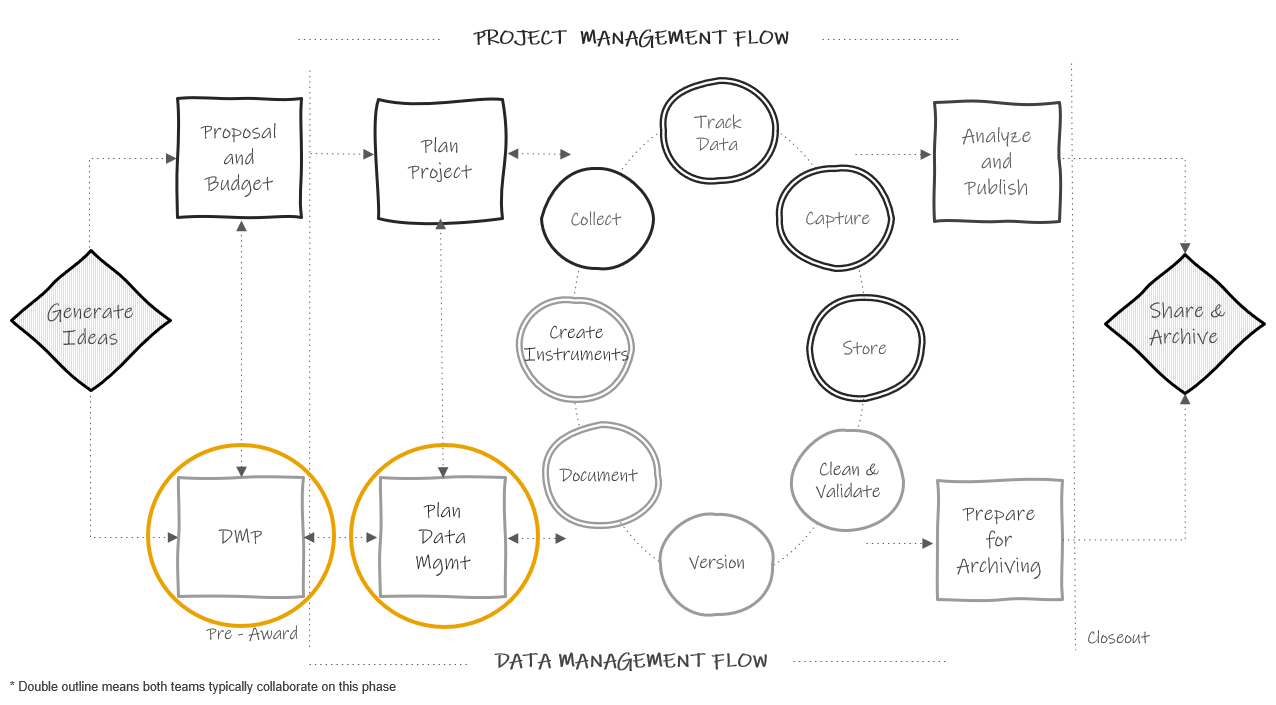


Figure 6.1: Planning in the research project life cycle

Part of the DMP and planning data management phase, as noted in previous chapters, will include assigning roles and responsibilities. In terms of data management, it is important to assign and document roles, not just presume roles[[92]](#footnote-305) for many reasons including for the security of data, the continuity of practices, and for the standardization of workflows.

## 6.1 Typical roles in a research project

Before diving in to how to assign and document roles for a project, it is important to get an understanding of typical roles on an education research project team. Your team may be lucky enough to have all of (or multiple of) these roles. Other times, just one person, such as the Principal Investigator (PI), may take on all or multiple of these roles. With that said, if your budget allows it, I highly recommend hiring individuals to fill each of the roles mentioned below to allow team members to specialize and excel in their area of expertise. While learning all aspects of a project is highly recommended to create a cohesive team that works collaboratively, team members that take on too many project roles can be spread too thin and project goals may suffer.

### 6.1.1 PI and Co-PI

The PIs (or project directors), as well as Co-PIs, are the individuals who prepare and submit the grant proposal and are responsible for the administration of that grant. There are often more than one PI on a project including at least someone with content area knowledge as well as a methodologist. PIs and Co-PIs have varying levels of involvement in research projects and are typically, not always, more hands off in the day to day administration. Even if some tasks are delegated to other research staff, PIs and Co-PIs are ultimately responsible for Institutional Review Board (IRB) submissions and for meeting IRB requirements, as well as for submitting MOUs, budgets, effort reporting, continuing review reports, and any final technical finding reports.

### 6.1.2 Project Coordinator

The project coordinator (or project manager) is an essential member of the research team. As the name implies, this person typically coordinates all research activities and ensures compliance with agencies such as the Institutional Review Board. Tasks they may oversee include recruitment and consenting of participants, creation of data collection materials, creation of protocols, training data collectors, data collection scheduling, and more. The project coordinator may also supervise many of the other research team roles, such as research assistants.

### 6.1.3 Data Manager

The data manager is also an essential member of the team. This person is responsible for the organizing, cleaning, documenting, storing, and dissemination of research project data. This team member works very closely with the project coordinator, as well as the PI, to ensure that data management is considered throughout the project life cycle. Tasks a data manager may oversee include data storage, security and access, building data collection and tracking tools, cleaning and validating data, data documentation, and organizing data for sharing purposes.

This role is vital in maintaining the standardization of data practices. If you do not have the budget to hire a full-time data manager, make sure to assign someone on your team to oversee the flow of data, ensuring that throughout the project, data is documented, collected, entered, cleaned and stored consistently and securely.

### 6.1.4 Project Team Members

This role refers to any staff hired to help implement a research project which may include full-time staff members, with titles such a research or project assistants for instance, or it may include part-time graduate students. Project team members are typically out in the field, collecting data, or they may also assist in other areas such as preparing data collection materials or assisting with data management. Senior project team members may also assist in implementing training or acting as data collection leads in the field.

### 6.1.5 Other Roles

The size of a research team and the roles that exist are dependent on factors such as funding, the type of research study, the intervention being studied, or the organization of your specific research institution. Some teams may include additional roles, not mentioned above, such as research director, lab manager, software engineer, database manager, postdoc, analyst, statistician, administrative professional, hourly data collector, outreach coordinator, or coach/interventionist, all who may assist in the research cycle in other ways. Some of these roles will assist in the research data life cycle as seen in the diagram above. Some may be on a path that is hidden from the diagram but still happening, behind the scenes, alongside the process. Take for instance, the role of a coach implementing an intervention that is being studied. Their tasks aren’t shown on the original diagram but their work is happening alongside the data collection cycle.

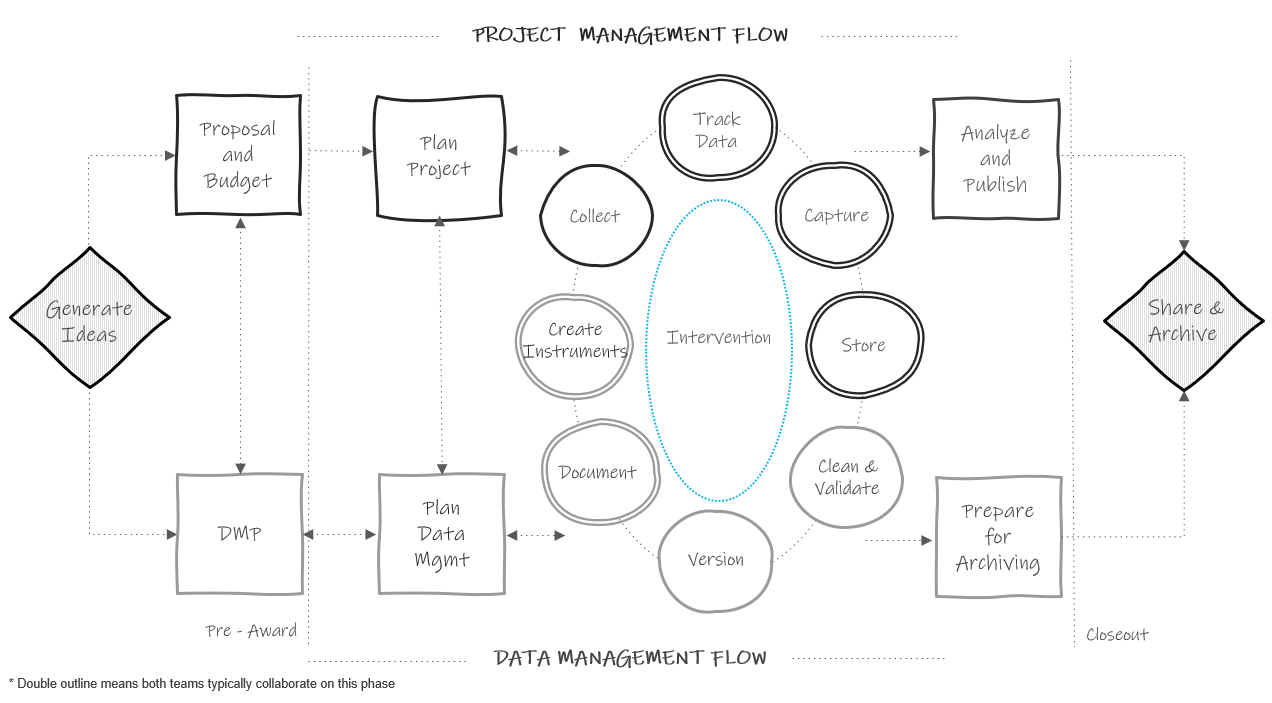


Figure 6.2: Life cycle diagram updated to show hidden processes

## 6.2 Goals of Assigning Roles and Responsibilities

There are several goals of assigning roles and responsibilities.[[93]](#footnote-316)

1. Appoint specific team members to project roles and delineate the responsibilities of those roles
2. Assess equity in responsibilities
   * Including time needed to complete tasks and number of responsibilities assigned to each team member (not overloading any one team member)
3. Assess the skills needed for responsibilities
4. Assess any training need to fill gaps in knowledge
5. Estimate costs associated with roles
6. Develop contingency plans
   * For transitions out of the role or for absenteeism

Early on in the project you will start to generally assign roles in your data management plan. Remember you are often required to state who will be responsible for tasks such as data integrity and security. Then, once your project is funded and you start to have a better idea of your goals and your budget, you can flesh out the details of your roles. During the planning phase, using tools such as your planning checklists will help you think through more specific responsibilities and tasks associated with each role.

## 6.3 Assigning roles and responsibilities

When assigning roles and responsibilities, there are several factors to consider.

1. Required skillset

In assigning roles and responsibilities, make sure to consider the skills that are needed to be successful in each position. For example, when considering the role of a data manager and the responsibilities associated with that role, you may look for skill sets in the following buckets:

* Interpersonal skills (Detail oriented, organized, good communicator)
* Domain skills (Experience working with education data, understands data privacy - FERPA, HIPAA)
* Technical skills (Understanding of database structure, experience building data pipelines, coding experience, specific software/tool experience)

The specific skills needed for each role will depend on your project needs as well as the skill sets of the other members of the team.

1. Training needs

In addition to considering skills needed for certain roles, also consider what training is needed to fulfill assigned responsibilities. In roles that work with data, training may include mandated courses from a program like the Collaborative Institutional Training Initiative (CITI) or it may be signing up for training on how to use a specific tool or software. Make sure that your team members are well-equipped to perform their responsibilities before the project begins.

1. Estimate costs

If you are working on roles and responsibilities after your grant has been funded, then your grant budget has already submitted. However, it can still be helpful to thinking through costs associated with overall roles (based on the experience/skillset of the person filling the role) or even broken down by associated responsibilities (based on things like percent effort or time to complete each task). If discrepancies between the original budget and updated costs are found, often funders will allow PIs to amend budgets.

1. Contingency plans

You should also beging thinking through backup plans should a staff member leave the project or be absent for an extended period of time. This may include cross training staff or a plan for training replacement staff.

## 6.4 Documenting roles and responsibilities

Once roles and responsibilities have been assigned, those decisions should be documented to avoid any ambiguity about who is doing what. While documentation is a topic that will be covered in the next chapter, I think it is helpful to break the rules and discuss just this one document here while we are covering the topic of assigning roles.

A general roles and responsibilities document is one way to document your decisions.[[94]](#footnote-320) In addition to being a planning document that can help you assign tasks to the appropriate staff member, it can also serve as a reference document that allows your team to easily see who is on the project team, what roles they play, and who to contact for questions regarding various project aspects (for example - who to contact for data storage access).

This document can be laid in any format that conveys the information clearly to your team. Figure 6.3 and Figure 6.4 are two example templates. Note that these templates only list overarching responsibilities, not specific tasks. Specific actionable steps will be laid out in other process documentation such as [standard operating procedures](#sop) where names are attached to each task.

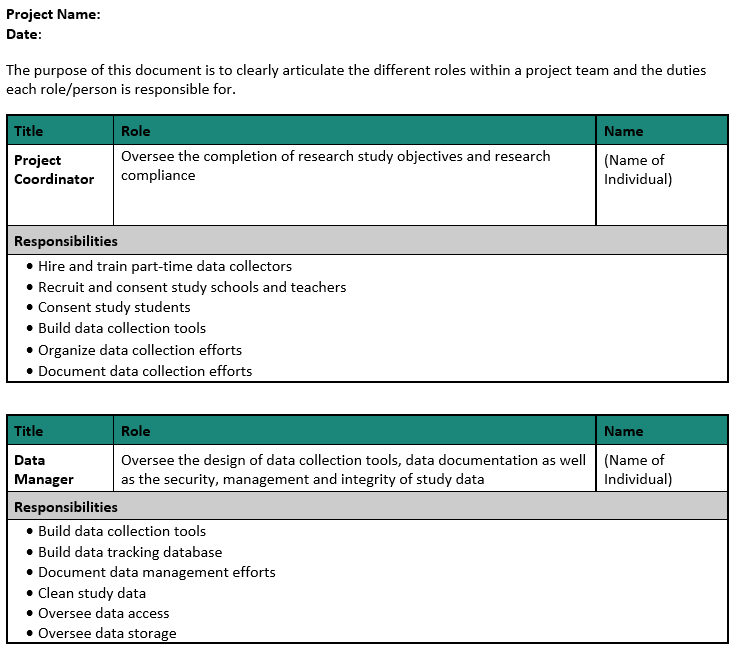


Figure 6.3: Roles and responsibilities document organized by role

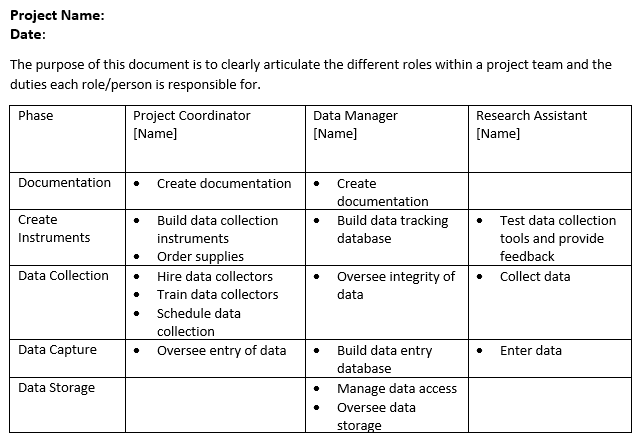


Figure 6.4: Roles and responsibilities document organized by phase

Reviewing roles and responsibilities in a format like this allows you to start to see how responsibilities are assigned and decide if tasks need to be redistributed in any way. You can also then start to fill in more details as needed.

Since there is no one template for creating a roles and responsibility document, you can really add whatever information helps to most clearly convey the information. Some additional columns you may consider adding include:

* Links to related standard operating procedures (ex: for building a participant tracking database you may link to the specific SOP that lays out steps for building the tool)
* Names of other staff members (if any) that assist with or also contribute to each responsibility
* Timing of each responsibility (ex: weekly, ongoing, the month of February)

## 6.5 Data Management Role

Like I mentioned earlier, I highly recommend hiring a full-time data manager if you are able to budget for this as it allows each team member to have more narrow responsibilities and to implement their tasks with better precision. However, not everyone will have the capacity to do this. If so, it will be vitally important to still assign those data management responsibilities to specific team members. In choosing who to assign these tasks to, you will want to consider several things such as appropriate skill set to manage the data, interest in data management tasks, and time to commit to data management. Oftentimes this responsibility falls to a full-time project coordinator as they are the ones who are intimately familiar with the data, and since they are full-time, they are able to carve out hours for data management tasks. Other times it may be a collaboration between a project coordinator and another staff member, such as a part-time graduate student (who may have more technical skills in terms of data wrangling). No matter who you assign these roles to, just ensure that they are documented and the information is disseminated to the team.

# 7 Documentation

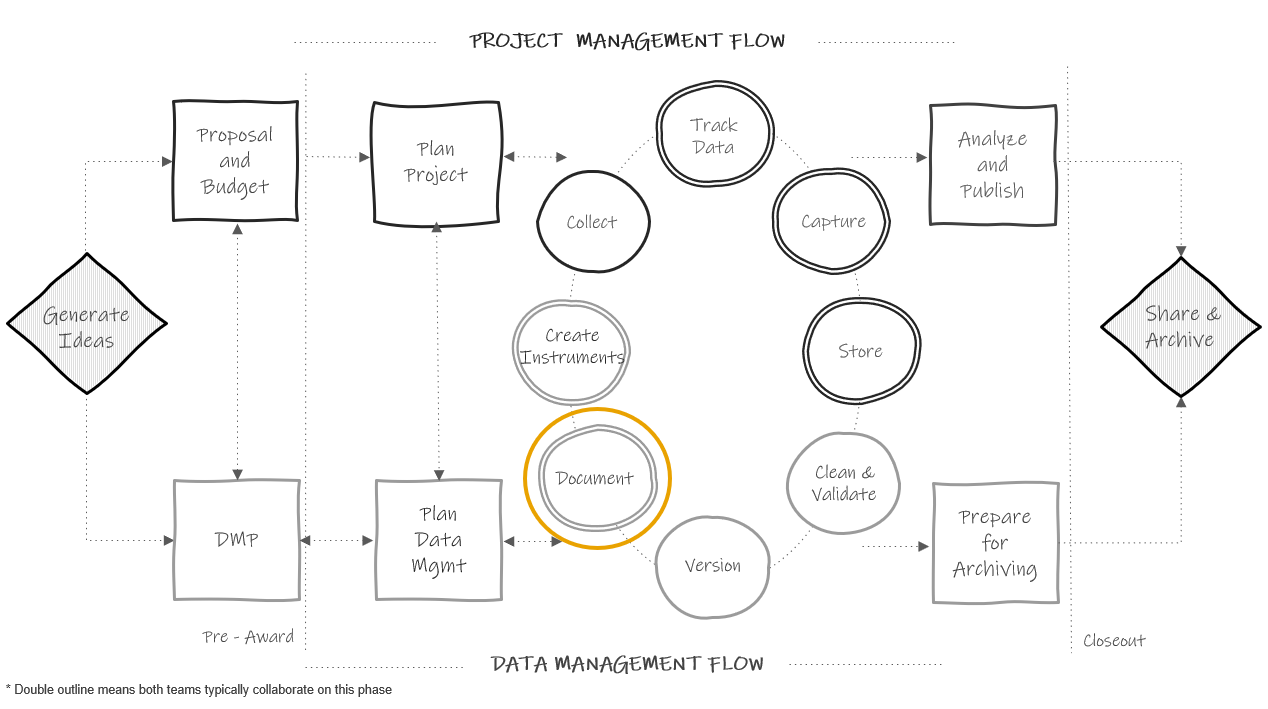


Figure 7.1: Data documentation in the research project life cycle

Documentation is a collection of files that contain procedural and descriptive information about your team, your project, your workflows, and your data. Collecting thorough documentation during your study is unequivocally as important as collecting your data. Documentation serves many purposes including:

* Standardizing procedures
* Securing data and protecting confidentiality
* Tracking data provenance
* Discovering errors
* Enabling reproducibility
* Ensuring others use and interpret data accurately
* Providing searchability through metadata

We are going to cover four levels of documents in this chapter: team level, project level, dataset level, and variable level. While most of the documentation we will discuss does fall within the documentation phase in the research life cycle, some documents will be created earlier or later and the timing will be discussed in each section. During a project, while you are actively using your documents, the format of these documents does not matter. Choose a format that is human-readable and works well for your team (ex: Word, PDF, plain text file, Google Doc, Excel, HTML, OneNote, etc.). When projects are closing out and you are preparing to share your data, you can consider, at that time, how to best make your documents more sustainable, interoperable, and searchable. See the chapter on [Data Sharing](#share) for more information.

The documents below are all recommended and will help you successfully run your project. You can choose to create as many or as few of these documents as you wish. You should choose which documents to create based on what is best for your project and your team, as well as what is required by your funder (see [Data Management Plan](#dmp)), and other governing bodies such as your Institutional Review Board. No matter which documents you choose to implement, it is important that you template your documents and implement them consistently within, and even across projects. Implementing documentation using templates, or consistent formats and fields, reduces duplication in efforts (no need to reinvent the wheel) and allows your team to more easily interpret the document. These documents are best created by the team member that directly oversees the process and sometimes that may include a collaborative effort (for example both a project coordinator and a data manager may build documents together).

Each type of documentation discussed below is a living document to be updated as procedures change or new information is received. As seen in the cyclical section of Figure ?? above, team members should revisit documentation each time new data is collected, or more often if needed. If changes are made and not added to documentation over long periods of time, you will find that you no longer remember what happened and that information will be lost. It will also be important to version your documents along the way so that staff know that they are working with the most recent version and can see when documents have been updated and why.

Creating and maintaining these documents **is an investment**. Make sure to account for this time in your proposal budget. With that said, the return for the investment is well worth the effort.

## 7.1 Team Level

Team level data documentation typically contain data governance rules that apply to the entire team, across all projects. While these documents can be amended at any time, they should really be started long before you apply for a grant, when your lab, center, or institution is formed.

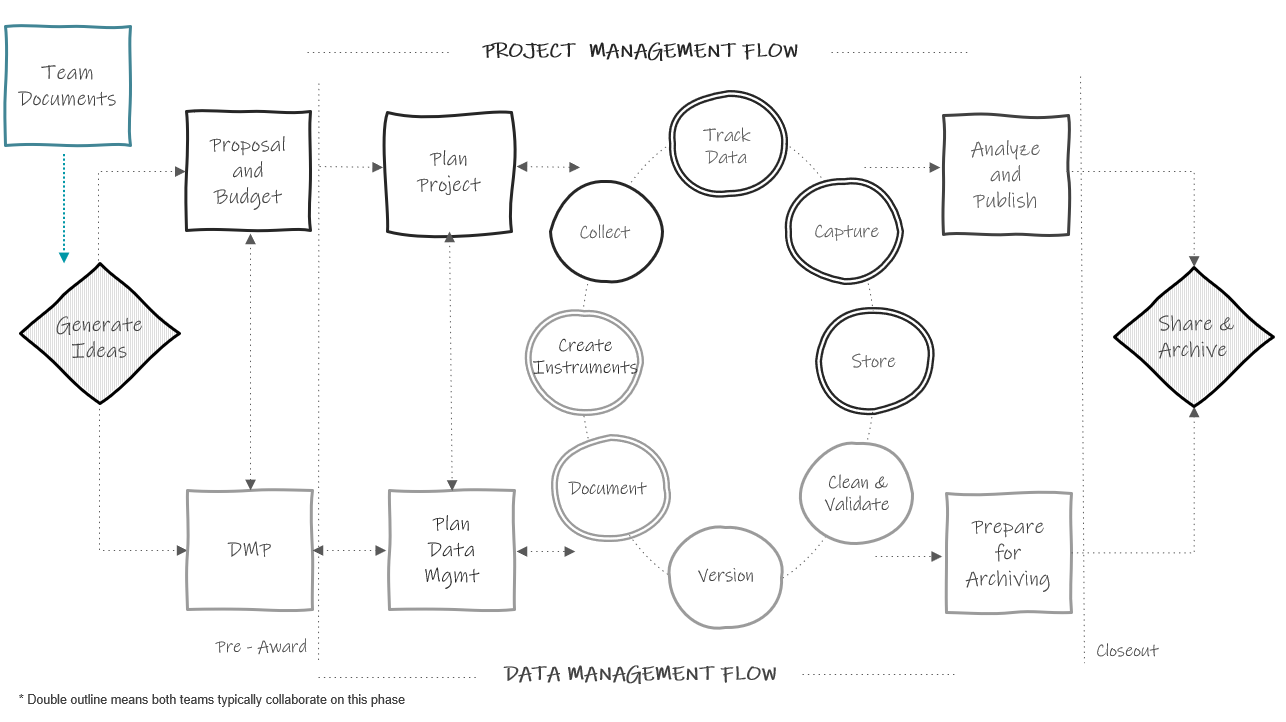


Figure 7.2: Team level documentation in the research project life cycle

### 7.1.1 Lab Manual

One example of a team level document is a lab manual, or team handbook. A lab manual creates common knowledge across your team.[[95]](#footnote-338) It provides staff with consistent information about how the team works and why they do the things they do. It sets expectations, provides guidelines, and can even be a place for passing along helpful career advice.[[96]](#footnote-340) While a lab manual will mostly consist of administrative, procedural, and interpersonal types of information, it can be helpful to also include data management content as well, including general rules about how to access, store, share, and work with data securely and ethically.

**Example lab manuals**

| Document | Description |
| --- | --- |
| Crowdsourced lab manual template[[97]](#footnote-342) | Lab manual template |
| Faylab Lab Manual[[98]](#footnote-343) | Example public lab manual |
| The RAISE Lab Manual[[99]](#footnote-345) | Example public lab manual |
| Common Topics in Lab Handbooks[[100]](#footnote-347) | Common topics covered in lab manuals |

### 7.1.2 Wiki

Another option that can either be created alongside the lab manual or as an alternative to the lab manual is a team wiki. A wiki is a webpage that allows users to collaboratively edit and manage content. It can be created and housed in many tools such as SharePoint, Teams, Notion, GitHub, OSF, and more. While some lab wikis are public (as you’ll see in the examples below), most are not and can be restricted to only invited users. This is a great way to keep disparate documents and pieces of information, for both administrative and data related purposes, organized in a central, accessible location. Your wiki can include links to important documents, or you can also add text directly to the wiki to describe certain procedures. Rather than sending team members to multiple different folders for frequently requested information, you can refer them to your one wiki page.

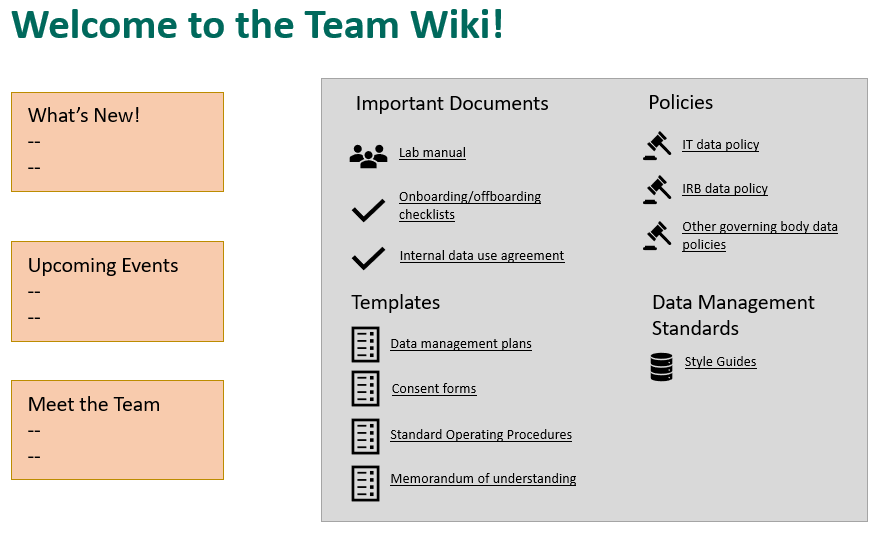


Figure 7.3: Example team wiki with links to frequently requested information

Project level wikis can also be created and be very useful in centralizing frequently referenced information pertaining to specific projects.

**Example wikis**

| Document | Description |
| --- | --- |
| Aly Lab Wiki[[101]](#footnote-353) | Example public lab wiki |
| SYNC Lab Wiki[[102]](#footnote-355) | Example public lab wiki |

### 7.1.3 Onboarding/Offboarding

While **onboarding** checklists will mostly consist of non-data related, administrative information such as how to sign up for an email or how to get set up on your laptop, it should also contain several data specific pieces of information to get all new staff generally acclimated to working with data, for any project, in their new role.

Similarly, while **offboarding** checklists will contain a lot of procedural information about returning equipment and so forth, it should also contain some basic data tasks that help maintain data integrity and security.

Data related topics to consider adding to your onboarding and offboarding checklists are included in Figure 7.4.

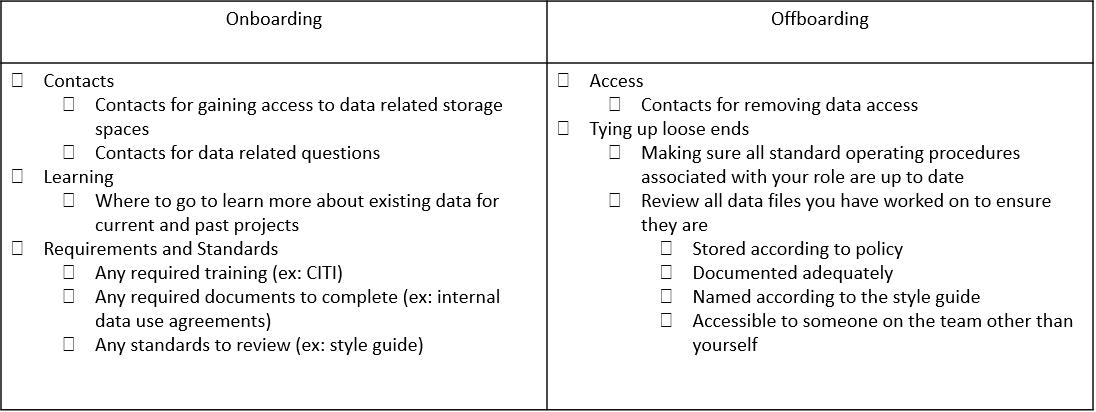


Figure 7.4: Sample data topics to add to onboarding and offboarding checklists

### 7.1.4 Data Use Agreement

Typically when we think of a data use agreement (DUA) we think of a document that we draft in conjunction with an external partner that would like to access our data (or we want to access theirs). It usually covers the terms for how someone is allowed to use data, considering things like access controls, research participant privacy, data destruction rules, and so on.[[103]](#footnote-362) However, it can be really helpful to document the terms and conditions of data use and have staff, at minimum, review or even sign an internal statement saying that they have reviewed all team policies regarding working securely with data.[[104]](#footnote-364) These rules for working with data can be added to a lab manual, as many people do, or they can be added to a separate data use agreement where staff members can sign or check a box acknowledging that they have read and understand the policies.

Ideas of content to include in a DUA are included in Figure 7.5.

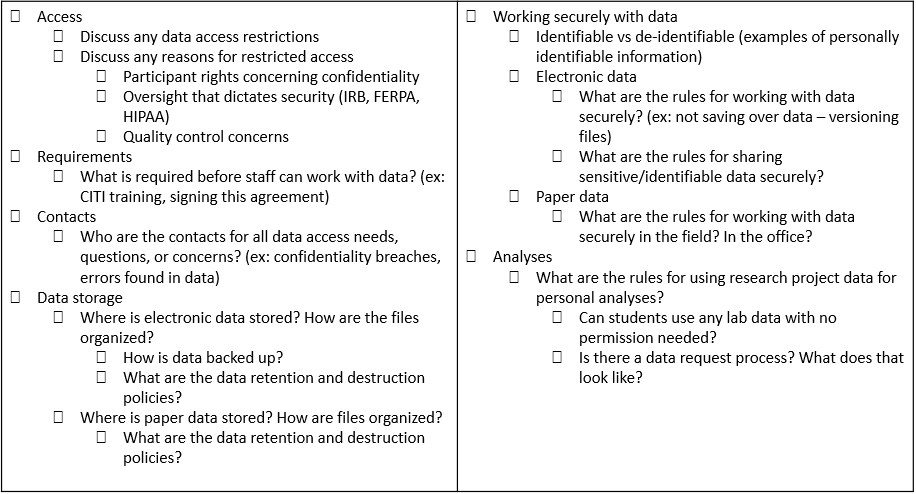


Figure 7.5: Example of content to include in an internal data use agreement

### 7.1.5 Style Guide

A style guide is a set of standards for the formatting of information.[[105]](#footnote-370) It improves consistency within and across files and projects. This document includes conventions for procedures such as variable naming, variable value coding, file naming, versioning, file structure, and even coding practices. It can be created in one large document or separate files for each type of procedure. I highly recommend applying your style guide consistently across all projects, hence why this is included in the team documentation. Since style guides are so important, and there are so many recommended practices to cover, I have given this document its own chapter. See the chapter on [Style Guides](#style) for more information.

**Example style guides**

| Document | Description |
| --- | --- |
| Strategic Data Project style guide[[106]](#footnote-372) | Example style guide for practices including folder structure, file naming, variable naming and coding |
| The tidyverse style guide[[107]](#footnote-374) | Example R coding style guide |

## 7.2 Project Level

Project level documentation is where all descriptive information about your project is contained, as well as any planning decisions and process documentation specifically related to your project. Again, while most of these documents are created in the documentation phase, some documents such as the DMP (started before your project is funded), checklists and meeting notes (started during the planning phase), or the CONSORT diagram (started after data is collected) will begin at other points throughout the cycle.

### 7.2.1 Data Management Plan

This is the first project level document that will be created in your research project life cycle. See [Data Management Plan](#dmp) to review details about this document. The only other note to mention here is that your DMP can continue to be modified throughout your entire study. If any major changes are made, it may be helpful to reach out to your program officer to keep them in the loop as well.

### 7.2.2 Checklists and meeting notes

Since we’ve already discussed these documents in a previous [chapter](#checklist) I won’t say much more here other than to acknowledge that these documents are also part of your portfolio of documentation and are really key planning documents as you start to build your other project as well as data and variable level documents.

### 7.2.3 Roles and responsibilites document

Using the checklists during your planning phase, you hopefully decided on and assigned some roles and responsibilities for your project. Now is the time to formally document those decisions in a way that you can share with the others. In the previous [chapter](#roledoc) we reviewed ways to structure this document. Once this document is created, make sure to store it in a central location for easy referral and update the document as needed.

**Example roles and responsibilities**

| Document | Description |
| --- | --- |
| Johns Hopkins Institute for Clinical and Translational Research Best Practices for Research Data Management[[108]](#footnote-380) | Example roles and responsibilities are assigned throughout this resource |

### 7.2.4 Research Protocol

The research protocol will be a comprehensive project summary document. If you are submitting your study to your Institutional Review Board, you will most likely be required to submit this document as part of your application. A research protocol provides a means for the board to determine if your methods provide adequate protection for human subjects. In addition to serving this required purpose, the research protocol is also an excellent document to deposit along with your data at the time of data sharing, as well as an excellent resource for you when writing technical reports or manuscripts. This document provides all context needed for you and others to effectively interpret and use your data. It generally provides the what, who, when, where, and how of your study. Make sure to follow your university’s specific template if provided, but common items typically included in a protocol are provided in Figure 7.6.

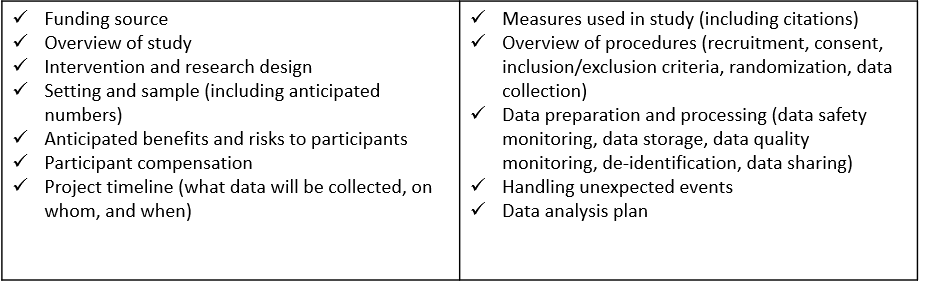


Figure 7.6: Common research protocol elements

When it comes time to submit your data in a repository, the protocol can be revised to contain information helpful for a data end user. Content such as risks and benefits to participants might be removed, and numbers such as final study sample count should be updated to show your final numbers. Additional [supplemental information](#supplement) can also be added as needed.

**Example protocols**

| Document | Description |
| --- | --- |
| University of Washington Protocol Template[[109]](#footnote-386) | Protocol template |
| Ohio State Protocol Template[[110]](#footnote-388) | Protocol Template |
| University of Missouri Protocol Template[[111]](#footnote-390) | Protocol Template |
| SCOPE User Guide[[112]](#footnote-392) | Example data sharing documentation that can be created from a research protocol |

### 7.2.5 Supplemental Documents

There is a series of documents, that while they can absolutely be standalone documents, I am calling supplemental documents here because they can be added to your research protocol as an addendum at any point to further clarify specifics of your project.

1. Timeline

The first supplemental document that I highly recommend creating is a visual representation of your data collection timeline. This can be both a helpful planning tool (for both project and data teams) in preparing for times of heavier and lighter workloads, as well as an excellent document to share with future data users to better understand waves of data collection. There is no one format for how to create this document. Figure 7.7 is an example of one way to visualize a data collection timeline.

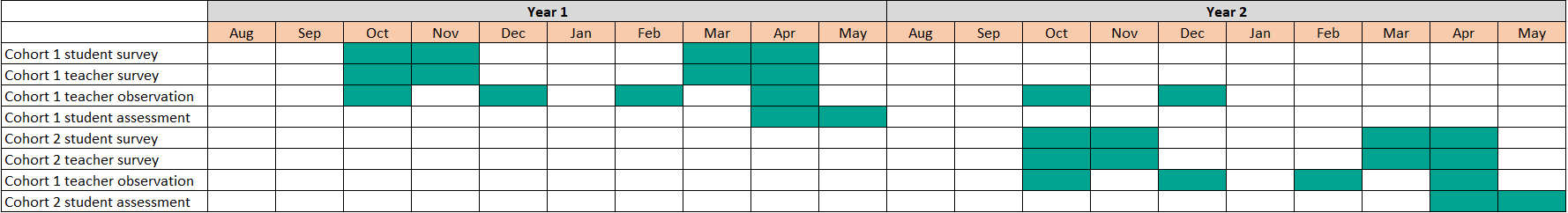


Figure 7.7: Example data collection timeline

1. CONSORT Diagram

A CONSORT (Consolidated Standards of Reporting Trials) diagram displays the flow of participants through a program.[[113]](#footnote-398) It visually portrays enrollment, randomization, as well as attrition in the study. As you can imagine though, this diagram cannot be created until at least one wave of data has been collected, and must be updated as more waves are collected. Your participant tracking database, which we will discuss in our [Data Tracking](#track) chapter, will inform the creation of this diagram.

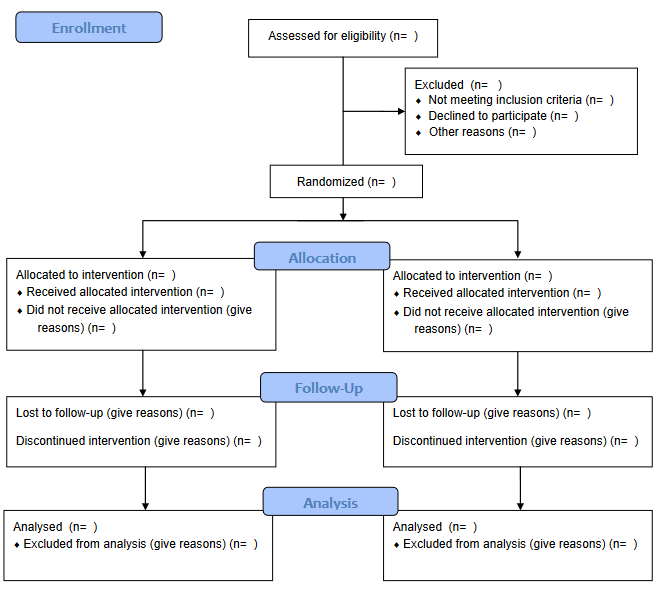


Figure 7.8: 2010 CONSORT flow diagram template

1. Instruments

Actual copies of instrument can be included as supplemental documentation. This includes copies of surveys, assessments, forms, and so forth. It can also include any technical documents associated with your instruments or measures (i.e. a technical document for an assessment or a publication associated with a measure you used).

1. Flowchart of data collection instruments/screeners

You can also include flowcharts of how participants were provided or assigned to different instruments or screeners to help users better understand issues such as missing data.[[114]](#footnote-403)

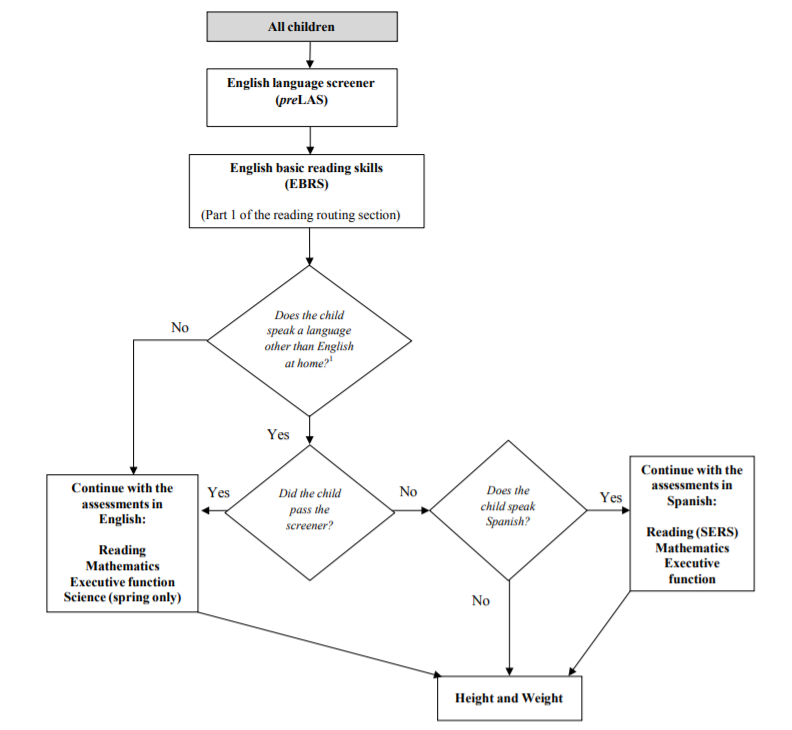


Figure 7.9: Flowchart of an ECLS-K:2011 kindergarten assessment

1. Consent Forms

Consent forms can also be added as an addendum to research protocols to give further insight into what information was provided to study participants.

1. Related publications

You may also choose to attach any publications that have come from your data as an addendum to your protocol.

### 7.2.6 Standard Operating Procedures

While the research protocol provides summary information for all procedures associated with a project, we still need documents to inform how the procedures are actually implemented on a daily basis.[[115]](#footnote-409) If you recall from our [planning chapter](#plan), every step that we added to a data collection workflow is then added to a standard operating procedure (SOP) and the details fleshed out. Not only will you have an SOP for each type of data you are collecting (survey, assessments, observations), you can also have SOPs for other types of decisions as well. Many of the decisions laid out in your protocol, will be further detailed in an SOP. Examples of procedures to include in an SOP are provided in Figure 7.10.

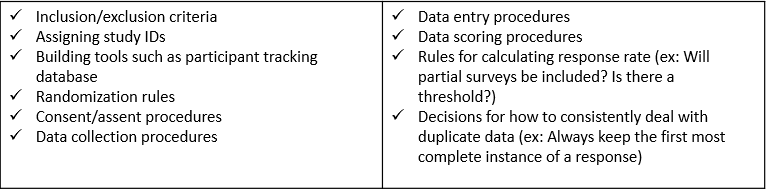


Figure 7.10: Examples of processes or decisions to develop an SOP for

SOPs not only help staff know how to perform tasks, they also create transparency, allow for continuity when staff turnover or go out on leave, create standardization in practices, and last, because an SOP should include versioning information, they allow you to accurately report changes in project procedures throughout the project. You will want to create a template that is used consistently across all procedures, by all staff who build SOPs.

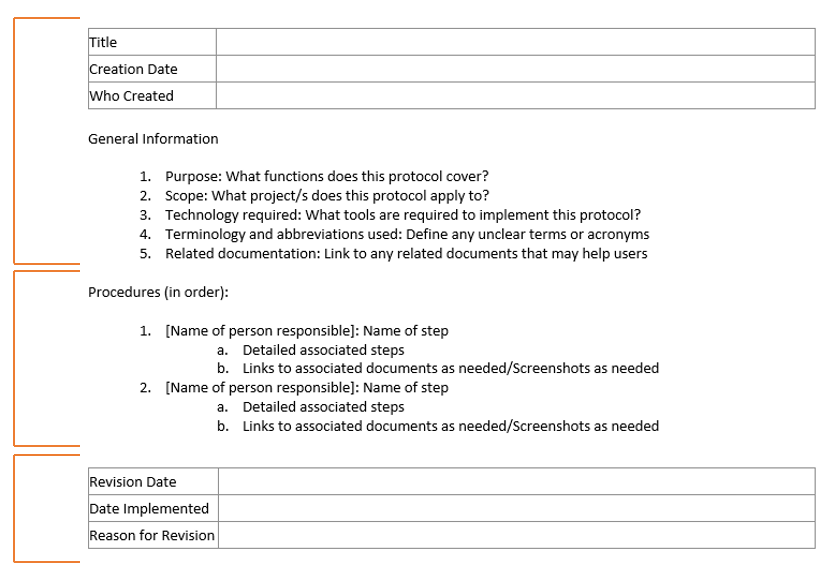


Figure 7.11: Standard operating procedure minimal template

In developing your SOP template, like the one in Figure 7.11, you should begin with **general information** about the scope and purpose of the procedure, as well as any tools or terminology. This provides context for the user and gives them the background to use and interpret the SOP. The next section, **procedures**, lists all procedures in order. Each step provides the name of the staff member/s associated with that step to ensure that there is no ambiguity. Each step should be as detailed as possible so that you could hand your SOP over to any new staff member, with no background in this process, and they can implement the procedure with little trouble. Specifics such as names of files and links to their locations, names of contacts, methods of communication (ex: email vs instant message), and so forth should be included. Additions such as screenshots, links to other SOPs, or even links to tutorials can be embedded as well. Last, any time an update is made to the SOP, clarifying information about the update is added to the **revision** section. This allows you to keep track of changes over time: when were changes made, who made those changes, and why.

**Example SOPS**

| Document | Description |
| --- | --- |
| Northwestern University Writing Standard Operating Procedures and Templates[[116]](#footnote-417) | SOP Template |
| IMPACCT Trials Coordination Centre Standard Operating Procedure for Allocation of Participant Identification Numbers[[117]](#footnote-419) | Example SOP for assigning IDs |
| CITI Template and Sample SOP[[118]](#footnote-421) | Sample Activitiy Monitor Configuration SOP |

## 7.3 Dataset Level

Our next type of documentation applies solely to your datasets and includes information about what data they contain and how they are related. It also includes things such as planned transformations for the data, potential issues to be aware of, and any alterations to the data. In addition to being helpful descriptive documentation to keep, a huge reason for creating dataset documentation is for authenticity. Datasets go through many iterations of processing which can result in multiple versions of a dataset.[[119]](#footnote-425) Preserving data lineage by tracking data errors and transformations is key to ensuring that you know where your data come from, what processing has already been completed, and that you are using the correct version of the data.

Not **all** of your dataset level documentation will be created in the documentation phase and we will talk about the timing as we review each document.

### 7.3.1 Readme

A readme is a plain text document that contains information about your files. These grew out of computer science but are now prevalent in the research world. These documents are a way to convey pertinent information to collaborators in a simple, no frills manner. Readmes can be used in many different ways but I am going to cover three ways they are often used in the context of data management.

1. For conveying information to your colleagues
   * An example of this is if a study participant reaches out to a project coordinator to let them know that they entered the incorrect ID in their survey. When the project coordinator downloads the raw data file to be cleaned by the data manager, they also create a file named “readme.txt” that contains this information and is saved alongside the file in the raw data folder. That way when the data manager goes to retrieve the file, they will see that a readme is included and know to review that document first.
     + ID 5051 entered incorrectly. Should be 5015.
     + ID 5089 completed the survey twice
       - First survey is only partially completed
2. For conveying steps in a process (sometimes also called a setup file)
   * There may be times that a specific data pipeline or reporting process requires multiple steps, opening different files and running different scripts. This information **can** go in an SOP, but if it is a programmatic type process done using a series of scripts, it might be easiest to put a simple file named “readme\_setup.txt” in the same folder as your scripts so that someone can easily open the file to see what they need to run.
   * Step 1: Run the file 01\_clean\_data.R to clean the data  
     Step 2: Run the file 02\_check\_errors.R to check for errors  
     Step 3: Run the file 03\_run\_report.R to create report
3. For providing information about a set of files in a directory
   * If colleagues are accessing your clean datasets in your project directory, it can be helpful to add readmes to the top of those directories to provide information about what datasets are available in the directory, as well as pertinent information about those datasets, including how the datasets are related/can be linked.[[120]](#footnote-427)

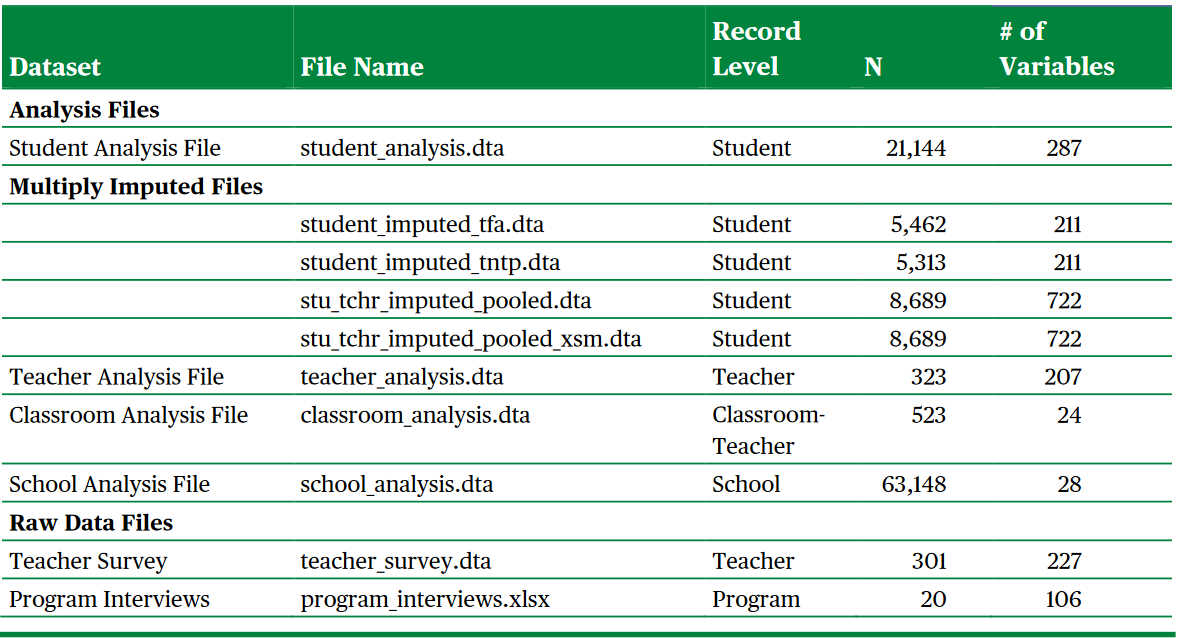


Figure 7.12: Institute of Education Sciences example Readme for conveying information on files in a directory

### 7.3.2 Changelog

A changelog is a record of all of the versions of your data and code. While there are automatic ways to track your data and code through programs such as Git and GitHub, in the field of education where researchers are often working with human subject, identifiable data, users are most often not keeping their study data during an active project, in a remote repository. Data are usually kept in an institution approved storage location. Even if your storage location has versioning such as Box or SharePoint, unless you also have a way to commit messages along with those versions (like a commit message with Git), you will still want to keep a changelog.

A changelog provides data lineage, allowing the user to understand where the data originated as well as all transformations made to the data. It also supports data confidence, allowing the user to understand what version of the data they are currently using and to see if more recent versions have been created and why.

In its simplest form a changelog should contain the file name (versioned consistently), the date it was created, and a description of the dataset (including what changes were made compared to the previous version). It could include additional information as well such as who made the change, and a link to any code used to transform the data.[[121]](#footnote-432)

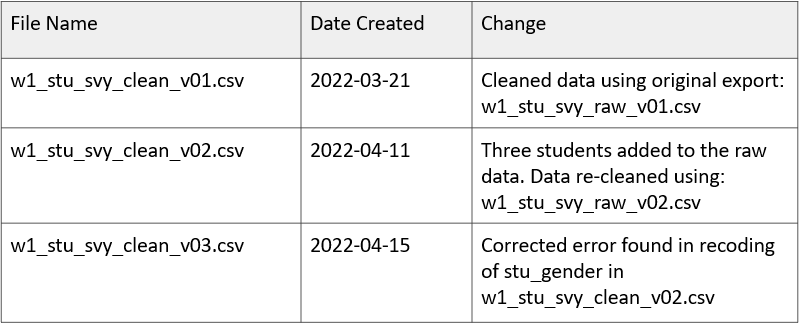


Figure 7.13: Example simple changelog for a clean student survey data file

These changelogs will most likely not be created until the data capture and data cleaning phases of the life cycle when data transformations begin happening, and can be updated at any point as needed.

### 7.3.3 Data Cleaning Plan

A data cleaning plan is a written proposal outlining how you plan to transform your raw data into clean, usable data. This document contains no code and is not technical skills dependent. A data cleaning plan is created for each dataset that you plan to collect (ex: student survey, student assessment, teacher survey, district student demographic data). Because this document lays out your intended transformations for each raw dataset, it allows any team member to provide feedback on the data cleaning process.

This document can be started in the documentation phase, but will most likely continue to be updated throughout the study, especially as you start digging in to your collected raw data and seeing what additional transformations are needed. Typically the person who has the responsibility of cleaning the data will write out the data cleaning plans, but those documents can then be brought to a planning meeting allowing other team members, such as PIs, to provide input on the plan. This ensures that everyone agrees on the transformations to be performed. Once finalized, this data cleaning plan serves as a guide in the cleaning process. We will talk much more about what types of transformations **should** go into a data cleaning plan in the [Data Cleaning](#clean) chapter of this book.

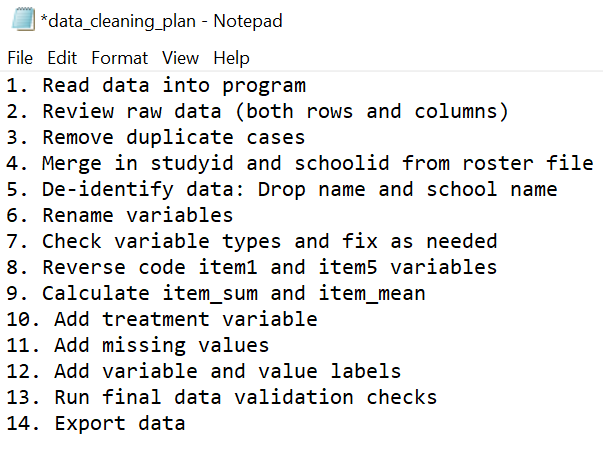


Figure 7.14: A simplistic data cleaning plan

## 7.4 Variable Level

Our last category of documentation is variable level documentation. When we think about data management, I think this is most likely the first type of documentation that pops into people’s minds. This is documentation that tells us all pertinent information about the variables that exist in our datasets: variable names, descriptions, types, and so forth. While variable level documentation is often used for the interpretation of existing datasets, it can also serve many other vital purposes including guiding the construction of data collection instruments, assisting in data cleaning, or validating the accuracy of data,[[122]](#footnote-443) and we will discuss this more throughout the chapters in this book.

### 7.4.1 Data dictionary

A data dictionary is a rectangular format collection of names, definitions, and attributes about variables in a dataset.[[123]](#footnote-445) This document is both a planning tool and a tool used for interpretation, and is most useful if created in the documentation phase, before a project begins, because it is integral to many other phases of a study.

A data dictionary is typically created in a rectangular format. What tool you use to build your data dictionary is up to you, but there are key pieces of information that should be included, as well as optional fields that can be helpful as well.[[124]](#footnote-447)

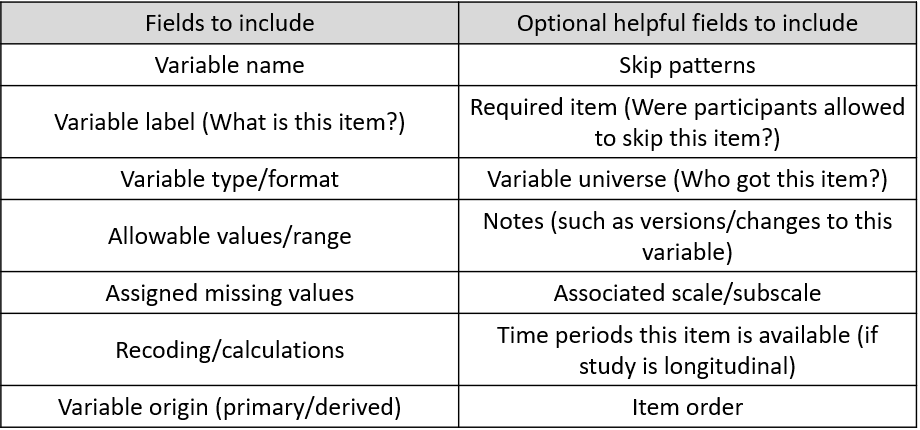


Figure 7.15: Fields to include in a data dictionary

#### 7.4.1.1 Creating a data dictionary for an original data source

Before you begin to build these dictionaries you will need to have the following:

1. Your style guide already created: We will talk more about [style guides](#style) in the next chapter, but this document will provide standards for how you should name variables and code response values.
2. Documentation for your measures: If you are collecting data using existing measures, you will want to collect any documentation on those measures such as technical documents or copies of instruments. You will want your documentation to provide information such as:
   * What items make up the measures/scales? What is the exact wording of items?
   * How are items coded?
   * Are there any calculations/reverse coding needed?

You will then build one data dictionary for each instrument you plan to collect (ex: student survey data dictionary, teacher survey data dictionary, student assessment data dictionary). All measures/items for each instrument will be included in the data dictionary.

As you build your data dictionary, consider the following:

* Are your variable names meeting the requirements laid out in your style guide?
* If your items come from an existing scale, does your value coding align with the coding laid out in the documentation? If your items do not come from an existing scale, does your value coding align with the requirements in your style guide?
* What additional items will make up your final dataset (it could be items that you plan to add to the data after it is collected, i.e. treatment, unique identifiers, calculated variables)?

For demonstration purposes only, the data dictionary in Figure 7.16 uses items from Patterns of Adaptive Learning Scales (PALS).[[125]](#footnote-452) In an actual research study your dictionary would most likely include many more items and a variety of measures.

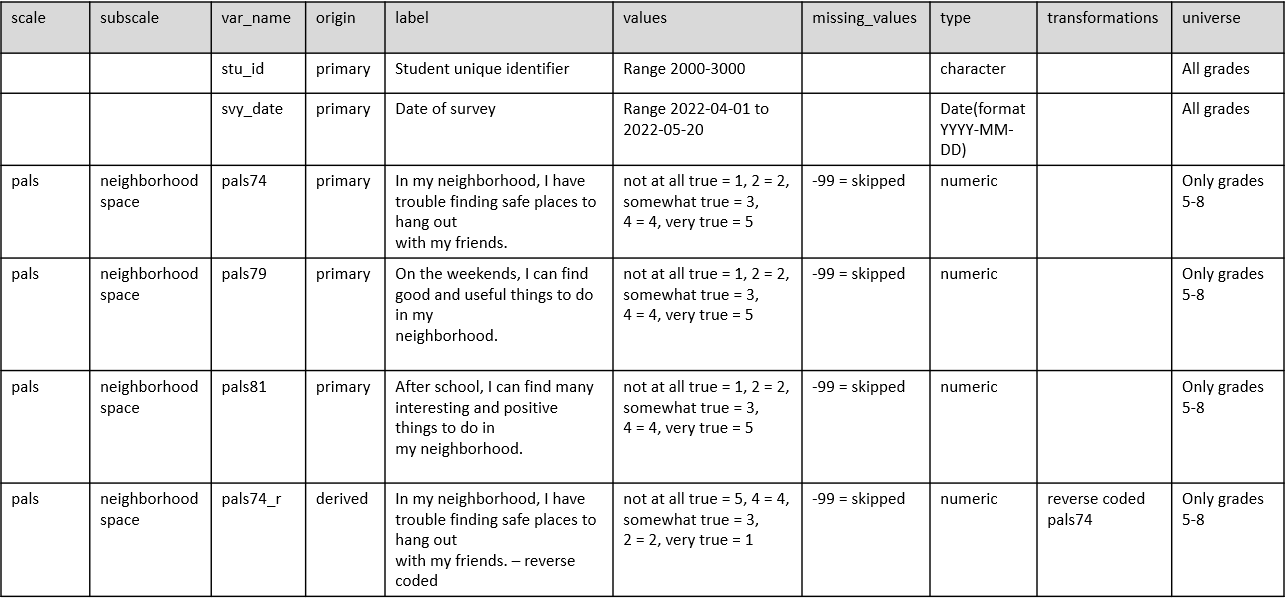


Figure 7.16: Example student survey data dictionary

The last step of creating your data dictionary, as it should be for every document you create in this documentation phase, is to do a review with your team.

* Is everyone in agreement about how variables are named, how values are coded, and our variable types?
* Is everyone in agreement about who gets each item?
* Does the team want to adjust any of the question/item wording?
* You’ll also want to confirm that the data dictionary includes everything the team plans to collect and no items are missing.
* If additional items are added to instruments at later time points, adding fields such as (time periods available), can be really helpful to future users in understanding why some items may be missing data in certain time points.

#### 7.4.1.2 Creating a data dictionary from an existing data source

Not all research study data will be gathered through original data collection methods. You may be collecting external data sources from organizations like school districts or state departments of education. In these cases you will begin building your data dictionaries later in the cycle, when data is received, and rather than the forward moving flow we discussed before where the dictionary is built first, we will now have to work backwards to answer questions about our data.

The first step in building your data dictionary now is to review your existing data. Yet, it turns out that all this tells you is what **does** exist in the data, not what **should** exist in the data. Items could be incorrectly coded, columns could be assigned the incorrect variable type, and so forth. As you review your data, start to collect questions such as:

1. What do these variables represent?
   * What was the wording of these item?
2. Who received the items?
3. What do these values represent?
   * Am I seeing the full range of values/categorical options for each item? Or was the range larger than what I am seeing?
   * Do I have values in my data that don’t make sense for an item?
4. What types are the items currently? What types should they be?

In order to answer those questions, you may need to do some additional detective work.

1. Contact the person who originally collected the data to learn more about the instrument and the data.
2. Contact the person who cleaned the data (if cleaned) to see what transformations they completed on the raw data.
3. Request access to the original instruments to review exact question wording, item response options, skip patterns, and so forth.
4. Request any documentation they have. Do they have their own data dictionaries, codebooks, or syntax that might help you understand what is going on in the data?

Ultimately you should end up with a data dictionary structured similarly to the one above. You may add additional fields that help you keep track of further changes (ex: a column for the old variable name and a column for your new variable name), and your transformations section may become more verbose as the values assigned previously may not align with the values you prefer based on your style guide or the existing measures. Otherwise, the data dictionary should still be constructed in the same manner mentioned above.

#### 7.4.1.3 Time well spent

The process described in this section is a manual, time consuming process. This is intentional. Building your data dictionary is an information seeking journey where you take time to understand your dataset, create standardization of items, and plan for data transformations. Spending time manually creating this document before collecting data prevents many potential errors and time lost fixing data in the future. While there are absolutely ways you can automate the creation of a data dictionary using an existing dataset, the only time I can imagine that being useful is when you have a clean dataset that you have confidently already verified is accurate and ready to be shared. However, a data dictionary, as mentioned before, is so much more than a document to be shared alongside a public dataset. It is a tool for guiding many other processes in your research data life cycle.

**Example data dictionaries**

| Document | Description |
| --- | --- |
| USDA data dictionary template[[126]](#footnote-459) | Example data dictionary template |
| OSF data dictionary template[[127]](#footnote-461) | Example data dictionary template |

### 7.4.2 Codebook

A codebook documents the contents, structure, and layout of a data file[[128]](#footnote-465). It enables the user to quickly ascertain some of the details about a dataset without ever opening the file. Unlike a data dictionary, a codebook is created **after** your data is collected and cleaned and its value lies in data interpretation and data validation.

The codebook contains some information that overlaps with a data dictionary, but is more of a summary document of what actually exists in your dataset.

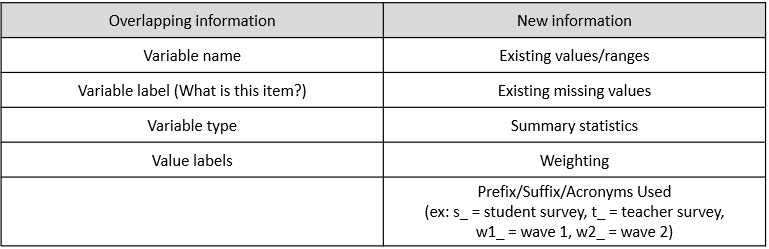


Figure 7.17: Codebook content that overlaps and is unique to a data dictionary

Ultimately you want to export a codebook that contains variable level information like this document in Figure 7.18 from the United States Department of Health and Human Services.[[129]](#footnote-470)

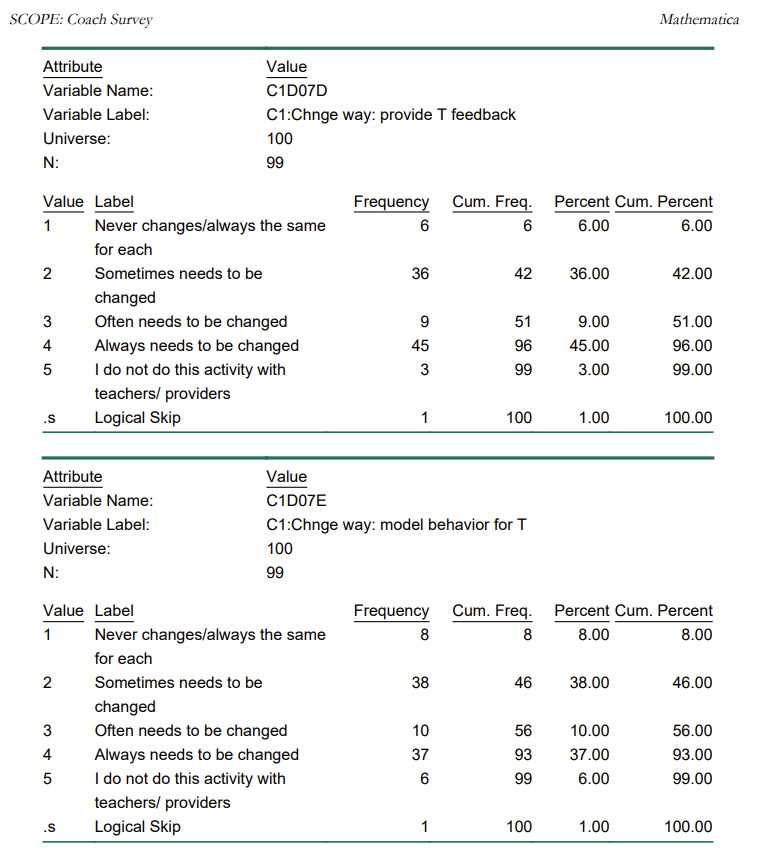


Figure 7.18: Example codebook content from the SCOPE Coach Survey

You can see how in addition to being an excellent resource for users to review your data without ever opening the file, this document may also help you catch errors in your data if out of range or unexpected values appear.

You can either create separate codebooks per datasets or have them all contained in one document, clickable through a table of contents. Unlike a data dictionary which I recommend to create manually, a codebook should be created through an automated process. Automating codebooks will not only save you tons of time, but it will also reduce errors that are made in manual entry. You can use many tools to create codebooks, including point and click statistical programs such as SPSS, or with a little programming knowledge you can more flexibly design codebooks using programs like R or SAS. Fo example, the R programming language has many packages that will create and export codebooks in a variety of formats from your existing dataset by just running a few functions[[130]](#footnote-474).

Last, you may notice as you review codebooks, many will start with several pages of text, usually containing information about the project. It’s common for people, when it comes time to share their data, to combine information from their research protocol or readme files, into their codebooks, rather than sharing separate documents.

**Example codebooks**

| Document | Description |
| --- | --- |
| ICPSR Guide to Codebooks[[131]](#footnote-476) | Guide for creating codebooks |
| SCOPE Codebook, US Department of Health and Human Services[[132]](#footnote-477) | A codebook for this study can be obtained from their ICPSR repository |

## 7.5 Metadata

The last type of documentation to discuss is metadata, which is created in the “prepare for archiving” phase. When it comes time to deposit your data in a repository, you will submit two types of documentation, human-readable documentation, which includes any of the documents we’ve previously discussed, and metadata. Metadata is documentation that is meant to be processed by machines and serves the purpose of making your files searchable.[[133]](#footnote-480) Metadata aids in the cataloging, citing, discovering, and retrieving of data and its creation is a critical step in creating FAIR data[[134]](#footnote-483).

For the most part, no additional work is needed on your part to create metadata when depositing your data in a repository. It will simply be created as part of the depositing process.[[135]](#footnote-486) As you deposit your data, the repository may have you fill out a form that contains descriptive (description of project and files - enables discovery), administrative (licensing and ownership), and structural metadata (technical considerations).[[136]](#footnote-488) The information from this form will become your metadata.[[137]](#footnote-490)

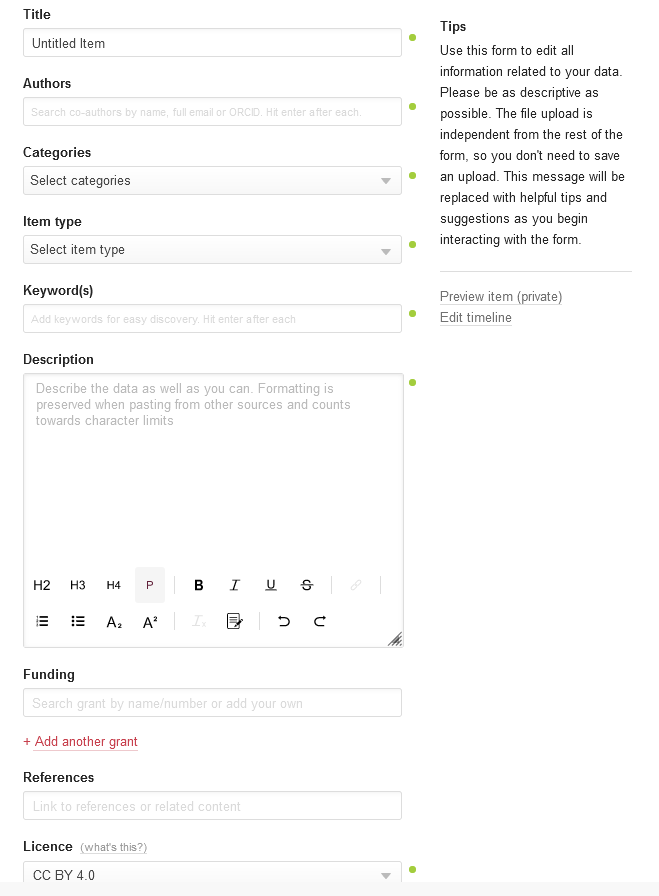


Figure 7.19: Example intake metadata form for Figshare repository

The most common metadata elements[[138]](#footnote-495) are included in Figure 7.20.

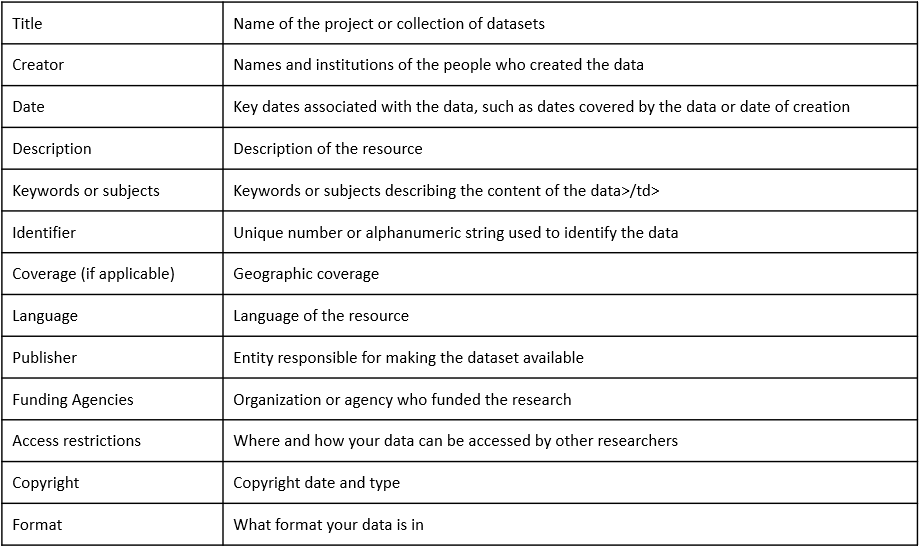


Figure 7.20: Common metadata elements

Depending on the repository, at minimum you will enter basic project level metadata similar to above, but you may be required or have the option to enter more comprehensive information, such as project-level information covered in your research protocol. You may also have the option to enter additional levels of metadata that will help make each level more searchable, such as dataset-level or and variable-level metadata.[[139]](#footnote-500) All of the information needed for this metadata can be gathered from the documents we’ve discussed earlier in this chapter.

Once entered into the form, the repository converts entries into both human-readable and machine-readable, searchable formats such as XML[[140]](#footnote-504) or JSON-LD. We can see what this metadata looks like to humans once it is submitted. Figure 7.21 an example of how ICPSR Open displays the metadata information on a project page.[[141]](#footnote-505) Notice we even have the option to download the XML formatted metadata files in one of two [standards](#metastandards) if we want as well.

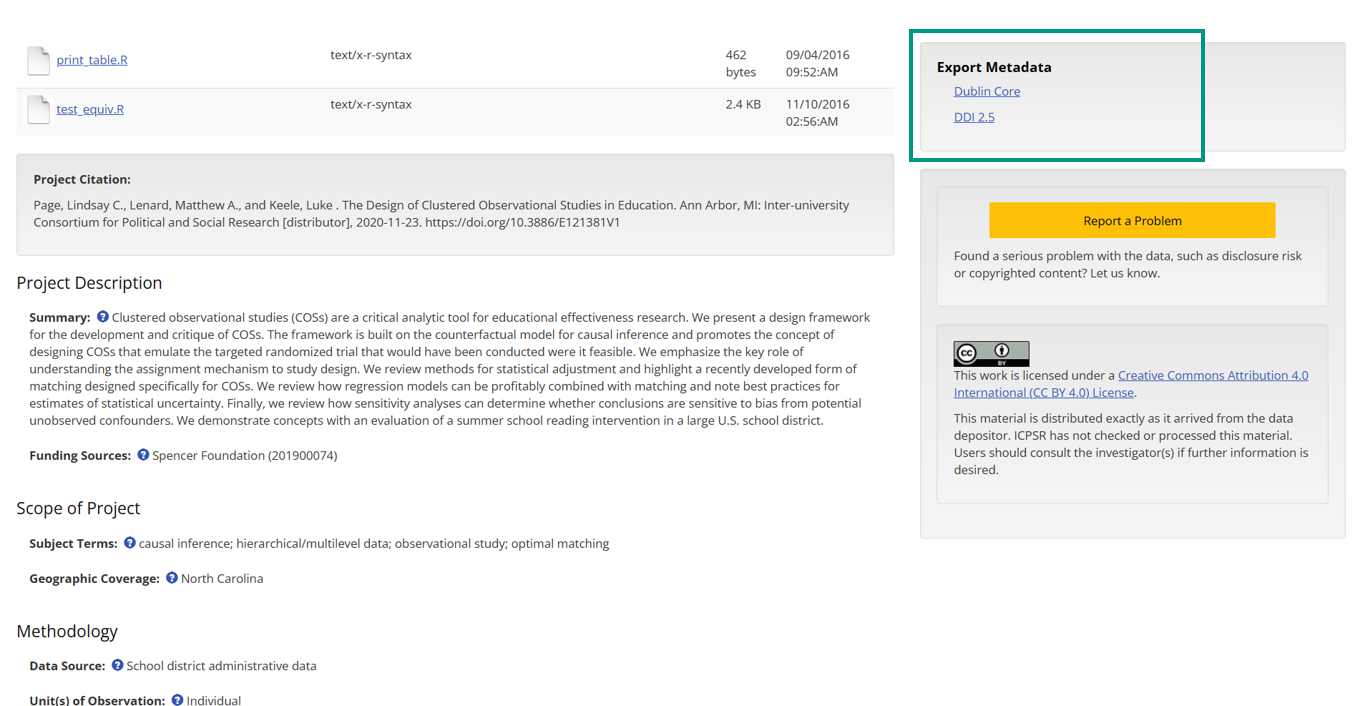


Figure 7.21: Example metadata displayed on an ICPSR Open project page

There are other ways metadata can be gathered as well. For instance, for variable-level metadata, rather than having users input metadata, repositories may create metadata from the deposited statistical data files that contain inherent metadata (such as variable types or labels) or from deposited documentation such as data dictionaries or codebooks.[[142]](#footnote-510)

If your repository provides limited forms for metadata entry, you can also choose to increase the searchability of your files by creating your own machine-readable documents. There are several tools to help users create machine-readable codebooks and data dictionaries that will be findable through search engines such as Google Dataset Search.[[143]](#footnote-511)

### 7.5.1 Standards

Metadata standards, typically field specific, establish common structuring and meaning of data and improve data interoperability in addition to increasing the ability of users to find and understand data.[[144]](#footnote-515) Metadata standards can be applied in several ways.[[145]](#footnote-516)

1. Formats: What machine-readable format should metadata be in?
2. Schema: What fields are recommended verses mandatory for project, dataset and variable level metadata?
3. Controlled vocabularies: A controlled list of terms used to index and retrieve data.

Many fields have chosen metadata standards to adhere to. Some fields, like psychology,[[146]](#footnote-519) are developing their own metadata standards, including formats, schemas, and vocabularies grounded in the FAIR principles and the Schema.org schema.[[147]](#footnote-520) Yet, the Institute of Education Sciences recognizes that there are currently no agreed upon standards in the field of education.[[148]](#footnote-522)

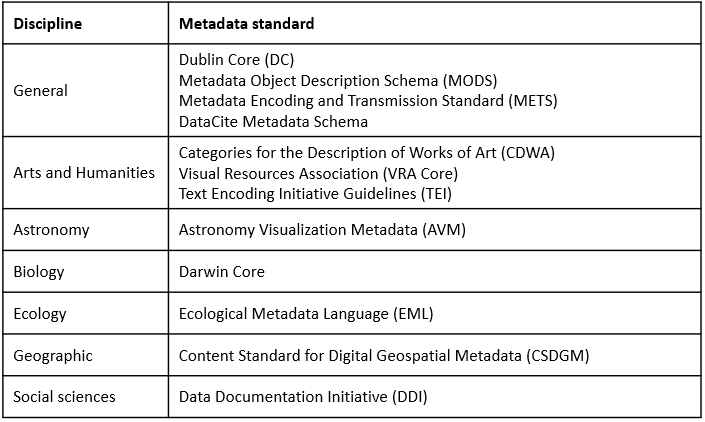


Figure 7.22: A sampling of field metadata standards

It can be helpful to see how standards differ as well as overlap. The DDI Alliance put together this table in Figure 7.23 for instance, mapping the DDI Elements (and vocabularies) to the Dublin Core,[[149]](#footnote-526) two commonly used standards.

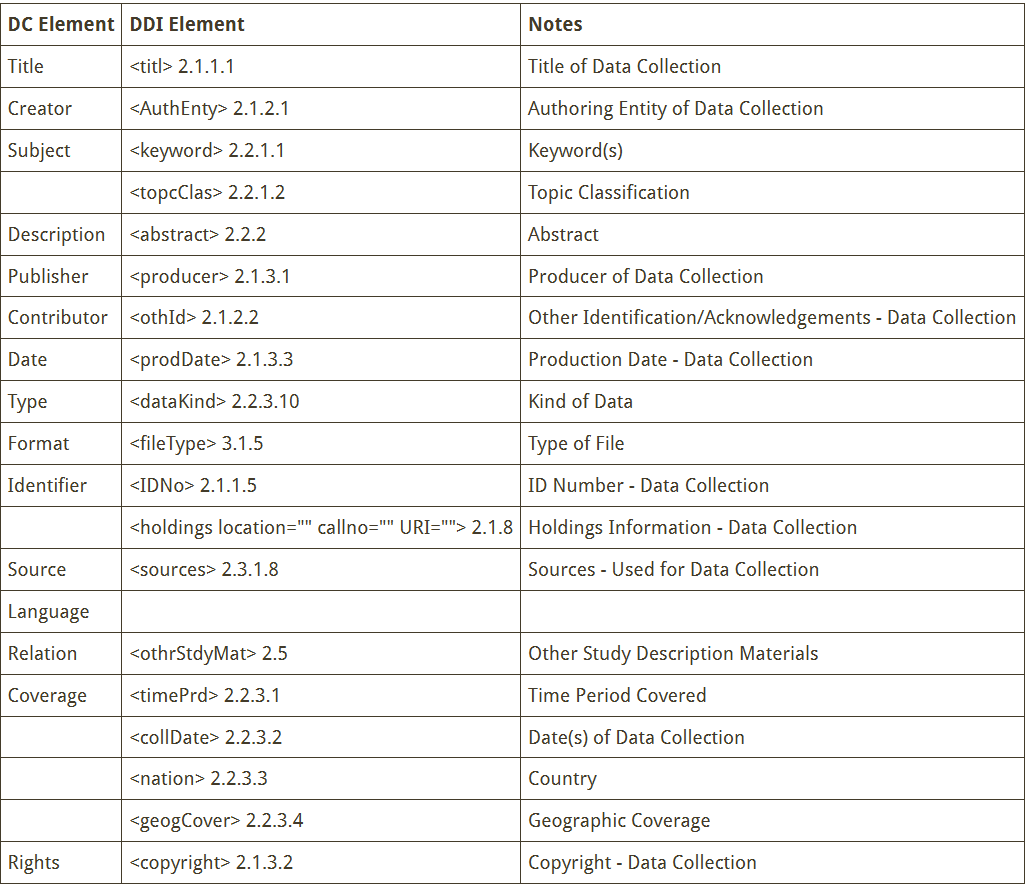


Figure 7.23: A comparison of DDI Version 2 standards to Dublin Core standards

We can see what this metadata comparison actually looks like if we download the Dublin Core and the DDI 2.5 XML format metadata files from the ICPSR Open project we saw above.[[150]](#footnote-531) You can start to see the differences and similarities across standards.

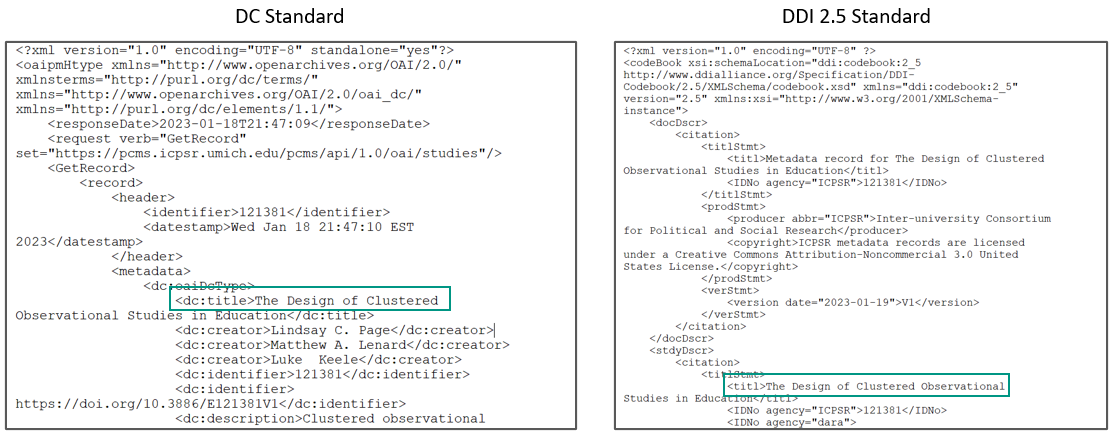


Figure 7.24: Metadata comparison from an AERA Open project

If you plan to archive your data, first check with your repository to see if they follow any standards. For example, the repository Dryad uses a combination of Dublin and Darwin Core,[[151]](#footnote-535) Figshare also uses a combination of schemas including Schema.Org, Dublin Core, and DataCite[[152]](#footnote-537), while ICPSR uses the DDI standard[[153]](#footnote-539). If the repository does use certain standards, work with them to ensure your metadata adheres to those standards. Some repositories may even provide curation support free or for a fee. But as I mentioned earlier, depending on your repository, adding metadata to your project may require no additional work on your part. The repository may simply have you enter information into a form and convert all information for you.

If no standards are provided by your repository and you plan to create your own metadata, you can choose any standard that works for you. Oftentimes researchers may choose to pick a more general standard such as DataCite or Dublin Core,[[154]](#footnote-540) and in the field of education, most researchers are at least familiar with the DDI standard so that is another good option. Remember, if you do choose to adhere to a standard, this decision should be documented in your data management plan.

## 7.6 Wrapping it up

At this point your head might be spinning from the amount of documents we’ve covered. It’s important to understand that while each document discussed provides a unique and meaningful purpose, you don’t have to create every document listed. Choose the documents that help you organize your project and your data processes in the best way. Each document you create that is well maintained, will improve your data management workflow, decrease errors, and enhance your understanding of your data.

# 8 Style guide

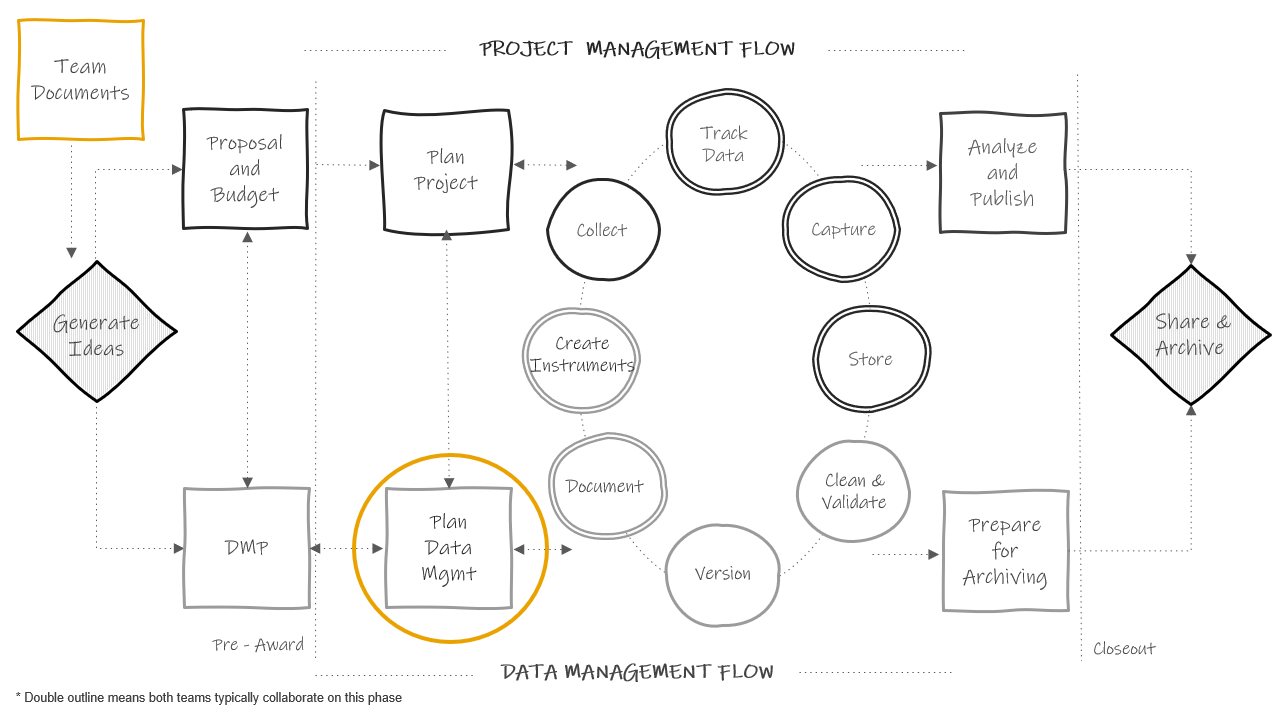


Figure 8.1: Style guide in the research project life cycle

A style guide provides general agreed upon rules for the formatting of information[[155]](#footnote-548). As mentioned in the previous [chapter](#styleshort), style guides can be created to standardize procedures such as variable naming, variable value coding, file naming, file versioning, file structure, and even coding practices.

The benefits for creating style guides and using them consistently include:

* Creating standardization (within and across projects)
* Improving interpretation: Consistent and clear structure, naming, and coding allows your files and variables to be findable and understandable to humans and computers.
* Increasing reproducibility: If the organization of your file paths, file naming, or variable naming constantly change it undermines the reproducibility of any data management or analysis code you have written.

Style guides can be created for individual projects, but they can also be created at the team level, to be applied across all projects. Most importantly, they should be created before a project kicks off so you can implement them as soon as your project begins. If you do not have a team-wide style guide already created, you most likely will want to create a project-level style guide during your planning phase so that you can begin setting up your directory structures and file naming standards before you start creating and saving project-related files.

Style guides can be housed in one large document, with a table of contents used to reference each section, or they can be created as separate documents. Either way, style guides should be stored in a central location that is easily accessible to all team members (such as a team or project [wiki](#wiki)), and all team members should be trained, and periodically retrained, on the style guide to ensure adherence to the rules. If all team members are not consistently implementing the style guide, then the benefits of the guide are lost.

For the remainder of this chapter, we will spend time reviewing some good practices for rules to add to your style guides for the following purposes:

1. Structuring directories
2. Naming files
3. Naming variables
4. Assigning variable values
5. Styling your syntax files

While some best practices will be provided below, ultimately the rules you choose to add to each style guide should be chosen based on which practices work best for your projects and your team. Whatever rules you settle on, write them in a style guide so that everyone is following the same rules within and across projects.

## 8.1 General good practices

Before we dive in to particular types of style guides, there are a few things to understand about how computers read names in order to understand the “why” behind some of these practices.

1. Avoid spaces.
   * While some applications (like Windows) recognize spaces, command line operations and some operating systems still do not support them so it is best to avoid them all together. Furthermore, they can often break a URL when shared
   * The underscore \_ and hyphen - are good delimiters to use in place of spaces
     + It is worth noting that \_ can be difficult to read when file paths are shared in links that are underlined to denote that the path is clickable (for example when sharing a SharePoint link to a document)
2. With the exception of \_ and -, avoid special characters
   * Examples include but are not limited to ?, ., \*, \, /, +, ', &, "
   * Computers assign specific meaning to many of these special characters
3. There are several existing naming conventions that you can choose to add to your style guide. Different naming conventions may work better for different purposes. Using these conventions help you to be consistent with both delimeters and capitalization which not only makes your names more human-readable but also allows your computer to read and search names easier.
   * Pascal case (ScaleSum)
   * Snake case (scale\_sum)
   * Camel case (scaleSum)
   * Kebab case (scale-sum)
   * Train case (Scale-Sum)
4. Character length matters. Computers are unable to read names that surpass a certain character length. This applies to file paths, file names, and variable names. Considerations for each type of limit are reviewed below.

## 8.2 Directory structure

When deciding how to structure your project directories (the organization of your operating systems folders and files), there are several things you want to consider.

**When structuring your folders:**

* First, consider organizing your directory into a hierarchical folder structure to clearly delineate segments of your projects and improve searchability
  + The alternative to using a folder structure is using metadata and tagging to organize and search for files[[156]](#footnote-550)
* When creating your folder structure, strike a balance between a deep and shallow structure
  + Too shallow leads to too many files in one folder which is difficult to sort through
  + Too deep leads to too many clicks to get to one file, plus file paths can max out with too many characters. A file path includes the full length of both folders and file names
    - An example file path with 69 characters W:\team\project\data\wave1\student\survey\projecta-w1-stu-svy-raw.csv
  + Examples of file path limits:
    - SharePoint/OneDrive path limit is 400 characters[[157]](#footnote-553)
    - Windows path limit is 260 characters[[158]](#footnote-555)
* Create folders that are specific enough that you can limit access
  + For example you will want to limit user access to folders that hold Personally Identifiable Information (PII)
  + To protect any files that you don’t want others to accidentally edit (for example your clean datasets), also consider making some files “read only”
* Decide if you want an “archive” folder to move old files into or if you want to leave previous versions in the same folder

**When naming your folders:**

* Consider setting a character limit on folder names (again to reduce problems with hitting path character limits)
* Make your folder names meaningful and easy to interpret
* Never use spaces in your folder names
  + Use \_ or - to separate words
* With the exception of - and \_, don’t use special characters in your folder names
* Be consistent with delimiters and capitalization. Follow an existing naming convention (as mentioned above).

**Example directory structure style guide**

1. All project directories follow this hierarchical metadata structure   
 - Level 1: Name of project   
 - Level 2: Life cycle folders   
 - Level 3: Data collection wave folders (if relevant)   
 - Level 4: Participant folder (if relevent)  
 - Level 5: Specific content folder   
 - Level 6: Archive folders   
2. All folders should be named according to these rules   
 - Meaningful name but no longer than 20 characters   
 - No spaces or special characters in folder names   
 - Only use lower case letters   
 - Use `-` to separate words   
3. All previous versions of files must be placed into their respective "archive" folder  
 - A changelog should be placed in each "archive" folder to document changes between document versions

**Example directory structure created using a style guide**

levelName  
1 project-new   
2 ¦--intervention   
3 ¦ °--cohort-1   
4 ¦ °--coaching\_materials   
5 ¦ °--archive   
6 ¦ °--changelog.txt   
7 ¦--project-mgmt   
8 ¦ °--cohort-1   
9 ¦ °--scheduling-materials   
10 ¦ °--archive   
11 ¦ °--changelog.txt   
12 ¦--documentation   
13 ¦ ¦--sops   
14 ¦ ¦ °--archive   
15 ¦ ¦ °--changelog.txt   
16 ¦ °--data-dictionaries   
17 ¦ °--archive   
18 ¦ °--changelog.txt   
19 ¦--data   
20 ¦ °--cohort-1   
21 ¦ °--student   
22 ¦ °--survey   
23 ¦ °--archive   
24 ¦ °--changelog.txt  
25 °--tracking   
26 °--cohort-1   
27 ¦--participant-database   
28 ¦ °--archive   
29 ¦ °--changelog.txt   
30 °--parent\_consents

## 8.3 File naming



Figure 8.2: xkcd comic on naming files

As xkcd[[159]](#footnote-561) so aptly points out in Figure 8.2, many of us are pretty bad at naming files in a consistent and usable way. We are often in a rush to save our files and maybe don’t consider how unclear our file names will be for future users (including ourselves).

Our file names alone should be able to answer questions such as:

* What are these documents?
* When were these documents created?
* Which document is the most recent version?

A file naming style guide helps us to name files in a way that allows us to answer these questions. You can have one overarching file naming guide, or you may have file naming guides for different purposes that need different organizational strategies (for example one naming guide for project meeting notes, another naming guide for project data files). Let’s walk through several conventions to consider when naming your files.

* Make names descriptive (a user should be able to understand the contents of the file without opening it)
* Never use spaces between words
  + Use - or \_ to separate words.
* With the exception of \_ and -, never use special characters
* Be consistent with delimiters and capitalization. Follow an existing naming convention.
* Consider limiting the number of allowable characters to prevent hitting your path limit (as mentioned above)
* Do not use / in dates and format them consistently. It is beneficial to format dates using the ISO 8601 standard in one of these two ways:
  + YYYY-MM-DD or YYYYMMDD
  + While the first way adds more characters to your variable names, it also may be clearer for users to interpret. Either of these date formats will be sortable.
* When versioning your files, pick a format and add it to your style guide
  + If you plan to version using a number, consider left padding with 0 before single digit numbers to keep the file name the same length as it grows (v01, v02).
  + As mentioned in our chapter on Documentation, it is possible to version programatically using tools like Git and GitHub. However, these tools are not always practical for education research. A more practical means of versioning may be to manually version files and track changes in a [changelog](#change).
* If your files need to be run in a sequential order, add the order number to the beginning of the file name, with leading zeros to ensure proper sorting (01\_, 02\_)
* Choose abbreviations and/or consistent terms to use for common names/phrases and add them to your style guide (student = stu).
  + This helps reduce file name character lengths and also creates standardized, searchable metadata, which can allow you to more easily, programmatically retrieve files (for example, retrieve all files containing the phrase “stu\_obs\_raw”).
* Keep redundant metadata in the file name
  + This reduces confusion if you ever move a file to a different folder or send a file to a collaborator. It also makes your files searchable.
  + For example, always put the data collection wave in a file name, even if the file is currently housed in a specific wave folder. Or always put the project name in the file name, even if the file is currently housed in a project folder.
* Choose an order for file name metadata (ex: project -> time -> participant -> measure)

**Example file naming style guide**

1. Never use spaces between words  
2. Never use special characters  
3. Use \_ to separate words  
4. Only use lower case letters  
5. Keep names under 35 characters  
6. Use the following metadata file naming order:  
 - Order of use (if relevant–and always add a 0 before single digits)  
 - Project  
 - Cohort/Wave (if relevant)  
 - Participant  
 - Measure  
 - Further description  
 - Date (always add)  
 - Version (if necessary)  
7. Format dates as YYYY-MM-DD  
8. If there are multiple versions of a document on the same date, version using v# with a leading 0.  
9. Use the following abbreviations  
 - student = stu  
 - survey = svy  
 - wave = w  
 - project math efficacy = me

**Example file names created using a style guide**

me\_stu\_svy\_sop\_2022-08-01.docx  
me\_w1\_stu\_svy\_data\_raw\_2022-11-03.csv  
me\_w1\_stu\_svy\_cleaning\_syntax\_2023-01-22v01.R  
me\_w1\_stu\_svy\_cleaning\_syntax\_2023-01-22v02.R

## 8.4 Variable naming

This style guide will be a necessary document to have before you start to create your data dictionaries. Below are several considerations to review before developing your variable naming style guide. These are broken into two types of rules, those that are non-negotiable requirements that really should be included in your style guide (if you do not follow these rules you will run into serious problems in interpretation for both humans and machines), and then best practices suggestions that are recommended but not required.

**Mandatory:**

* Don’t name a variable any keywords or functions used in any programming language (such as if, for, repeat)[[160]](#footnote-564)
* Set a character limit
  + Most statistical programs have a limit on variable name characters
    - SPSS is 64
    - Stata is 32
    - SAS is 32
    - Mplus is 8
    - R is 10,000
  + With this said, do not limit yourself to 8 characters based on the fact that one future user may use a program like Mplus. Consider the balance between character limit and interpretation. It is very difficult to make good human-readable variable names under 8 characters. It is much easier to make them under 32. And the majority of your users will be using a program with a limit of 32 or more. If you have one potential Mplus user, they can always rename your variables for their specific analysis.
* Use the same variable name across time in a project
  + If an item is named anx1 in the fall, name that same item anx1 again in the spring
* Don’t use spaces or special characters (except\_), they are not allowed in most programs
  + Even the - is not allowed in programs such as R and SPSS as it can be mistaken for a minus sign
  + While . is allowed in R and SPSS it is not allowed in Stata so it’s best to avoid using it
* Do not start a variable name with a number. This is not allowed in many statistical programs.
* All variable names should be unique
  + This absolutely applies to variables within the same dataset, but it should also apply to all variables across datasets within a project. The reason is, at some point you may merge data across forms and end up with identical variable names (which programs will not allow).
  + So, for example if you collect student gender from a survey and you also collect student gender from school records, differentiate between the two (s\_gender and d\_gender)
* If you substantively change an item (substantive wording OR response options change) after at least one round of data has been collected, version your variable names in order to reduce errors in interpretation.
  + For example revised anx1 becomes anx1\_v2

**Suggested:**

* Names should be meaningful
  + Instead of naming gender q1, name it gender
  + If a variable is a part of a scale, consider using an abbreviation of that scale plus the scale item number (anx1, anx2, anx3)
    - Not only does this allow you to easily associate an item with a scale, but it also allows you to programatically select and manipulate scale items (for example, sum all items that start with “anx”)
* If you have used the question/scale before, consider keeping the variable name the same across projects. This can be very useful if you ever want to combine data across projects.
* Be consistent with delimiters and capitalization. Follow an existing naming convention.
  + Snake case (scale\_sum) – preferred method for variable names
  + Kebab case (scale-sum) – don’t use for variable names
  + Train case (Scale-Sum) – don’t use for variable names
* Consider denoting reverse coding in the variable name to reduce confusion (anx1\_r)
* Choose abbreviations and standard phrases to use across all variables. Using controlled vocabularies improves interpretation and also makes data exploration and manipulation easier[[161]](#footnote-567).
  + mean = mean
  + scaled score = ss
  + percentile rank = pr
* Include an indication of the measure in the variable name (for example as a prefix) so you always know what instrument the item came from. This can also help with the unique variable name requirement above.
  + s = student self-report
  + t = teach report on students
  + s\_anx1, t\_conf2

**Example variable naming style guide**

1. Use snake case  
2. Keep names under 32 characters  
3. Use meaningful variable names  
4. If part of a scale, use scale abbreviation plus item number from the scale (not order number)  
5. Include an indication of the measure as a prefix in the variable name  
 - student self-report survey = s\_  
 - teacher self-report survey = t\_  
 - district student level data = d\_  
6. Denote reverse coded variables using suffix `\_r`

**Example variable names created using a style guide**

s\_anx1  
s\_anx1\_r  
s\_gender  
d\_gender  
t\_stress5

### 8.4.1 Time

Before moving on there is one last consideration for variable names. If your data is longitudinal, you may need to add rules for accounting for time in your variable names as well.

Depending on how you plan to combine your data over time, there are two different ways to account for time.

1. Concatenate time to your variable names. You do this if you plan to merge your data across time in [wide format](#structure). The reason you need to concatenate time to your variable names here is because your variable names will repeat (anx1 in wave 1, anx1 in wave 2). And remember from our guidelines above, all variable names in a dataset **must** be unique. In order to create unique variable names and correctly interpret when items were given, we add time to our variable names. The only variables you will not assign time to are your [linking variables](#columns) (such as student unique identifier, teacher unique identifier, and so on). Those variables need to stay identical for linking purposes and will only appear once in your data after merging.
2. Create time variables and add them to your data. You do this if you plan to append your data over time in [long format](#structure). Appending your data in long format requires no additional work in terms of variable naming. As discussed in our data structure chapter, you actually want your variables to be identically formatted and named across time when appending. So here, in order to differentiate when items were asked, we add a new variable such as time or wave and add the appropriate value for each row.

During an active project, it is actually best to not combine data and to store all datasets as distinct files until you are either ready to internally use your data or you are ready to publicly share your data (during the preparing for archiving phase). At that time you can make a decision on the best way to combine your data (if you need to combine them at all), and programatically add time to variable names (if necessary).[[162]](#footnote-569) Waiting to combine data has benefits:

1. Having variables named consistently over time (with no time component added) allows you to easily reuse your data collection and data capture tools, as well as your cleaning code, each wave.
2. Storing files separately prevents you from potentially wasting time combining your data in a way that ends up not actually being useful or from wasting time merging datasets that later need to be re-combined because you find an error in an individual dataset at some point.

#### 8.4.1.1 Time in variable names

While combining your datasets across time should not happen early on in your project, it is helpful to consider early on how you *might* combine data in the future. If you do plan to potentially merge data in a wide format, it can be helpful to go ahead and plan your rule for adding time to variable names, and add that rule to your style guide. Just be abundantly clear in your guide that this time component should only be added when datasets are combined.

There is no right or wrong way to assign time in your variable names necessarily. Just make sure you continue to follow the rules from above (such as never starting a variable name with a number). Below are some options for adding time to a sample variable, s\_gender.

* As a prefix or suffix with a generic abbreviation, such as w1 for wave 1, added with a delimiter \_
  + w1\_s\_gender or s\_gender\_w1
* As a prefix or suffix with a meaningful abbreviation, such as f21 for fall 2021, added with a delimiter (\_)
  + f21\_s\_gender or s\_gender\_f21
* One of the above options with no delimiter
  + w1s\_gender or s\_genderw1
* As a number embedded into your variable at a certain location, for instance, after an existing prefix such as s for student survey
  + s1\_gender, s2\_gender

While the first and second method do add additional characters to your variable name, there are also benefits to adding time in these ways. First, it can be easier to visually spot and interpret the time component when it is separated with a delimiter. Second, adding time as a standalone component also allows you to more easily, programmatically, manipulate the time component of your variable. This gives you more flexibility in working with your data, especially in selecting variables and restructuring your datasets.

## 8.5 Value Coding

In addition to naming variables in a standardized way, variables values also need to be added consistently. Value codes apply to any of your categorical variable. This may be numeric categorical values with associated labels (ex: “no” = 1) or it may be character categorical values with associated labels (ex: “no” = ‘n’).

First, if you are using a pre-existing measure, assign values and labels in the manner that the technical documentation tells you to assign codes. That will be important for any further derivations you need to make later on based on those measures. Otherwise, you will be assigning your own values and labels. Some guidelines for assigning codes and labels (as well as examples for how to apply those guidelines) are below.

* Values must be unique
  + Do: Assign “yes” = 1 and “no” = 0
  + Don’t: Assign “yes” = 1 **and** “no” = 1
* Values must be consistent within a variable
  + Do: For gender assign “male” = ‘m’
  + Don’t: For gender allow “male” = ‘m’ or ‘M’ or ‘Male’ or ‘male’
* Values must be consistent across time
  + Do: Assign anx1 values of “yes” = 1 and “no” = 0 in wave 1 **and** wave 2
  + Don’t: Assign anx1 values of “yes” = 1 and “no” = 0 in wave 1 and values of “yes” = 1 and “no” = 2 in wave 2
* Values should be consistent across the project
  + Do: Assign “yes” = 1 and “no” = 0 as the value for all yes/no items
  + Don’t: Assign “yes” = 1 and “no” = 0 for some variables, and “yes” = 1 and “no” = 2 for others
    - Unless a pre-existing measure determines how some variables are coded
* Order Likert-type scale response options in a logical way
  + Do: Assign “Strongly Disagree” = 1; “Disagree” = 2; “Agree” = 3; “Strongly Agree” = 4
  + Don’t: Assign “Strongly Disagree” = 1; “Disagree” = 3; “Agree” = 4; “Strongly Agree” = 2
    - Unless a pre-existing measure tells you to code variables in a different way
* Define missing values
  + You may choose to leave all missing values as blank, NA, or NULL and that is okay
  + However, you may care about the specific reason for missing data and need to consider defining missing values based on their properties
    - The key in this case is to use extreme values that do not actually occur in your data and to also use values that match your variable type (ex: numeric missing values for numeric variables)[[163]](#footnote-573)

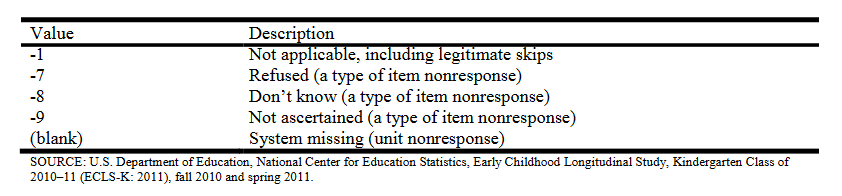


Figure 8.3: Missing values assigned in the ECLS-K:2011 data file

## 8.6 Coding

If your team plans to clean data using code, it can be very helpful to create a coding style guide. This style guide can be tailored to a specific language that all staff will use (such as R or Stata), or it can be written more generically to apply to any coding language staff use to clean data. Below is a small sampling of good coding practices to consider adding to your guide. If you are looking for guides for a specific language, it can be very helpful to google existing style guides in that language.

* Consider building and implementing coding templates[[164]](#footnote-578)
  + Templates can standardize the format of syntax files (such as using standard headers to break up code)
  + They also standardize the summary information provided at the beginning of your syntax (code author, project name, date created)
* Use comments throughout your code to clearly explain the purpose of each code chunk
  + The format of these comments will be dependent on your coding language
    - R uses #
    - SPSS and Stata use \*
* Improve code readability by using
  + spaces
  + indentation
  + setting a line limit for your code (ex: 80 characters)
* Use relative file paths for reproducibility
  + Setting absolute file paths in syntax reduces reproducibility because future users may have different file paths. It is important to set file paths relative the directory you are working in.[[165]](#footnote-581)
* If you create objects in your program (like you do in R or Python), consider adding object naming rules similar to variable naming rules
  + No spaces in object names
  + No special characters except \_ to separate words
  + No names that are existing program keywords (if, for, etc.)
* Reduce duplication, improve efficiency, and increase your ability to troubleshoot errors by using functions, loops, or macros for repetitive code chunks
* Record session information for future users
  + Record both version information as well as operating system information relevant to your code to increase the reproducibility of your code

# 9 Data Tracking

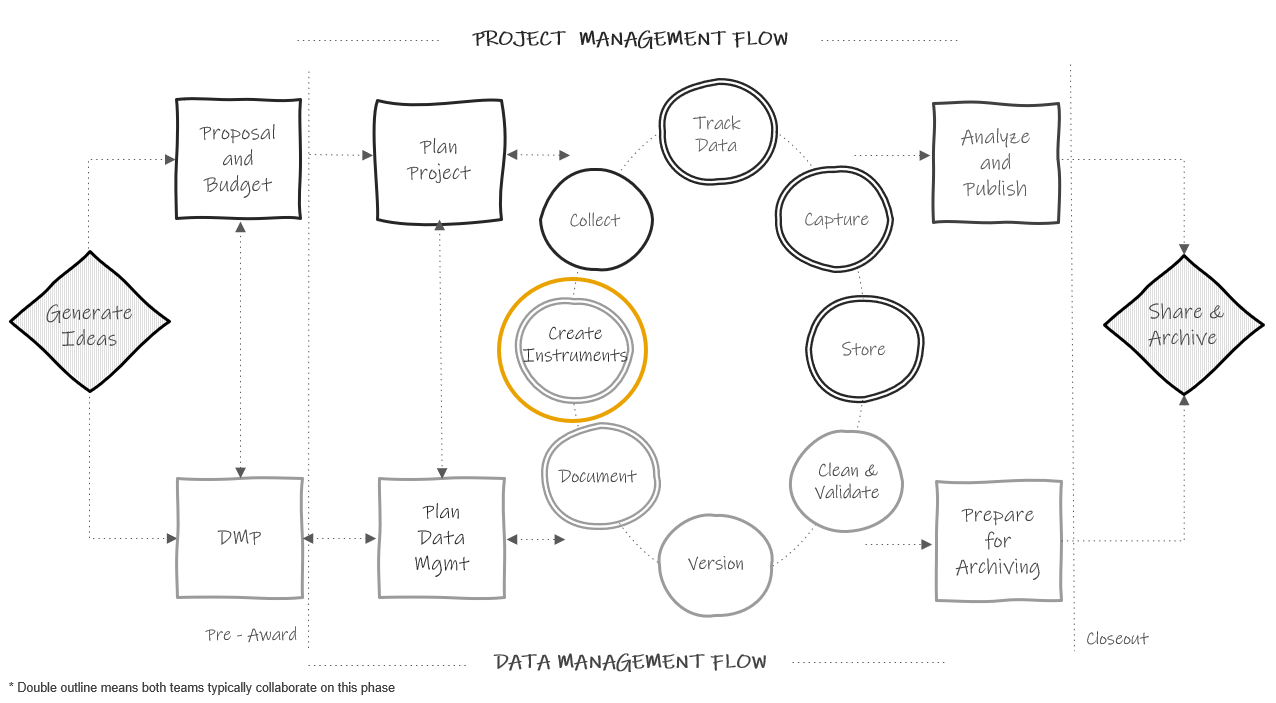


Figure 9.1: Tracking in the research project life cycle

During your project, you will want to be able to answer both progress and summary questions about your recruitment and data collection activities.

1. How many participants consented to be in our study? How many have we lost during our study and why?
2. How much progress have we made in this cycle of data collection? How much data do we have left to collect?
3. How many forms did we collect each cycle and why are we missing data for some forms?

Questions like these will arise many times throughout your study for both your own project coordination purposes, as well as for external progress reporting and publication purposes. Yet, how will you answer these questions? Will you dig through papers, search through emails, and download in progress data, each time you need to answer a question about the status of your project activities? A better solution is to track all project activities in a participant tracking database.

A participant tracking database is an essential component of both project management and data management. This database contains all study participants, their relevant study information, as well as tracking information about their completion of project milestones. This database has two underlying purposes.

1. To serve as a roster of study participants as well as a “master key”[[166]](#footnote-588) that houses both identifying participant information as well as assigned unique study identifiers.
2. To aid in project coordination and reporting, tracking the movement of participants as well as completion of study events throughout a project.

This database is considered your single source of truth concerning everything that happened throughout the duration of your study. Any time a participant consents to participate, drops from the study, changes their name, completes a data collection measure, is provided a payment, or moves locations, a project coordinator, or other designated team member, updates the information in this one location. Tracking this administrative information in this one database, rather than across disparate spreadsheets, emails and papers, ensures that you always have one definitive source to refer to when seeking answers about your sample and your project activities.

## 9.1 Benefits

A thorough and complete participant database that is updated regularly is beneficial for the following reasons:

1. Securing data de-identification
   * Assigning unique study identifiers that are only linked to a participant’s true identity within this one database is necessary for maintaining participant confidentiality. This database is stored in a [restricted secure location](#trackingstorage), separate from where de-identified study datasets are stored, and is typically destroyed at a period of time after a project’s completion,
2. Project coordination and record keeping
   * Checking and tracking activities on a daily basis allows your team to easily report the status of data collection activities (ex: as of today we have completed 124 out of 150 assessments)
   * Checking and tracking daily, compared to checking and tracking after data collection is complete, reduces the likelihood of missing data
   * Thorough tracking allows you to explain missing data in reports and publications (ex: Teacher 1234 went on maternity leave)
3. Consort diagram creation
   * This database aids in the creation of documentation including the flow of participants in your [consort diagram](#supplement)
4. Dataset verification
   * As part of your data cleaning process, all raw data sample sizes should compared against what is reported as complete in your participant database to ensure that no participants are missing from your final datasets

## 9.2 Building your database

It is beneficial to build this database before you begin recruiting participants, typically during the same time that you are building your data collection tools. This way, as your team recruits and consents participants, you can record their name, assign them a study ID, and enter any other necessary identifying contact information into the participant database. Depending on your database system, you may even be able to scan and upload copies of your consent forms into the database.

While a project coordinator can build this database, it can be helpful to consult with a data manager, or someone with relational database expertise, when creating this system. This ensures that your system is set up efficiently and comprehensively.

This database may be a standalone structure, used only for tracking and anonymization purposes. Or it may be integrated as part of your larger study system, where all study data is collected and/or entered as well.

### 9.2.1 Relational databases

Before we discuss how to build this database, it is helpful to have a basic understanding of the benefits of relational databases, first introduced in our [Data Structure](#database) chapter. A relational database organizes information into tables, made up of records (rows) and fields (columns), and tables are related through keys[[167]](#footnote-591). Using a normalized relational database[[168]](#footnote-594) to track participant information, compared to disparate, non-connected spreadsheets, has many benefits.

1. Reduces redundancy and improves efficiency
2. Increases performance
3. Decreases required storage
4. Reduces data entry errors

Let’s compare a very simple example of building a tracking database using a relational model and a non-relational model.

#### 9.2.1.1 Relational model

In Figure 9.2 we have three entities we need to track in our database: schools, teachers, and students. We built a very simple database with one table for each entity. Within each table we added fields that we need to collect on these participants. We have also set up our tables to include primary keys (which uniquely identify rows in each table) and foreign keys (which includes values that correspond to the primary key of another table). Our keys are all unique study identifiers that we have assigned to our study participants.

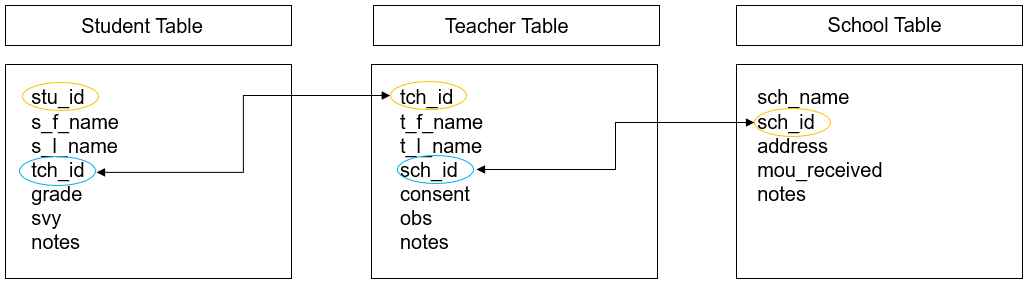


Figure 9.2: Participant database built using a relational model

We can see here that in each table we have no duplicated, repeating information. The student table only contains student level information, the teacher table only contains teacher level information, and the school table only contains school level information. This is a huge time saver. Imagine if a teacher’s last name changes. Rather than updating that name in multiple places, we now only update it once, in the teacher table. If we want to see a table with both student and teacher information, we can simply query our database to create a new table. In some programs, this type of querying may be a simple point and click option, in other programs it may require someone to write some simple queries that can then be used at any time by any user.

Say for example, we needed to pull a roster of students for each teacher. We could easily create and run a query, such as this SQL query, that joins the student and teacher tables above by tch\_id and then pulls the relevant teacher and student information from both tables:

SELECT Teacher.t\_l\_name, Teacher.t\_f\_name, Student.s\_l\_name, Student.s\_f\_name, Student.grade FROM Student INNER JOIN Teacher ON Student.tch\_id = Teacher.tch\_id ORDER BY Teacher.t\_l\_name, Teacher.t\_f\_name, Student.s\_l\_name, Student.s\_f\_name

#### 9.2.1.2 Non-relational model

Now imagine that we built a non-relational database (Figure 9.3), such as three tabs in an Excel spreadsheet, to track our participant information. Since we are unable to set up a system that links these tables together, we need to enter redundant information into each table (such as teacher or school name) in order to see that information within each table without having to flip back and forth across tables to find the information we need. Using this method we now have to enter repeating teacher and school names in the student table, and if any teacher names change, we will need to update it in both the teacher table and in the student table for every student associated with that teacher. This requires more entry time and creates the opportunity for more data entry errors.

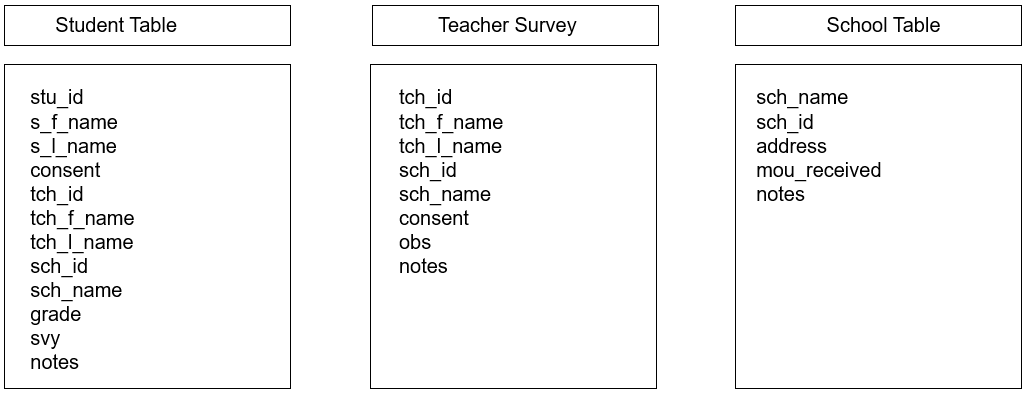


Figure 9.3: Participant database built in using a non-relational model

#### 9.2.1.3 Considerations

If your study includes a variety of related entities, tracked over waves of time, a relational database will be very helpful to build. If however, you are only tracking one entity (ex: just students) for one wave of data collection, then a database might be overkill and a simple spreadsheet will work just fine.

### 9.2.2 Structuring the database

Before you can begin to construct your database, you will need to collect the following pieces of information.

1. Who are your entities/units of analysis?
   * students, teachers, classrooms, districts, and so on
2. Are you collecting data longitudinally, across more than one wave?
3. Do you want to use a relational table structure?
   * If yes, how do you want to construct and relate your tables?
4. What fields do you want to include in each table?

Once you collect those pieces of information, you can begin to design your database structure. It can be helpful to visualize your database model during this process. In Figure 9.4 I am designing a database structure for a scenario where I will be collecting information from teachers in schools, over two waves of data collection.

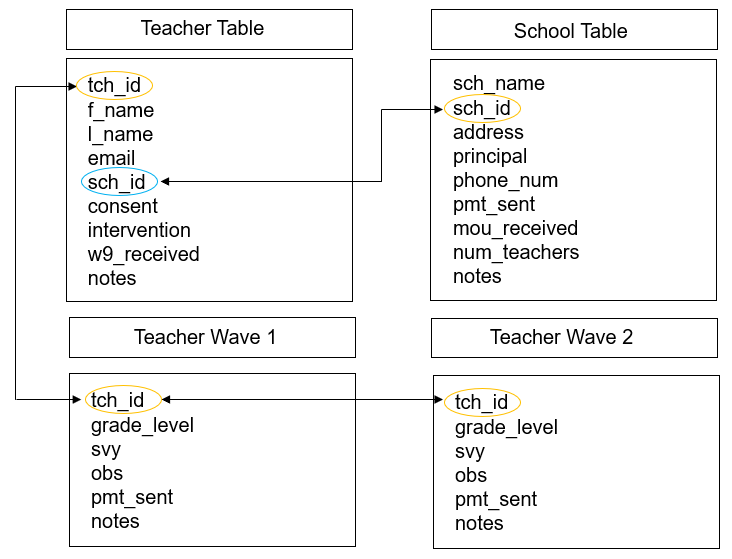


Figure 9.4: Example participant database model

I have designed my database model in this way:

1. I have four tables total
   * Two tables have information that should be constant (name, email, consent, one time payments sent out, one time documents received)
     + If at any time this constant information changes (ex: new last name, new principal), I would update that information in the appropriate table and make a note of when and why the change occurred in my “notes” field
   * Two tables are for my longitudinal information
     + This is where I will track my data collection activities each wave
2. I have connected my tables through primary and foreign keys (“tch\_id” and “sch\_id”)

While this example is for a fairly simple scenario, you can hopefully see how you might extrapolate this model to more entities and more waves of data collection, as well as how you might modify it to better meet the needs of your specific project.

It is possible that you have a study where you are collecting anonymous data. While you will no longer have participant level identifiable information, you can still construct a simple system, creating tables based on your sites for instance (school table, district table) to track your project management and data collection efforts (ex: number of student surveys received per school per wave, MOU received).

### 9.2.3 Choosing fields

As you design your database model, you will also need to choose what fields to include in each table. The fields you choose to include will be dependent on your particular study design. While your participant tracking database may be the same database you enter all of your study data, for the purposes of this chapter we are only considering fields that are relevant for project coordination and participant de-identification. We are not concerned with fields that are collected as part of your data collection measures (i.e. survey items). You can consider your participant tracking database as an **internal** database that is only used for coordination, summary, and linking purposes. This is not a database where you would export data for external data sharing.

**Ideas of static fields to collect:**

* Study IDs (Primary and foreign keys for a relational database)
* Names (Participants and sites)
* Contact information
* Other necessary linking identifiers (double IDs, district/school IDs)
* Information helpful for data collection scheduling (blocks, class times)
* Consent/assent status
* Randomization information (cohort, group)
* Condition (treatment/control)
* Summary information for rates (# of consents sent out, # of students in class)
* Administrative status (W-9 received, MOUs received)
* Movement/drop out status
* Notes
  + Reasons for changes (for example changes in name, email)
  + Reasons for movement/drop out

**Ideas of fields to collect longitudinally:**

* Data collection status (unique fields for each instrument)
* Incentive status (gift cards sent out)
* Other changing information relevant to project coordination (grade level)
* Notes
  + Communication with participants
  + Reasons for missing data
  + Errors in data

#### 9.2.3.1 Structuring fields

As you choose your fields you also need to make some decisions about how you will structure those fields.

1. Set data types for your fields (ex: character, integer, date)
   * Restrict entry values to only allowable data types to reduce errors
2. Set allowable values and ranges
   * For example, a categorical status field may only allow “complete”, “partially complete” or “incomplete”
3. Do not lump separate pieces of information together in a field
   * For example separate out first name and last name into two fields
4. Name your fields according to the variable naming rules we discussed in our [Style Guide](#varname) chapter

### 9.2.4 Choosing a tool

There are many criteria to consider when choosing a tool to build your database in.

* Choose a tool that is customizable to your needs
  + Can you build a relational table structure?
  + Can you export files? Can we connect to the database via application programming interfaces (APIs)?
  + Can you query data?
* Choose a tool that is user-friendly
  + You don’t want a tool with a steep learning curve for users.
* If you are running a project across multiple sites, consider the accessibility of the tool
  + For example, you may want a tool that is cloud-based so everyone can access it
  + You may also want to make sure multiple users can access it at the same time
* Choose a tool that is interoperable
  + For instance, some tools may have difficulties running on certain operating systems
* Consider cost and licensing
  + There are many free tools, but they may not provide all of the functionality you want
  + What products do you already have access to? Your institution has a license for?
* Consider security
  + Security in terms of participant confidentiality
    - Does the tool meet HIPAA/FERPA requirements?
    - Can we limit access to the entire database? To specific tables?
      * Since this database contains PII you will want to place restrictions on who can access and enter data
  + Protect data loss
    - Can we backup the system?
    - Can we protect against overwriting data?
* Data quality protection
  + Can we set up data quality constraints? For example, restrict input types/values.

There are many tool options you can choose from. A sampling of those options are below. These tools represent a wide range from the criteria above. Take some time to review your options to see which one best meets your needs.

* Microsoft Access
* Microsoft Excel
* QuickBase
* Airtable
* REDCap
* Claris FileMaker
* Google Sheets and Google Forms
* Forms that feed into a relational database, maintained using a SQL (structured query language) database engine such as SQLite, MySQL, or PostgreSQL

### 9.2.5 Entering data

Your last consideration when building your database will be, how do you want your team to enter data into your database? There are many ways to enter data including using SQL statements, importing data, integrating your data collection platform and your tracking database, or even scanning forms using QR codes. While some of those options may work great for your project, here we are going to talk about the two simplest and most common options: manually entering data into a spreadsheet view, and manually entering data into a form.

#### 9.2.5.1 Entering data in a spreadsheet view

Your first option is to manually enter data in a spreadsheet format for each participant in a row. This would be the most common (or only) option when using tools such as Microsoft Excel or Google Sheets. However, you can also use this option when entering into other database tools such as Microsoft Access. There are both pros and cons to this method.

* Pros: This is the quickest and easiest method. It also allows you to view all the data holistically.
* Cons: This method can lead to errors if someone enters data on the wrong row/record.

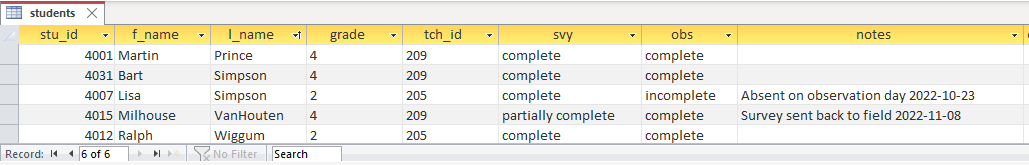


Figure 9.5: Example spreadsheet view data entry

#### 9.2.5.2 Entering data in a form

Your second option is to create a form that is linked to your tables. As you enter data in your forms, it automatically populates your tables with the information. This option is possible in many systems including Microsoft Access, RedCap, and even Google Forms which populates into Google Sheets.

* Pros: This method reduces data entry errors as you are only working on one participant form at a time
* Cons: Takes some time, and possibly expertise, to set up the data entry forms

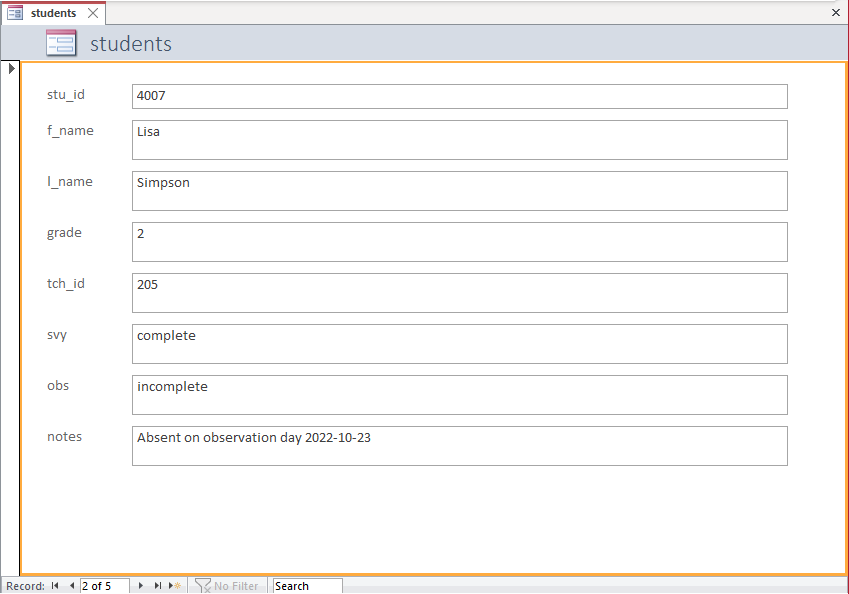


Figure 9.6: Example form view data entry

#### 9.2.5.3 Other considerations

If your participant tracking database is separate from your data collection tools, all information will need to be entered by your team using one of these ways. However, if your participant tracking tool is also your data collection/data capture tool (such as those who collect data using RedCap), fields such as data collection status (ex: survey completed), may not be manually entry. Rather they may be automated to populate as “complete” once a participant submits their responses in the data collection tool.

## 9.3 Creating unique identifiers

Assigning unique identifiers (UIDs) is an important part of ensuring the protection of the privacy of human participants. Hopefully at this point your team has already created your study id schema and documented it in an [SOP](#sop). However, since this is the point when you will begin actually assigning those IDs, this is a good time to discuss best practices for creating your unique identifier schema. There are several rules to consider when assigning your identifiers.

1. These identifiers are numeric or alphanumeric and typically range from 2-10 digits
2. They must be unique within entities and but also across entities to reduce confusion
   * For example, no duplicating IDs within students or across students and teachers
3. The identifier must be randomly assigned and be completely distinct from any participant information
   * The ID should not be associated with name, location, dob, income, grade level, and so forth
   * As an example, do not sort names alphabetically and then assign IDs in sequential order or do not include initials as part of the identifier
4. Participants must keep this same identifier for the entire project, never change a participant’s ID
   * This applies across all waves of data collection
   * This even applies in circumstances where a participant has the opportunity to be re-recruited into your study more than once. The participant still keeps the same ID throughout the study. You can use other variables to identify the unique instances of that participant (ex: cohort associated with that participant)
5. If confidentiality was promised to schools or districts, assign identifiers to sites as well

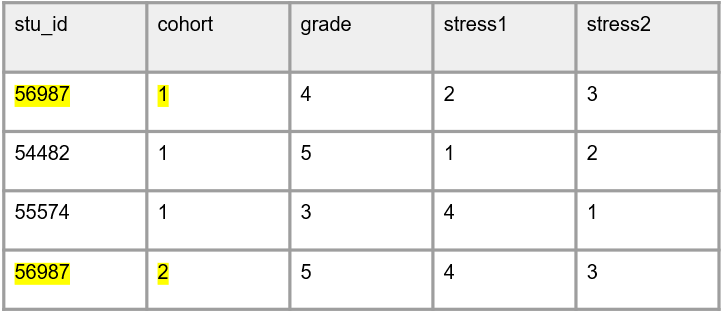


Figure 9.7: Example of keeping participant IDs for the entire study

The only time you will not assign unique identifiers is when you collect anonymous data. In this situation you will not be able to assign identifiers since you will not know who participants are. However, it is still possible to assign identifiers to known entities such as school sites if anonymity is required.

## 9.4 Tracking best practices

Last, it is important to review **how** to use the participant tracking database. As mentioned earlier, this database should be created before you begin recruiting participants so that as soon as participants begin to consent, they can be added to your database. Then once data collection begins, your Institutional Review Board (IRB) approved staff member can begin tracking data as it is collected. Your team may designate one person to track, or they may designate multiple. If you are working across multiple sites, with multiple teams, you will most likely have one or more persons at each site tracking data as it comes in.

Since this is a highly sensitive database, not only may you want to restrict who has access to your database, but if your tool allows it, it can be helpful to also limit access to specific tables with identifiable information. For instance, project coordinators may have access to all information in the database (including sensitive data), while research assistants have limited access to tables that only include unique identifiers where they can track data collection completion. No matter who is tracking information, it is important to develop a tracking routine, assign roles and responsibilities, add your routine to an [SOP](#sop).

Some tracking best practices include:

1. Update all relevant contact information in this database alone, not in other places.
   * Again, you can allow multiple people to update or designate one contact to funnel all updates through.
2. When tracking data (consents, paperwork, data collection forms) only track data that you physically have (paper or electronic)
   * Never track data as “complete” that someone tells you they collected
     + You can always mark this information in a “notes” field and track when you have the physical data
3. Track daily during data collection
   * Do not wait until the end of data collection to track what data was collected
   * This helps ensure that you don’t miss the opportunity to collect data that you *thought* you had but never actually collected
4. Only track complete data as “complete”
   * If a form is only partially completed and you plan to send it back out to the field for completion, mark this in the “notes” but do not mark it as “completed”. If you have a “partially completed” option, you can mark this option.

# 10 Data Collection

## 10.1 Why consider data management in data collection?

## 10.2 Consents

## 10.3 Electronic data collection instruments

## 10.4 Paper data collection instruments

## 10.5 Interviews/focus groups

# 11 Data Capture

## 11.1 Electronic data capture

## 11.2 Paper data capture

## 11.3 Extant data

# 12 Data Storage and Security

## 12.1 Types of data you’ll be storing

## 12.2 General security rules

## 12.3 Participant tracking database

## 12.4 Electronic data

## 12.5 Detachable media

## 12.6 Audio/visual data

## 12.7 Paper data

## 12.8 Sharing data

# 13 Data Cleaning

## 13.1 Foundational knowledge

## 13.2 Data structure

## 13.3 Data cleaning plan

## 13.4 Data validation

## 13.5 Why use code?

# 14 Data Sharing

## 14.1 Why share your data?

## 14.2 Considering FAIR principles

## 14.3 Best practices

## 14.4 Retractions and revisions

# 15 Wrapping It Up

## 15.1 Connecting practices to outcomes

## 15.2 Putting in the work

# 16 Call to Action

## 16.1 Last thoughts

## 16.2 Training for future researchers

## 16.3 Investing in data management and data managers

Academic Health Science Libraries, Association of, Association of American Medical Colleges, and Association of Research Libraries. “Institutional Strategies for the NIH Data Management and Sharing Policy: Infrastructure, Policies, and Services,” September 2022. <https://www.aamc.org/media/62881/download?attachment>.

“Advances in Methods and Practices in Psychological Science. Association for Psychological Science - APS.” Accessed February 20, 2023. <https://www.psychologicalscience.org/publications/ampps>.

Alliance, DDI. “Controlled Vocabularies - Overview Table of Latest Versions Data Documentation Initiative.” Accessed January 19, 2023. <https://ddialliance.org/controlled-vocabularies>.

———. “Mapping to Dublin Core (DDI Version 2).” Accessed January 19, 2023. <https://ddialliance.org/resources/ddi-profiles/dc>.

Alston, Jesse M., and Jessica A. Rick. “A Beginner’s Guide to Conducting Reproducible Research.” *The Bulletin of the Ecological Society of America* 102, no. 2 (April 2021). <https://doi.org/10.1002/bes2.1801>.

alvinashcraft. “Naming Files, Paths, and Namespaces - Win32 Apps,” December 15, 2022. <https://learn.microsoft.com/en-us/windows/win32/fileio/naming-a-file>.

Aly, Mariam. “Aly Lab Wiki,” August 28, 2018. <https://osf.io/mdh87/>.

Arslan, Ruben C. “How to Automatically Document Data with the Codebook Package to Facilitate Data Re-Use.” Preprint. PsyArXiv, September 5, 2018. <https://doi.org/10.31234/osf.io/5qc6h>.

Baker, Monya. “1,500 Scientists Lift the Lid on Reproducibility.” *Nature* 533, no. 7604 (May 1, 2016): 452–54. <https://doi.org/10.1038/533452a>.

Balazs, Aczel. “A Crowdsourced Effort to Develop a Lab Manual Template. Google Docs.” Accessed January 12, 2023. <https://docs.google.com/document/d/1LqGdtHg0dMbj9lsCnC1QOoWzIsnSNRTSek6i3Kls2Ik/edit?usp=embed_facebook>.

Board, The Ohio State University Institutional Review. “Guidelines for Writing a Research Protocol,” March 6, 2014. <http://orrp.osu.edu/files/2011/10/GuidelinesforWritingaResearchProtocol.pdf>.

Bolam, Mike. “Guides: Metadata & Discovery @ Pitt: Metadata Standards.” Accessed January 18, 2023. <https://pitt.libguides.com/metadatadiscovery/metadata-standards>.

Bordelon, Dominic. “Guides: Research Data Management @ Pitt: Understanding Research Data Management.” Accessed October 13, 2022. <https://pitt.libguides.com/managedata/understanding>.

Borghi, John A., and Ana E. Van Gulick. “Data Management and Sharing: Practices and Perceptions of Psychology Researchers.” *PLOS ONE* 16, no. 5 (May 21, 2021): e0252047. <https://doi.org/10.1371/journal.pone.0252047>.

Borghi, John, and Ana Van Gulick. “Promoting Open Science Through Research Data Management.” *Harvard Data Science Review*, July 28, 2022. <https://doi.org/10.1162/99608f92.9497f68e>.

Borycz, Joshua. “Implementing Data Management Workflows in Research Groups Through Integrated Library Consultancy.” *Data Science Journal* 20, no. 1 (February 17, 2021): 9. <https://doi.org/10.5334/dsj-2021-009>.

Briney, Kristin. *Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success*. Research Skills Series. Exeter, UK: Pelagic Publishing, 2015.

Briney, Kristin, Heather Coates, and Abigail Goben. “Foundational Practices of Research Data Management.” *Research Ideas and Outcomes* 6 (July 27, 2020): e56508. <https://doi.org/10.3897/rio.6.e56508>.

Broman, Karl W., and Kara H. Woo. “Data Organization in Spreadsheets.” *The American Statistician* 72, no. 1 (January 2, 2018): 2–10. <https://doi.org/10.1080/00031305.2017.1375989>.

Buchanan, Erin M., Sarah E. Crain, Ari L. Cunningham, Hannah R. Johnson, Hannah Stash, Marietta Papadatou-Pastou, Peder M. Isager, Rickard Carlsson, and Balazs Aczel. “Getting Started Creating Data Dictionaries: How to Create a Shareable Data Set.” *Advances in Methods and Practices in Psychological Science* 4, no. 1 (January 2021): 251524592092800. <https://doi.org/10.1177/2515245920928007>.

Burnard, Lou. *What Is the Text Encoding Initiative?: How to Add Intelligent Markup to Digital Resources*. OpenEdition Press, 2014. <https://doi.org/10.4000/books.oep.426>.

Butters, Oliver W, Rebecca C Wilson, and Paul R Burton. “Recognizing, Reporting and Reducing the Data Curation Debt of Cohort Studies.” *International Journal of Epidemiology* 49, no. 4 (August 1, 2020): 1067–74. <https://doi.org/10.1093/ije/dyaa087>.

Cakici, Tatiana Baquero. “Folders v. Metadata in SharePoint Document Libraries. Enterprise Knowledge,” April 5, 2017. <https://enterprise-knowledge.com/folders-v-metadata-sharepoint-document-libraries/>.

Campos-Varela, Isabel, and Alberto Ruano-Raviña. “Misconduct as the Main Cause for Retraction. A Descriptive Study of Retracted Publications and Their Authors.” *Gaceta Sanitaria* 33, no. 4 (July 1, 2019): 356–60. <https://doi.org/10.1016/j.gaceta.2018.01.009>.

CDISC. “CDISC Standards in the Clinical Research Process.” Accessed January 12, 2023. <https://www.cdisc.org/standards>.

CESSDA. “Data Authenticity - Data Management Expert Guide.” Accessed January 13, 2023. <https://dmeg.cessda.eu/Data-Management-Expert-Guide/3.-Process/Data-authenticity>.

———. “Documentation and Metadata.” Accessed February 9, 2023. <https://dmeg.cessda.eu/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata>.

Ceviren, A. Busra, and Jessica Logan. “Ceviren\_logan\_EHE\_forum\_2022.pdf.” Presentation. presentation, April 4, 2022. <https://doi.org/10.6084/m9.figshare.19514368.v1>.

Clinical {and} Translational Research, Johns Hopkins Institute for. “Best Practices for Research Data Management.” Accessed February 10, 2023. <https://ictrweb.johnshopkins.edu/ictr/dmig/Best_Practice/a8376318-ebd6-421f-be63-acf8c88376a1_e0996b36-932a-44d5-94b0-61ad04f5774d.html?v=65849&ip=qam562joeegjlraagoltna>.

———. “Data Dictionary/Codebook.” Accessed February 10, 2023. <https://ictrweb.johnshopkins.edu/ictr/dmig/Best_Practice/a8376318-ebd6-421f-be63-acf8c88376a1_6342a1c3-1a5d-4287-a46e-374824e3780e.html?v=65849&ip=hpdkvlttuiyioooqhw>.

Cofield, Melanie. “LibGuides: Metadata Basics: Key Concepts.” Accessed February 10, 2023. <https://guides.lib.utexas.edu/metadata-basics/key-concepts>.

CONSORT. “Consort - the CONSORT Flow Diagram.” Accessed January 13, 2023. <https://www.consort-statement.org/consort-statement/flow-diagram>.

Cowles, Wind. “Research Guides: Research Data Management at Princeton: Home.” Accessed September 15, 2022. <https://libguides.princeton.edu/c.php?g=102546&p=665862>.

Cruse, Trisha. “Best Practice: Provide Budget Information for Your Data Management Plan. DataOne,” September 1, 2011. <https://dataoneorg.github.io/Education/bestpractices/provide-budget-information>.

Dahdul, Wasila. “Research Guides: Research Data Management: Describing Data.” Accessed January 18, 2023. <https://guides.lib.uci.edu/datamanagement/describe>.

Daskalova, Gergana. “Coding Etiquette. Coding Club.” Accessed February 17, 2023. <https://ourcodingclub.github.io/tutorials/etiquette/>.

David T. Bourgeois, Ph D., and Bourgeois. *Information Systems for Business and Beyond*. Published through the Open Textbook Challenge by the Saylor Academy, 2014. <https://pressbooks.pub/bus206/>.

Desk, CSP Library Research. “CSP Library: Zotero Guide: Defining Your Research Workflow.” Accessed November 1, 2022. <https://library.csp.edu/Zotero/workflow>.

Dijk, Wilhelmina van, Christopher Schatschneider, and Sara A. Hart. “Open Science in Education Sciences.” *Journal of Learning Disabilities* 54, no. 2 (March 2021): 139–52. <https://doi.org/10.1177/0022219420945267>.

DMPTool. “DMP Templates.” Accessed February 9, 2023. <https://dmptool.org/public_templates?page=1&search=&sort_direction=asc&sort_field=templates.title>.

“Documents. Xkcd.” Accessed February 16, 2023. <https://xkcd.com/1459/>.

Doucette, Lise, and Bruce Fyfe. “Drowning in Research Data: Addressing Data Management Literacy of Graduate Students - PDF Free Download,” 2013. <https://docplayer.net/8853333-Drowning-in-research-data-addressing-data-management-literacy-of-graduate-students.html>.

Duru, Maya, and Sarah Kopper. “Gantt Chart Template,” n.d. <https://www.povertyactionlab.org/sites/default/files/research-resources/rr_grantprop_Template_Gantt_Chart.pdf>.

Eaker, C. “What Could Possibly Go Wrong? The Impact of Poor Data Management.” *In Federer, L. (Ed.). The Medical Library Association’s Guide to Data Management for Librarians*, 2016. <https://trace.tennessee.edu/cgi/viewcontent.cgi?article=1023&context=utk_libpub>.

Education Sciences, Institute of. “Data Sharing.” Accessed October 27, 2022. <https://ies.ed.gov/funding/datasharing_implementation.asp>.

———. “Frequently Asked Questions about Providing Public Access to Data.” Accessed October 21, 2022. <https://ies.ed.gov/funding/datasharing_faq.asp>.

Education Sciences, Institute of. “IES Funding Opportunities.” Accessed January 9, 2023. <https://ies.ed.gov/funding/23rfas.asp>.

Embse, Nathaniel von der. “The RAISE Lab Manual,” August 23, 2019. <https://uploads.strikinglycdn.com/files/f3dfcf89-32f2-48cb-94cd-3070f733c53d/RAISE%20Lab%20Guidelines_Final.pdf>.

“FAIR Principles. GO FAIR.” Accessed October 21, 2022. <https://www.go-fair.org/fair-principles/>.

Farewell, Dr Timothy S. “My Easy r Script Header Template – Tim Farewell,” April 10, 2018. <https://timfarewell.co.uk/my-r-script-header-template/>.

Fay, Gavin, and Ashley Novak. “Faylab Lab Manual,” December 7, 2022. <https://thefaylab.github.io/lab-manual/>.

Figshare. “Fair-Figshare.” Accessed February 21, 2023. <https://knowledge.figshare.com/publisher/fair-figshare>.

———. “How to Write a Data Management Plan (DMP) and Include Figshare in Your Data Sharing Plans - a Help Article for Using Figshare.” Accessed February 21, 2023. <https://help.figshare.com/article/how-to-write-a-data-management-plan-dmp-and-include-figshare-in-your-data-sharing-plans>.

“Figshare - Credit for All Your Research. Figshare.” Accessed January 18, 2023. <https://figshare.com/>.

Foster, Erin D., and Ariel Deardorff. “Open Science Framework (OSF).” *Journal of the Medical Library Association : JMLA* 105, no. 2 (April 2017): 203–6. <https://doi.org/10.5195/jmla.2017.88>.

Foundation, National Science. “Data Management for NSF EHR Directorate Proposals and Awards,” 2017. <https://www.nsf.gov/bfa/dias/policy/dmpdocs/ehr.pdf>.

Gentzkow, Matthew, and Jesse Shapiro. “Code and Data for the Social Sciences: A Practitioner’s Guide,” 2014. <https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf>.

Geraghty, Louise, and Laura Feeney. “Formalize Research Partnership and Establish Roles and Expectations. The Abdul Latif Jameel Poverty Action Lab (j-PAL).” Accessed January 12, 2023. <https://www.povertyactionlab.org/resource/formalize-research-partnership-and-establish-roles-and-expectations>.

Gilmore, Rick O., Joy Lorenzo Kennedy, and Karen E. Adolph. “Practical Solutions for Sharing Data and Materials from Psychological Research.” *Advances in Methods and Practices in Psychological Science* 1, no. 1 (March 2018): 121–30. <https://doi.org/10.1177/2515245917746500>.

Gonzales, Sara, Matthew B. Carson, and Kristi Holmes. “Ten Simple Rules for Maximizing the Recommendations of the NIH Data Management and Sharing Plan.” *PLOS Computational Biology* 18, no. 8 (August 3, 2022): e1010397. <https://doi.org/10.1371/journal.pcbi.1010397>.

Grace-Martin, Karen. “The Wide and Long Data Format for Repeated Measures Data. The Analysis Factor,” October 4, 2013. <https://www.theanalysisfactor.com/wide-and-long-data/>.

Hansen, Karsten Kryger. *DataFlowToolkit.dk*, 2017. <https://doi.org/10.5278/16k4-4n24>.

Hart, Sara. “Example of a Data Management Plan.” figshare, November 11, 2020. <https://doi.org/10.6084/m9.figshare.13218743.v1>.

Health, National Institutes of. “Budgeting for Data Management & Sharing Data Sharing.” Accessed January 11, 2023. <https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/budgeting-for-data-management-sharing#after>.

———. “Writing a Data Management & Sharing Plan.” Accessed October 27, 2022. <https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-DMS/writing-a-data-management-and-sharing-plan>.

House, The White. “Executive Order – Making Open and Machine Readable the New Default for Government Information. Whitehouse.gov,” May 9, 2013. <https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government->.

Hubbard, Aleata. *Data Cleaning in Mathematics Education Research: The Overlooked Methodological Step*, 2017. <https://eric.ed.gov/?id=ED583982>.

Humanities, National Endowment for the. “Data Managment Plans for NEH Office of Digital Humanities Proposals and Awards,” 2018. <https://www.neh.gov/sites/default/files/2018-06/data_management_plans_2018.pdf>.

ICPSR. “Guide to Codebooks 1st Edition.” Ann Arbor, MI, 2011. <https://www.icpsr.umich.edu/files/deposit/Guide-to-Codebooks_v1.pdf>.

———. “Guide to Social Science Data Preparation and Archiving: 6th Ed.” Accessed January 9, 2023. <https://www.icpsr.umich.edu/web/pages/deposit/guide/>.

———. “ICPSR, Data Management, Metadata.” Accessed January 18, 2023. <https://www.icpsr.umich.edu/web/pages/datamanagement/lifecycle/metadata.html>.

———. “ICPSR Template - Version 1,” February 16, 2023. <https://www.icpsr.umich.edu/files/ICPSR/nih/FINAL_ICPSR-NIH-DMS-Plan-Template_2023.docx>.

Iowa Libraries, University of. “University Libraries the University of Iowa.” Accessed February 10, 2023. <https://www.lib.uiowa.edu/>.

(ITCC), IMPACCT Trials Coordination Centre. “Standard Operating Procedure,” n.d. <https://www.uts.edu.au/sites/default/files/2020-07/5.5.5%20Allocation%20of%20Participant%20ID%20Numbers%20V2.5_20200520.pdf>.

Justice, National Institute of. “Data Archiving. National Institute of Justice.” Accessed October 27, 2022. <https://nij.ojp.gov/funding/data-archiving>.

Kawakita, Todd, and Jared Silver. “Strategic Data Project Adopt: Coding Style Guide,” 2013. <https://hwpi.harvard.edu/files/sdp/files/sdp-toolkit-coding-style-guide.pdf>.

Kline, Melissa. “A Technical Specification for Psychological Datasets. Google Docs,” 2018. <https://docs.google.com/document/d/1u8o5jnWk0Iqp_J06PTu5NjBfVsdoPbBhstht6W0fFp0/edit?usp=embed_facebook>.

Kovacs, Marton, Rink Hoekstra, and Balazs Aczel. “The Role of Human Fallibility in Psychological Research: A Survey of Mistakes in Data Management.” *Advances in Methods and Practices in Psychological Science* 4, no. 4 (October 2021): 251524592110459. <https://doi.org/10.1177/25152459211045930>.

LDbase. “Information to Gather Before Uploading Your Data LDbase.” Accessed January 18, 2023. <https://www.ldbase.org/resources/user-guide/information-to-gather>.

Lewis, Crystal. “Codebook Package Comparison.” Accessed January 17, 2023. <https://github.com/Cghlewis/codebook-pkg-comparison>.

———. “Data Capture Checklist.” Accessed November 2, 2022. <https://cghlewis.github.io/mpsi-data-training/capture.html>.

———. “Data Cleaning Checklist.” Accessed November 2, 2022. <https://cghlewis.github.io/mpsi-data-training/clean.html>.

———. “Data Collection Checklist.” Accessed November 2, 2022. <https://cghlewis.github.io/mpsi-data-training/collection.html>.

———. “Data Sharing Checklist.” Accessed November 2, 2022. <https://cghlewis.github.io/mpsi-data-training/share.html>.

———. “Data Storage and Security Checklist.” Accessed November 2, 2022. <https://cghlewis.github.io/mpsi-data-training/storage.html>.

———. “Data Tracking Checklist.” Accessed November 2, 2022. <https://cghlewis.github.io/mpsi-data-training/tracking.html>.

———. “Documentation Checklist.” Accessed November 2, 2022. <https://cghlewis.github.io/mpsi-data-training/document.html>.

———. “Roles and Responsibilities Checklist.” Accessed November 1, 2022. <https://cghlewis.github.io/mpsi-data-training/randr.html>.

———. “Task Management Checklist.” Accessed January 10, 2023. <https://cghlewis.github.io/mpsi-data-training/taskmgmt.html>.

———. “Using a Data Dictionary as Your Roadmap to Quality Data. Crystal Lewis.” Accessed January 17, 2023. <https://cghlewis.com/blog/data_dictionary/>.

Library, Lamar Soutter. “NECDMC - Modules for Managing Research Data Lamar Soutter Library - University of Massachusetts Medical School - Module 3.” Accessed February 10, 2023. <https://library.umassmed.edu/resources/necdmc/modules>.

Library, Samuel J. Wood. “Research Data Management, Retention, and Sharing Weill Cornell Medicine Samuel j. Wood Library.” Accessed January 18, 2023. <https://library.weill.cornell.edu/research-support/research-data-management-retention-and-sharing>.

Library, UC Merced. “What Is a Data Dictionary? UC Merced Library.” Accessed January 17, 2023. <https://library.ucmerced.edu/data-dictionaries>.

Lindsay C. Page, University of Pittsburgh, Harvard Graduate School of Education Matthew A. Lenard, and University of Pennsylvania Luke Keele. “The Design of Clustered Observational Studies in Education,” November 23, 2020. <https://doi.org/10.3886/E121381V1>.

Logan, Jessica A. R., Sara A. Hart, and Christopher Schatschneider. “Data Sharing in Education Science.” *AERA Open* 7 (January 2021): 233285842110064. <https://doi.org/10.1177/23328584211006475>.

Markowetz, Florian. “Five Selfish Reasons to Work Reproducibly.” *Genome Biology* 16, no. 1 (December 8, 2015): 274. <https://doi.org/10.1186/s13059-015-0850-7>.

MaryQiu1987. “Database Normalization Description - Office,” May 5, 2022. <https://learn.microsoft.com/en-us/office/troubleshoot/access/database-normalization-description>.

Mehr, Samuel. “Common Topics in Lab Handbooks,” April 2, 2020. <https://www.rsb.org.uk/images/biologist/2020/Apr_May_2020_67-2/67.2_Handbook.pdf>.

———. “How to… Write a Lab Handbook. RSB.” Accessed January 12, 2023. <https://www.rsb.org.uk//biologist-features/how-to-write-a-lab-handbook>.

Microsoft. “Restrictions and Limitations in OneDrive and SharePoint - Microsoft Support.” Accessed February 17, 2023. <https://support.microsoft.com/en-us/office/restrictions-and-limitations-in-onedrive-and-sharepoint-64883a5d-228e-48f5-b3d2-eb39e07630fa>.

Midgley, Carol. “Manual for the Patterns of Adaptive Learning Scales,” 2000. <http://websites.umich.edu/~pals/PALS%202000_V13Word97.pdf>.

Mines, Colorado School of. “Table of Responsibilities Timeline,” n.d. <https://libguides.mines.edu/ld.php?content_id=49061255>.

Minnesota Libraries, University of. “Examples of Data Management Plans University of Minnesota Libraries.” Accessed February 9, 2023. <https://www.lib.umn.edu/services/data/dmp-examples>.

Missouri, University of. “Social/Behavioral/Educational Research Protocol.” Accessed January 12, 2023. <https://research.missouri.edu/human-subjects-research/researcher-resources>.

Neild, R. C., D. Robinson, and J. Agufa. “Sharing Study Data: A Guide for Education Researchers. (NCEE 2022-004).” *U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.*, 2022. <https://ies.ed.gov/ncee/pubs/2022004/pdf/2022004.pdf>.

Nguyen, Kim. “Relational Database Schema Design Overview. Medium,” October 4, 2017. <https://medium.com/@kimtnguyen/relational-database-schema-design-overview-70e447ff66f9>.

NUCATS. “Standard Operating Procedures (SOPs).” Accessed January 13, 2023. <https://www.nucats.northwestern.edu/docs/cecd/overview-of-sops.pdf>.

———. “Writing Standard Operating Procedures.” Accessed January 13, 2023. <https://www.nucats.northwestern.edu/docs/cecd/writing-sops-and-templates.pdf>.

One, PLOS. “Data Availability,” n.d. <https://journals.plos.org/plosone/s/data-availability>.

Open Science, Center for. “COS Engagement with the Education Community,” 2022. <https://docs.google.com/presentation/d/1LpyVOj8oJPr3SVkRM2GfCFnl2Qeo10YbbqcqwtwrVUM>.

———. “Creating a Data Management Plan (DMP) Document - OSF Support.” Accessed January 9, 2023. <https://help.osf.io/article/144-creating-a-data-management-plan-dmp-document>.

———. “How to Make a Data Dictionary - OSF Support.” Accessed February 10, 2023. <https://help.osf.io/article/217-how-to-make-a-data-dictionary>.

Oregon, Pacific University. “Data Security and Storage. Pacific University,” March 31, 2014. <https://www.pacificu.edu/academics/research/scholarship-and-sponsored-projects/research-compliance-integrity/institutional-review-board/irb-policies-recommended-practices/data-security-storage>.

Program, CITI. “CITI Good Clinical Practice SOP Template,” n.d. <https://cghlewis.github.io/mpsi-data-training/img/sample_sop.pdf>.

“R: The r Base Package.” Accessed February 16, 2023. <https://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html>.

Renbarger, Rachel, Jill L. Adelson, Joshua Rosenberg, Sondra M Stegenga, Olivia Lowrey, Pamela Rose Buckley, and Qiyang Zhang. “Champions of Transparency in Education: What Journal Reviewers Can Do to Encourage Open Science Practices.” Preprint. EdArXiv, November 15, 2022. <https://doi.org/10.35542/osf.io/xqfwb>.

Research Data Management, Danish National Forum for. “Metadata - How to FAIR.” Accessed January 18, 2023. <https://howtofair.dk/how-to-fair/metadata/>.

“Retraction Watch.” Accessed January 9, 2023. <https://retractionwatch.com/>.

Reynolds, Tara, Christopher Schatschneider, and Jessica Logan. “The Basics of Data Management.” figshare, April 26, 2022. <https://doi.org/10.6084/m9.figshare.13215350.v2>.

Riederer, Emily. “Column Names as Contracts. Emily Riederer,” September 6, 2020. <https://emilyriederer.netlify.app/post/column-name-contracts/>.

“Schema.org - Schema.org.” Accessed January 19, 2023. <https://www.schema.org/>.

Science {and} Technology Policy, Office of. “OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay. The White House,” 2022. <https://www.whitehouse.gov/ostp/news-updates/2022/08/25/ostp-issues-guidance-to-make-federally-funded-research-freely-available-without-delay/>.

Science, Center for Open. “Center for Open Science.” Accessed October 21, 2022. <https://www.cos.io>.

Service, UK Data. “Costing. UK Data Service.” Accessed January 11, 2023. <https://ukdataservice.ac.uk/learning-hub/research-data-management/plan-to-share/costing/>.

———. “Data Management Costing Tool and Checklist,” 2022. <https://ukdataservice.ac.uk//app/uploads/costingtool.pdf>.

———. “Metadata. UK Data Service.” Accessed January 18, 2023. <https://ukdataservice.ac.uk/learning-hub/research-data-management/document-your-data/metadata/>.

———. “Quality. UK Data Service.” Accessed January 13, 2023. <https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/quality/>.

———. “Roles and Responsibilities. UK Data Service.” Accessed January 10, 2023. <https://ukdataservice.ac.uk/learning-hub/research-data-management/plan-to-share/roles-and-responsibilities/>.

———. “Versioning. UK Data Service.” Accessed January 17, 2023. <https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/versioning/>.

Staff, Editorial. “How to Add and Use Tags in Windows 10? WebNots,” March 5, 2019. <https://www.webnots.com/how-to-add-and-use-tags-in-windows-10/>.

“Standards for Excellence in Education Research - Standards for Excellence in Education Research.” Accessed October 21, 2022. <https://ies.ed.gov/seer/index.asp>.

Stangroom, Jeremy. “Rules for Naming Variables in SPSS - Quick SPSS Tutorial. EZ SPSS Tutorials,” August 7, 2019. <https://ezspss.com/rules-for-naming-variables-in-spss/>.

Strand, Julia. “Error Tight: Exercises for Lab Groups to Prevent Research Mistakes.” Accessed October 31, 2022. <https://psyarxiv.com/rsn5y/>.

“Style Guide.” In *Wikipedia*, January 4, 2023. <https://en.wikipedia.org/w/index.php?title=Style_guide&oldid=1131556370>.

“SYNC Lab Wiki.” Accessed January 12, 2023. <https://eur-synclab.github.io/>.

Tenopir, Carol, Suzie Allard, Priyanki Sinha, Danielle Pollock, Jess Newman, Elizabeth Dalton, Mike Frame, and Lynn Baird. “Data Management Education from the Perspective of Science Educators.” *International Journal of Digital Curation* 11, no. 1 (October 6, 2016): 232–51. <https://doi.org/10.2218/ijdc.v11i1.389>.

Tourangeau, Karen. “Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-k:2011).” Institute of Education Sciences, April 2015. <https://nces.ed.gov/pubs2015/2015074.pdf>.

Twente, University of. “How to Estimate Research Data Management (RDM) Costs,” n.d. <https://www.utwente.nl/en/service-portal/services/lisa/resources/files/library-public/dcc-rdm-costs-estimation.pdf>.

UCDavis. “PMLC - Project Roles and Responsibilities.” Accessed January 10, 2023. <https://tinyurl.com/nhba9mhc>.

United States Department Of Health And Human Services. Administration For Children And Families. Office Of Planning, Research And Evaluation. “Study of Coaching Practices in Early Care and Education Settings (SCOPE), United States, 2019: Version 1 SCOPE Data User Guide.” ICPSR - Interuniversity Consortium for Political; Social Research, 2022. <https://doi.org/10.3886/ICPSR38290.V1>.

University, Utrecht. “Costs of Data Management - Research Data Management Support - Utrecht University.” Accessed October 27, 2022. <https://www.uu.nl/en/research/research-data-management/guides/costs-of-data-management>.

USDA. “Data Dictionary - Blank Template Ag Data Commons.” Accessed February 10, 2023. <https://data.nal.usda.gov/data-dictionary-blank-template>.

USGS. “Tools for Creating Metadata Records. USGS,” n.d. <https://www.usgs.gov/data-management/metadata-creation#tools>.

———. “What Are the Differences Between Data, a Dataset, and a Database? u.s. Geological Survey.” Accessed October 17, 2022. <https://www.usgs.gov/faqs/what-are-differences-between-data-dataset-and-database>.

Valentine, Theresa. “Best Practice: Define Roles and Assign Responsibilities for Data Management. DataOne.” Accessed January 10, 2023. <https://dataoneorg.github.io/Education/bestpractices/define-roles-and>.

Washington, University of. “Developing a Protocol,” n.d. <https://depts.washington.edu/wildfire/resources/protckl.pdf>.

Wickham, Hadley. “Tidy Data.” *Journal of Statistical Software* 59 (September 12, 2014): 1–23. <https://doi.org/10.18637/jss.v059.i10>.

———. *Welcome the Tidyverse Style Guide*. Accessed February 17, 2023. <https://style.tidyverse.org/index.html>.

Wickham, Hadley, and Garrett Grolemund. *8 Workflow: Projects r for Data Science*. Accessed February 17, 2023. <https://r4ds.had.co.nz/workflow-projects.html>.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3, no. 1 (March 15, 2016): 160018. <https://doi.org/10.1038/sdata.2016.18>.

1. Aleata Hubbard, *Data Cleaning in Mathematics Education Research: The Overlooked Methodological Step*, 2017, <https://eric.ed.gov/?id=ED583982>. [↑](#footnote-ref-21)
2. Kristin Briney, *Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success*, Research Skills Series (Exeter, UK: Pelagic Publishing, 2015); Office of Science {and} Technology Policy, “OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay. The White House,” 2022, <https://www.whitehouse.gov/ostp/news-updates/2022/08/25/ostp-issues-guidance-to-make-federally-funded-research-freely-available-without-delay/>. [↑](#footnote-ref-24)
3. John A. Borghi and Ana E. Van Gulick, “Data Management and Sharing: Practices and Perceptions of Psychology Researchers,” *PLOS ONE* 16, no. 5 (May 21, 2021): e0252047, <https://doi.org/10.1371/journal.pone.0252047>. [↑](#footnote-ref-26)
4. A. Busra Ceviren and Jessica Logan, “Ceviren\_logan\_EHE\_forum\_2022.pdf” (presentation, presentation, April 4, 2022), <https://doi.org/10.6084/m9.figshare.19514368.v1>. [↑](#footnote-ref-28)
5. R. C. Neild, D. Robinson, and J. Agufa, “Sharing Study Data: A Guide for Education Researchers. (NCEE 2022-004),” *U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.*, 2022, <https://ies.ed.gov/ncee/pubs/2022004/pdf/2022004.pdf>; Tara Reynolds, Christopher Schatschneider, and Jessica Logan, “The Basics of Data Management” (figshare, April 26, 2022), <https://doi.org/10.6084/m9.figshare.13215350.v2>. [↑](#footnote-ref-30)
6. Melissa Kline, “A Technical Specification for Psychological Datasets. Google Docs,” 2018, <https://docs.google.com/document/d/1u8o5jnWk0Iqp_J06PTu5NjBfVsdoPbBhstht6W0fFp0/edit?usp=embed_facebook>. [↑](#footnote-ref-33)
7. John Borghi and Ana Van Gulick, “Promoting Open Science Through Research Data Management,” *Harvard Data Science Review*, July 28, 2022, <https://doi.org/10.1162/99608f92.9497f68e>. [↑](#footnote-ref-35)
8. Isabel Campos-Varela and Alberto Ruano-Raviña, “Misconduct as the Main Cause for Retraction. A Descriptive Study of Retracted Publications and Their Authors,” *Gaceta Sanitaria* 33, no. 4 (July 1, 2019): 356–60, <https://doi.org/10.1016/j.gaceta.2018.01.009>. [↑](#footnote-ref-38)
9. Lise Doucette and Bruce Fyfe, “Drowning in Research Data: Addressing Data Management Literacy of Graduate Students - PDF Free Download,” 2013, <https://docplayer.net/8853333-Drowning-in-research-data-addressing-data-management-literacy-of-graduate-students.html>. [↑](#footnote-ref-40)
10. Marton Kovacs, Rink Hoekstra, and Balazs Aczel, “The Role of Human Fallibility in Psychological Research: A Survey of Mistakes in Data Management,” *Advances in Methods and Practices in Psychological Science* 4, no. 4 (October 2021): 251524592110459, <https://doi.org/10.1177/25152459211045930>. [↑](#footnote-ref-42)
11. Center for Open Science, “COS Engagement with the Education Community” (2022), <https://docs.google.com/presentation/d/1LpyVOj8oJPr3SVkRM2GfCFnl2Qeo10YbbqcqwtwrVUM>. [↑](#footnote-ref-44)
12. Monya Baker, “1,500 Scientists Lift the Lid on Reproducibility,” *Nature* 533, no. 7604 (May 1, 2016): 452–54, <https://doi.org/10.1038/533452a>. [↑](#footnote-ref-46)
13. Dominic Bordelon, “Guides: Research Data Management @ Pitt: Understanding Research Data Management,” accessed October 13, 2022, <https://pitt.libguides.com/managedata/understanding>. [↑](#footnote-ref-57)
14. National Endowment for the Humanities, “Data Managment Plans for NEH Office of Digital Humanities Proposals and Awards,” 2018, <https://www.neh.gov/sites/default/files/2018-06/data_management_plans_2018.pdf>. [↑](#footnote-ref-59)
15. Borghi and Van Gulick, “Promoting Open Science Through Research Data Management.” [↑](#footnote-ref-62)
16. CDISC, “CDISC Standards in the Clinical Research Process,” accessed January 12, 2023, <https://www.cdisc.org/standards>. [↑](#footnote-ref-63)
17. Lou Burnard, *What Is the Text Encoding Initiative?: How to Add Intelligent Markup to Digital Resources* (OpenEdition Press, 2014), <https://doi.org/10.4000/books.oep.426>. [↑](#footnote-ref-65)
18. Kline, “A Technical Specification for Psychological Datasets.” [↑](#footnote-ref-67)
19. Institute of Education Sciences, “Frequently Asked Questions about Providing Public Access to Data,” accessed October 21, 2022, <https://ies.ed.gov/funding/datasharing_faq.asp>. [↑](#footnote-ref-68)
20. Carol Tenopir et al., “Data Management Education from the Perspective of Science Educators,” *International Journal of Digital Curation* 11, no. 1 (October 6, 2016): 232–51, <https://doi.org/10.2218/ijdc.v11i1.389>. [↑](#footnote-ref-70)
21. Science {and} Technology Policy, “OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay.” [↑](#footnote-ref-72)
22. Borghi and Van Gulick, “Promoting Open Science Through Research Data Management.” [↑](#footnote-ref-73)
23. Science {and} Technology Policy, “OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay.” [↑](#footnote-ref-75)
24. Borghi and Van Gulick, “Promoting Open Science Through Research Data Management.” [↑](#footnote-ref-76)
25. PLOS One, “Data Availability,” n.d., <https://journals.plos.org/plosone/s/data-availability>. [↑](#footnote-ref-77)
26. “Advances in Methods and Practices in Psychological Science. Association for Psychological Science - APS,” accessed February 20, 2023, <https://www.psychologicalscience.org/publications/ampps>. [↑](#footnote-ref-79)
27. Association of Academic Health Science Libraries, Association of American Medical Colleges, and Association of Research Libraries, “Institutional Strategies for the NIH Data Management and Sharing Policy: Infrastructure, Policies, and Services,” September 2022, <https://www.aamc.org/media/62881/download?attachment>. [↑](#footnote-ref-81)
28. Rachel Renbarger et al., “Champions of Transparency in Education: What Journal Reviewers Can Do to Encourage Open Science Practices,” preprint (EdArXiv, November 15, 2022), <https://doi.org/10.35542/osf.io/xqfwb>. [↑](#footnote-ref-83)
29. Jesse M. Alston and Jessica A. Rick, “A Beginner’s Guide to Conducting Reproducible Research,” *The Bulletin of the Ecological Society of America* 102, no. 2 (April 2021), <https://doi.org/10.1002/bes2.1801>. [↑](#footnote-ref-85)
30. Alston and Rick. [↑](#footnote-ref-87)
31. Borghi and Van Gulick, “Promoting Open Science Through Research Data Management”; Wind Cowles, “Research Guides: Research Data Management at Princeton: Home,” accessed September 15, 2022, <https://libguides.princeton.edu/c.php?g=102546&p=665862>; C. Eaker, “What Could Possibly Go Wrong? The Impact of Poor Data Management,” *In Federer, L. (Ed.). The Medical Library Association’s Guide to Data Management for Librarians*, 2016, <https://trace.tennessee.edu/cgi/viewcontent.cgi?article=1023&context=utk_libpub>. [↑](#footnote-ref-88)
32. Oliver W Butters, Rebecca C Wilson, and Paul R Burton, “Recognizing, Reporting and Reducing the Data Curation Debt of Cohort Studies,” *International Journal of Epidemiology* 49, no. 4 (August 1, 2020): 1067–74, <https://doi.org/10.1093/ije/dyaa087>. [↑](#footnote-ref-92)
33. Florian Markowetz, “Five Selfish Reasons to Work Reproducibly,” *Genome Biology* 16, no. 1 (December 8, 2015): 274, <https://doi.org/10.1186/s13059-015-0850-7>. [↑](#footnote-ref-94)
34. Borghi and Gulick, “Data Management and Sharing”; Cowles, “Research Guides.” [↑](#footnote-ref-96)
35. Markowetz, “Five Selfish Reasons to Work Reproducibly.” [↑](#footnote-ref-97)
36. Kovacs, Hoekstra, and Aczel, “The Role of Human Fallibility in Psychological Research”; Julia Strand, “Error Tight: Exercises for Lab Groups to Prevent Research Mistakes,” accessed October 31, 2022, <https://psyarxiv.com/rsn5y/>. [↑](#footnote-ref-98)
37. “Retraction Watch,” accessed January 9, 2023, <https://retractionwatch.com/>. [↑](#footnote-ref-100)
38. “FAIR Principles. GO FAIR,” accessed October 21, 2022, <https://www.go-fair.org/fair-principles/>. [↑](#footnote-ref-104)
39. Mark D. Wilkinson et al., “The FAIR Guiding Principles for Scientific Data Management and Stewardship,” *Scientific Data* 3, no. 1 (March 15, 2016): 160018, <https://doi.org/10.1038/sdata.2016.18>. [↑](#footnote-ref-106)
40. “Standards for Excellence in Education Research - Standards for Excellence in Education Research,” accessed October 21, 2022, <https://ies.ed.gov/seer/index.asp>. [↑](#footnote-ref-109)
41. Wilhelmina van Dijk, Christopher Schatschneider, and Sara A. Hart, “Open Science in Education Sciences,” *Journal of Learning Disabilities* 54, no. 2 (March 2021): 139–52, <https://doi.org/10.1177/0022219420945267>. [↑](#footnote-ref-112)
42. Center for Open Science, “Center for Open Science,” accessed October 21, 2022, <https://www.cos.io>. [↑](#footnote-ref-114)
43. Erin D. Foster and Ariel Deardorff, “Open Science Framework (OSF),” *Journal of the Medical Library Association : JMLA* 105, no. 2 (April 2017): 203–6, <https://doi.org/10.5195/jmla.2017.88>. [↑](#footnote-ref-116)
44. Institute of Education Sciences, “IES Funding Opportunities,” accessed January 9, 2023, <https://ies.ed.gov/funding/23rfas.asp>. [↑](#footnote-ref-118)
45. USGS, “What Are the Differences Between Data, a Dataset, and a Database? u.s. Geological Survey,” accessed October 17, 2022, <https://www.usgs.gov/faqs/what-are-differences-between-data-dataset-and-database>. [↑](#footnote-ref-128)
46. Karl W. Broman and Kara H. Woo, “Data Organization in Spreadsheets,” *The American Statistician* 72, no. 1 (January 2, 2018): 2–10, <https://doi.org/10.1080/00031305.2017.1375989>; Hadley Wickham, “Tidy Data,” *Journal of Statistical Software* 59 (September 12, 2014): 1–23, <https://doi.org/10.18637/jss.v059.i10>. [↑](#footnote-ref-140)
47. USGS, “What Are the Differences Between Data, a Dataset, and a Database? u.s. Geological Survey.” [↑](#footnote-ref-162)
48. Karen Grace-Martin, “The Wide and Long Data Format for Repeated Measures Data. The Analysis Factor,” October 4, 2013, <https://www.theanalysisfactor.com/wide-and-long-data/>. [↑](#footnote-ref-179)
49. The White House, “Executive Order – Making Open and Machine Readable the New Default for Government Information. Whitehouse.gov,” May 9, 2013, <https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government->. [↑](#footnote-ref-188)
50. Science {and} Technology Policy, “OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay.” [↑](#footnote-ref-190)
51. National Institutes of Health, “Writing a Data Management & Sharing Plan,” accessed October 27, 2022, <https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-DMS/writing-a-data-management-and-sharing-plan>. [↑](#footnote-ref-191)
52. Center for Open Science, “Creating a Data Management Plan (DMP) Document - OSF Support,” accessed January 9, 2023, <https://help.osf.io/article/144-creating-a-data-management-plan-dmp-document>. [↑](#footnote-ref-193)
53. Sara Gonzales, Matthew B. Carson, and Kristi Holmes, “Ten Simple Rules for Maximizing the Recommendations of the NIH Data Management and Sharing Plan,” *PLOS Computational Biology* 18, no. 8 (August 3, 2022): e1010397, <https://doi.org/10.1371/journal.pcbi.1010397>; ICPSR, “Guide to Social Science Data Preparation and Archiving: 6th Ed.” accessed January 9, 2023, <https://www.icpsr.umich.edu/web/pages/deposit/guide/>; Open Science, “Creating a Data Management Plan (DMP) Document - OSF Support.” [↑](#footnote-ref-197)
54. Institute of Education Sciences, “Data Sharing,” accessed October 27, 2022, <https://ies.ed.gov/funding/datasharing_implementation.asp>. [↑](#footnote-ref-200)
55. Health, “Writing a Data Management & Sharing Plan.” [↑](#footnote-ref-202)
56. National Institute of Justice, “Data Archiving. National Institute of Justice,” accessed October 27, 2022, <https://nij.ojp.gov/funding/data-archiving>. [↑](#footnote-ref-203)
57. National Science Foundation, “Data Management for NSF EHR Directorate Proposals and Awards,” 2017, <https://www.nsf.gov/bfa/dias/policy/dmpdocs/ehr.pdf>. [↑](#footnote-ref-205)
58. DMPTool, “DMP Templates,” accessed February 9, 2023, <https://dmptool.org/public_templates?page=1&search=&sort_direction=asc&sort_field=templates.title>. [↑](#footnote-ref-209)
59. Sara Hart, “Example of a Data Management Plan” (figshare, November 11, 2020), <https://doi.org/10.6084/m9.figshare.13218743.v1>. [↑](#footnote-ref-211)
60. University of Minnesota Libraries, “Examples of Data Management Plans University of Minnesota Libraries,” accessed February 9, 2023, <https://www.lib.umn.edu/services/data/dmp-examples>. [↑](#footnote-ref-213)
61. ICPSR, “ICPSR Template - Version 1,” February 16, 2023, <https://www.icpsr.umich.edu/files/ICPSR/nih/FINAL_ICPSR-NIH-DMS-Plan-Template_2023.docx>. [↑](#footnote-ref-215)
62. Figshare, “How to Write a Data Management Plan (DMP) and Include Figshare in Your Data Sharing Plans - a Help Article for Using Figshare,” accessed February 21, 2023, <https://help.figshare.com/article/how-to-write-a-data-management-plan-dmp-and-include-figshare-in-your-data-sharing-plans>. [↑](#footnote-ref-217)
63. UK Data Service, “Costing. UK Data Service,” accessed January 11, 2023, <https://ukdataservice.ac.uk/learning-hub/research-data-management/plan-to-share/costing/>. [↑](#footnote-ref-220)
64. Samuel J. Wood Library, “Research Data Management, Retention, and Sharing Weill Cornell Medicine Samuel j. Wood Library,” accessed January 18, 2023, <https://library.weill.cornell.edu/research-support/research-data-management-retention-and-sharing>. [↑](#footnote-ref-222)
65. National Institutes of Health, “Budgeting for Data Management & Sharing Data Sharing,” accessed January 11, 2023, <https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/budgeting-for-data-management-sharing#after>. [↑](#footnote-ref-224)
66. UK Data Service, “Data Management Costing Tool and Checklist,” 2022, <https://ukdataservice.ac.uk//app/uploads/costingtool.pdf>. [↑](#footnote-ref-226)
67. University of Twente, “How to Estimate Research Data Management (RDM) Costs,” n.d., <https://www.utwente.nl/en/service-portal/services/lisa/resources/files/library-public/dcc-rdm-costs-estimation.pdf>. [↑](#footnote-ref-228)
68. Utrecht University, “Costs of Data Management - Research Data Management Support - Utrecht University,” accessed October 27, 2022, <https://www.uu.nl/en/research/research-data-management/guides/costs-of-data-management>. [↑](#footnote-ref-230)
69. Trisha Cruse, “Best Practice: Provide Budget Information for Your Data Management Plan. DataOne,” September 1, 2011, <https://dataoneorg.github.io/Education/bestpractices/provide-budget-information>. [↑](#footnote-ref-232)
70. Joshua Borycz, “Implementing Data Management Workflows in Research Groups Through Integrated Library Consultancy,” *Data Science Journal* 20, no. 1 (February 17, 2021): 9, <https://doi.org/10.5334/dsj-2021-009>. [↑](#footnote-ref-240)
71. Strand, “Error Tight.” [↑](#footnote-ref-242)
72. Eaker, “What Could Possibly Go Wrong? The Impact of Poor Data Management.” [↑](#footnote-ref-243)
73. Alston and Rick, “A Beginner’s Guide to Conducting Reproducible Research.” [↑](#footnote-ref-244)
74. Crystal Lewis, “Roles and Responsibilities Checklist,” accessed November 1, 2022, <https://cghlewis.github.io/mpsi-data-training/randr.html>. [↑](#footnote-ref-247)
75. Crystal Lewis, “Task Management Checklist,” accessed January 10, 2023, <https://cghlewis.github.io/mpsi-data-training/taskmgmt.html>. [↑](#footnote-ref-249)
76. Crystal Lewis, “Documentation Checklist,” accessed November 2, 2022, <https://cghlewis.github.io/mpsi-data-training/document.html>. [↑](#footnote-ref-251)
77. Crystal Lewis, “Data Collection Checklist,” accessed November 2, 2022, <https://cghlewis.github.io/mpsi-data-training/collection.html>. [↑](#footnote-ref-253)
78. Crystal Lewis, “Data Tracking Checklist,” accessed November 2, 2022, <https://cghlewis.github.io/mpsi-data-training/tracking.html>. [↑](#footnote-ref-255)
79. Crystal Lewis, “Data Capture Checklist,” accessed November 2, 2022, <https://cghlewis.github.io/mpsi-data-training/capture.html>. [↑](#footnote-ref-257)
80. Crystal Lewis, “Data Storage and Security Checklist,” accessed November 2, 2022, <https://cghlewis.github.io/mpsi-data-training/storage.html>. [↑](#footnote-ref-259)
81. Crystal Lewis, “Data Cleaning Checklist,” accessed November 2, 2022, <https://cghlewis.github.io/mpsi-data-training/clean.html>. [↑](#footnote-ref-261)
82. Crystal Lewis, “Data Sharing Checklist,” accessed November 2, 2022, <https://cghlewis.github.io/mpsi-data-training/share.html>. [↑](#footnote-ref-263)
83. Borghi and Van Gulick, “Promoting Open Science Through Research Data Management.” [↑](#footnote-ref-265)
84. Borghi and Van Gulick. [↑](#footnote-ref-266)
85. CSP Library Research Desk, “CSP Library: Zotero Guide: Defining Your Research Workflow,” accessed November 1, 2022, <https://library.csp.edu/Zotero/workflow>. [↑](#footnote-ref-273)
86. Borycz, “Implementing Data Management Workflows in Research Groups Through Integrated Library Consultancy.” [↑](#footnote-ref-275)
87. Maya Duru and Sarah Kopper, “Gantt Chart Template,” n.d., <https://www.povertyactionlab.org/sites/default/files/research-resources/rr_grantprop_Template_Gantt_Chart.pdf>. [↑](#footnote-ref-285)
88. Kristin Briney, Heather Coates, and Abigail Goben, “Foundational Practices of Research Data Management,” *Research Ideas and Outcomes* 6 (July 27, 2020): e56508, <https://doi.org/10.3897/rio.6.e56508>. [↑](#footnote-ref-290)
89. Borghi and Van Gulick, “Promoting Open Science Through Research Data Management.” [↑](#footnote-ref-292)
90. Karsten Kryger Hansen, *DataFlowToolkit.dk*, 2017, <https://doi.org/10.5278/16k4-4n24>. [↑](#footnote-ref-294)
91. Matthew Gentzkow and Jesse Shapiro, “Code and Data for the Social Sciences: A Practitioner’s Guide,” 2014, <https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf>. [↑](#footnote-ref-298)
92. UK Data Service, “Roles and Responsibilities. UK Data Service,” accessed January 10, 2023, <https://ukdataservice.ac.uk/learning-hub/research-data-management/plan-to-share/roles-and-responsibilities/>. [↑](#footnote-ref-305)
93. Theresa Valentine, “Best Practice: Define Roles and Assign Responsibilities for Data Management. DataOne,” accessed January 10, 2023, <https://dataoneorg.github.io/Education/bestpractices/define-roles-and>. [↑](#footnote-ref-316)
94. UCDavis, “PMLC - Project Roles and Responsibilities,” accessed January 10, 2023, <https://tinyurl.com/nhba9mhc>; Colorado School of Mines, “Table of Responsibilities Timeline,” n.d., <https://libguides.mines.edu/ld.php?content_id=49061255>. [↑](#footnote-ref-320)
95. Samuel Mehr, “How to… Write a Lab Handbook. RSB,” accessed January 12, 2023, <https://www.rsb.org.uk//biologist-features/how-to-write-a-lab-handbook>. [↑](#footnote-ref-338)
96. Aczel Balazs, “A Crowdsourced Effort to Develop a Lab Manual Template. Google Docs,” accessed January 12, 2023, <https://docs.google.com/document/d/1LqGdtHg0dMbj9lsCnC1QOoWzIsnSNRTSek6i3Kls2Ik/edit?usp=embed_facebook>. [↑](#footnote-ref-340)
97. Balazs. [↑](#footnote-ref-342)
98. Gavin Fay and Ashley Novak, “Faylab Lab Manual,” December 7, 2022, <https://thefaylab.github.io/lab-manual/>. [↑](#footnote-ref-343)
99. Nathaniel von der Embse, “The RAISE Lab Manual,” August 23, 2019, <https://uploads.strikinglycdn.com/files/f3dfcf89-32f2-48cb-94cd-3070f733c53d/RAISE%20Lab%20Guidelines_Final.pdf>. [↑](#footnote-ref-345)
100. Samuel Mehr, “Common Topics in Lab Handbooks,” April 2, 2020, <https://www.rsb.org.uk/images/biologist/2020/Apr_May_2020_67-2/67.2_Handbook.pdf>. [↑](#footnote-ref-347)
101. Mariam Aly, “Aly Lab Wiki,” August 28, 2018, <https://osf.io/mdh87/>. [↑](#footnote-ref-353)
102. “SYNC Lab Wiki,” accessed January 12, 2023, <https://eur-synclab.github.io/>. [↑](#footnote-ref-355)
103. Louise Geraghty and Laura Feeney, “Formalize Research Partnership and Establish Roles and Expectations. The Abdul Latif Jameel Poverty Action Lab (j-PAL),” accessed January 12, 2023, <https://www.povertyactionlab.org/resource/formalize-research-partnership-and-establish-roles-and-expectations>. [↑](#footnote-ref-362)
104. CESSDA, “Data Authenticity - Data Management Expert Guide,” accessed January 13, 2023, <https://dmeg.cessda.eu/Data-Management-Expert-Guide/3.-Process/Data-authenticity>. [↑](#footnote-ref-364)
105. “Style Guide,” in *Wikipedia*, January 4, 2023, <https://en.wikipedia.org/w/index.php?title=Style_guide&oldid=1131556370>. [↑](#footnote-ref-370)
106. Todd Kawakita and Jared Silver, “Strategic Data Project Adopt: Coding Style Guide,” 2013, <https://hwpi.harvard.edu/files/sdp/files/sdp-toolkit-coding-style-guide.pdf>. [↑](#footnote-ref-372)
107. Hadley Wickham, *Welcome the Tidyverse Style Guide*, accessed February 17, 2023, <https://style.tidyverse.org/index.html>. [↑](#footnote-ref-374)
108. Johns Hopkins Institute for Clinical {and} Translational Research, “Best Practices for Research Data Management,” accessed February 10, 2023, <https://ictrweb.johnshopkins.edu/ictr/dmig/Best_Practice/a8376318-ebd6-421f-be63-acf8c88376a1_e0996b36-932a-44d5-94b0-61ad04f5774d.html?v=65849&ip=qam562joeegjlraagoltna>. [↑](#footnote-ref-380)
109. University of Washington, “Developing a Protocol,” n.d., <https://depts.washington.edu/wildfire/resources/protckl.pdf>. [↑](#footnote-ref-386)
110. The Ohio State University Institutional Review Board, “Guidelines for Writing a Research Protocol,” March 6, 2014, <http://orrp.osu.edu/files/2011/10/GuidelinesforWritingaResearchProtocol.pdf>. [↑](#footnote-ref-388)
111. University of Missouri, “Social/Behavioral/Educational Research Protocol,” accessed January 12, 2023, <https://research.missouri.edu/human-subjects-research/researcher-resources>. [↑](#footnote-ref-390)
112. United States Department Of Health And Human Services. Administration For Children And Families. Office Of Planning, Research And Evaluation., “Study of Coaching Practices in Early Care and Education Settings (SCOPE), United States, 2019: Version 1 SCOPE Data User Guide” (ICPSR - Interuniversity Consortium for Political; Social Research, 2022), <https://doi.org/10.3886/ICPSR38290.V1>. [↑](#footnote-ref-392)
113. CONSORT, “Consort - the CONSORT Flow Diagram,” accessed January 13, 2023, <https://www.consort-statement.org/consort-statement/flow-diagram>. [↑](#footnote-ref-398)
114. Karen Tourangeau, “Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-k:2011)” (Institute of Education Sciences, April 2015), <https://nces.ed.gov/pubs2015/2015074.pdf>. [↑](#footnote-ref-403)
115. NUCATS, “Standard Operating Procedures (SOPs),” accessed January 13, 2023, <https://www.nucats.northwestern.edu/docs/cecd/overview-of-sops.pdf>. [↑](#footnote-ref-409)
116. NUCATS, “Writing Standard Operating Procedures,” accessed January 13, 2023, <https://www.nucats.northwestern.edu/docs/cecd/writing-sops-and-templates.pdf>. [↑](#footnote-ref-417)
117. IMPACCT Trials Coordination Centre (ITCC), “Standard Operating Procedure,” n.d., <https://www.uts.edu.au/sites/default/files/2020-07/5.5.5%20Allocation%20of%20Participant%20ID%20Numbers%20V2.5_20200520.pdf>. [↑](#footnote-ref-419)
118. CITI Program, “CITI Good Clinical Practice SOP Template,” n.d., <https://cghlewis.github.io/mpsi-data-training/img/sample_sop.pdf>. [↑](#footnote-ref-421)
119. CESSDA, “Data Authenticity - Data Management Expert Guide”; UK Data Service, “Quality. UK Data Service,” accessed January 13, 2023, <https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/quality/>. [↑](#footnote-ref-425)
120. Neild, Robinson, and Agufa, “Sharing Study Data.” [↑](#footnote-ref-427)
121. UK Data Service, “Versioning. UK Data Service,” accessed January 17, 2023, <https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/versioning/>. [↑](#footnote-ref-432)
122. Crystal Lewis, “Using a Data Dictionary as Your Roadmap to Quality Data. Crystal Lewis,” accessed January 17, 2023, <https://cghlewis.com/blog/data_dictionary/>. [↑](#footnote-ref-443)
123. UC Merced Library, “What Is a Data Dictionary? UC Merced Library,” accessed January 17, 2023, <https://library.ucmerced.edu/data-dictionaries>. [↑](#footnote-ref-445)
124. Johns Hopkins Institute for Clinical {and} Translational Research, “Data Dictionary/Codebook,” accessed February 10, 2023, <https://ictrweb.johnshopkins.edu/ictr/dmig/Best_Practice/a8376318-ebd6-421f-be63-acf8c88376a1_6342a1c3-1a5d-4287-a46e-374824e3780e.html?v=65849&ip=hpdkvlttuiyioooqhw>. [↑](#footnote-ref-447)
125. Carol Midgley, “Manual for the Patterns of Adaptive Learning Scales,” 2000, <http://websites.umich.edu/~pals/PALS%202000_V13Word97.pdf>. [↑](#footnote-ref-452)
126. USDA, “Data Dictionary - Blank Template Ag Data Commons,” accessed February 10, 2023, <https://data.nal.usda.gov/data-dictionary-blank-template>. [↑](#footnote-ref-459)
127. Center for Open Science, “How to Make a Data Dictionary - OSF Support,” accessed February 10, 2023, <https://help.osf.io/article/217-how-to-make-a-data-dictionary>. [↑](#footnote-ref-461)
128. @ ICPSR, “Guide to Codebooks 1st Edition” (Ann Arbor, MI, 2011), <https://www.icpsr.umich.edu/files/deposit/Guide-to-Codebooks_v1.pdf>. [↑](#footnote-ref-465)
129. United States Department Of Health And Human Services. Administration For Children And Families. Office Of Planning, Research And Evaluation., “Study of Coaching Practices in Early Care and Education Settings (SCOPE), United States, 2019.” [↑](#footnote-ref-470)
130. Crystal Lewis, “Codebook Package Comparison,” accessed January 17, 2023, <https://github.com/Cghlewis/codebook-pkg-comparison>. [↑](#footnote-ref-474)
131. ICPSR, “Guide to Codebooks 1st Edition.” [↑](#footnote-ref-476)
132. United States Department Of Health And Human Services. Administration For Children And Families. Office Of Planning, Research And Evaluation., “Study of Coaching Practices in Early Care and Education Settings (SCOPE), United States, 2019.” [↑](#footnote-ref-477)
133. CESSDA, “Documentation and Metadata,” accessed February 9, 2023, <https://dmeg.cessda.eu/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata>; Danish National Forum for Research Data Management, “Metadata - How to FAIR,” accessed January 18, 2023, <https://howtofair.dk/how-to-fair/metadata/>. [↑](#footnote-ref-480)
134. “FAIR Principles”; UK Data Service, “Metadata. UK Data Service,” accessed January 18, 2023, <https://ukdataservice.ac.uk/learning-hub/research-data-management/document-your-data/metadata/>; Jessica A. R. Logan, Sara A. Hart, and Christopher Schatschneider, “Data Sharing in Education Science,” *AERA Open* 7 (January 2021): 233285842110064, <https://doi.org/10.1177/23328584211006475>. [↑](#footnote-ref-483)
135. CESSDA, “Documentation and Metadata”; University of Iowa Libraries, “University Libraries the University of Iowa,” accessed February 10, 2023, <https://www.lib.uiowa.edu/>. [↑](#footnote-ref-486)
136. Research Data Management, “Metadata - How to FAIR”; Melanie Cofield, “LibGuides: Metadata Basics: Key Concepts,” accessed February 10, 2023, <https://guides.lib.utexas.edu/metadata-basics/key-concepts>. [↑](#footnote-ref-488)
137. “Figshare - Credit for All Your Research. Figshare,” accessed January 18, 2023, <https://figshare.com/>. [↑](#footnote-ref-490)
138. Wasila Dahdul, “Research Guides: Research Data Management: Describing Data,” accessed January 18, 2023, <https://guides.lib.uci.edu/datamanagement/describe>. [↑](#footnote-ref-495)
139. LDbase, “Information to Gather Before Uploading Your Data LDbase,” accessed January 18, 2023, <https://www.ldbase.org/resources/user-guide/information-to-gather>; ICPSR, “ICPSR, Data Management, Metadata,” accessed January 18, 2023, <https://www.icpsr.umich.edu/web/pages/datamanagement/lifecycle/metadata.html>; Rick O. Gilmore, Joy Lorenzo Kennedy, and Karen E. Adolph, “Practical Solutions for Sharing Data and Materials from Psychological Research,” *Advances in Methods and Practices in Psychological Science* 1, no. 1 (March 2018): 121–30, <https://doi.org/10.1177/2515245917746500>. [↑](#footnote-ref-500)
140. ICPSR, “ICPSR, Data Management, Metadata.” [↑](#footnote-ref-504)
141. University of Pittsburgh Lindsay C. Page, Harvard Graduate School of Education Matthew A. Lenard, and University of Pennsylvania Luke Keele, “The Design of Clustered Observational Studies in Education,” November 23, 2020, <https://doi.org/10.3886/E121381V1>. [↑](#footnote-ref-505)
142. ICPSR, “ICPSR, Data Management, Metadata.” [↑](#footnote-ref-510)
143. Erin M. Buchanan et al., “Getting Started Creating Data Dictionaries: How to Create a Shareable Data Set,” *Advances in Methods and Practices in Psychological Science* 4, no. 1 (January 2021): 251524592092800, <https://doi.org/10.1177/2515245920928007>; Ruben C. Arslan, “How to Automatically Document Data with the Codebook Package to Facilitate Data Re-Use,” preprint (PsyArXiv, September 5, 2018), <https://doi.org/10.31234/osf.io/5qc6h>; USGS, “Tools for Creating Metadata Records. USGS,” n.d., <https://www.usgs.gov/data-management/metadata-creation#tools>. [↑](#footnote-ref-511)
144. Iowa Libraries, “University Libraries the University of Iowa.” [↑](#footnote-ref-515)
145. Mike Bolam, “Guides: Metadata & Discovery @ Pitt: Metadata Standards,” accessed January 18, 2023, <https://pitt.libguides.com/metadatadiscovery/metadata-standards>; DDI Alliance, “Controlled Vocabularies - Overview Table of Latest Versions Data Documentation Initiative,” accessed January 19, 2023, <https://ddialliance.org/controlled-vocabularies>. [↑](#footnote-ref-516)
146. Kline, “A Technical Specification for Psychological Datasets.” [↑](#footnote-ref-519)
147. “Schema.org - Schema.org,” accessed January 19, 2023, <https://www.schema.org/>. [↑](#footnote-ref-520)
148. Education Sciences, “Frequently Asked Questions about Providing Public Access to Data.” [↑](#footnote-ref-522)
149. DDI Alliance, “Mapping to Dublin Core (DDI Version 2),” accessed January 19, 2023, <https://ddialliance.org/resources/ddi-profiles/dc>. [↑](#footnote-ref-526)
150. Lindsay C. Page, Matthew A. Lenard, and Luke Keele, “The Design of Clustered Observational Studies in Education.” [↑](#footnote-ref-531)
151. Lamar Soutter Library, “NECDMC - Modules for Managing Research Data Lamar Soutter Library - University of Massachusetts Medical School - Module 3,” accessed February 10, 2023, <https://library.umassmed.edu/resources/necdmc/modules>. [↑](#footnote-ref-535)
152. Figshare, “Fair-Figshare,” accessed February 21, 2023, <https://knowledge.figshare.com/publisher/fair-figshare>. [↑](#footnote-ref-537)
153. ICPSR, “ICPSR, Data Management, Metadata.” [↑](#footnote-ref-539)
154. Iowa Libraries, “University Libraries the University of Iowa.” [↑](#footnote-ref-540)
155. “Style Guide.” [↑](#footnote-ref-548)
156. Tatiana Baquero Cakici, “Folders v. Metadata in SharePoint Document Libraries. Enterprise Knowledge,” April 5, 2017, <https://enterprise-knowledge.com/folders-v-metadata-sharepoint-document-libraries/>; Editorial Staff, “How to Add and Use Tags in Windows 10? WebNots,” March 5, 2019, <https://www.webnots.com/how-to-add-and-use-tags-in-windows-10/>. [↑](#footnote-ref-550)
157. Microsoft, “Restrictions and Limitations in OneDrive and SharePoint - Microsoft Support,” accessed February 17, 2023, <https://support.microsoft.com/en-us/office/restrictions-and-limitations-in-onedrive-and-sharepoint-64883a5d-228e-48f5-b3d2-eb39e07630fa>. [↑](#footnote-ref-553)
158. Alvinashcraft, “Naming Files, Paths, and Namespaces - Win32 Apps,” December 15, 2022, <https://learn.microsoft.com/en-us/windows/win32/fileio/naming-a-file>. [↑](#footnote-ref-555)
159. “Documents. Xkcd,” accessed February 16, 2023, <https://xkcd.com/1459/>. [↑](#footnote-ref-561)
160. “R: The r Base Package,” accessed February 16, 2023, <https://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html>; Jeremy Stangroom, “Rules for Naming Variables in SPSS - Quick SPSS Tutorial. EZ SPSS Tutorials,” August 7, 2019, <https://ezspss.com/rules-for-naming-variables-in-spss/>. [↑](#footnote-ref-564)
161. Emily Riederer, “Column Names as Contracts. Emily Riederer,” September 6, 2020, <https://emilyriederer.netlify.app/post/column-name-contracts/>. [↑](#footnote-ref-567)
162. Reynolds, Schatschneider, and Logan, “The Basics of Data Management.” [↑](#footnote-ref-569)
163. Tourangeau, “Early Childhood Longitudinal Study, Kindergarten Class of 2010–11 (ECLS-k.” [↑](#footnote-ref-573)
164. Dr Timothy S. Farewell, “My Easy r Script Header Template – Tim Farewell,” April 10, 2018, <https://timfarewell.co.uk/my-r-script-header-template/>; Gergana Daskalova, “Coding Etiquette. Coding Club,” accessed February 17, 2023, <https://ourcodingclub.github.io/tutorials/etiquette/>. [↑](#footnote-ref-578)
165. Hadley Wickham and Garrett Grolemund, *8 Workflow: Projects r for Data Science*, accessed February 17, 2023, <https://r4ds.had.co.nz/workflow-projects.html>. [↑](#footnote-ref-581)
166. Pacific University Oregon, “Data Security and Storage. Pacific University,” March 31, 2014, <https://www.pacificu.edu/academics/research/scholarship-and-sponsored-projects/research-compliance-integrity/institutional-review-board/irb-policies-recommended-practices/data-security-storage>. [↑](#footnote-ref-588)
167. MaryQiu1987, “Database Normalization Description - Office,” May 5, 2022, <https://learn.microsoft.com/en-us/office/troubleshoot/access/database-normalization-description>; Ph D. David T. Bourgeois and Bourgeois, *Information Systems for Business and Beyond* (Published through the Open Textbook Challenge by the Saylor Academy, 2014), <https://pressbooks.pub/bus206/>. [↑](#footnote-ref-591)
168. Kim Nguyen, “Relational Database Schema Design Overview. Medium,” October 4, 2017, <https://medium.com/@kimtnguyen/relational-database-schema-design-overview-70e447ff66f9>. [↑](#footnote-ref-594)