

Data Management in Large-Scale Education Research

Crystal Lewis

2022-10-31

Contents

Preamble	5
Introduction	5
Why this book	6
About this book	8
Who this book is for	9
Final note	9
Acknowledgements	10
 Research Data Management	 11
What is research data management?	11
Standards	11
Why care about research data management?	12
Existing Frameworks	15
Terminology	17
The Research Life Cycle	18
 Data Structure	 21
Basics of a dataset	21
Dataset organization rules	23
Linking data	27
File types	32

Data Management Plan	33
History and purpose	33
What is it?	34
Getting help	36
Budgeting	37
Planning Data Management	39
Why spend time on planning?	39
Planning checklists	39
How to move from a planning checklist to a workflow	39
Project Roles and Responsibilities	41
Why it's important to assign roles	41
Typical roles in a research project	41
Documentation	43
What is documentation?	43
Why is documentation important?	43
Team Level	43
Project Level	43
Dataset Level	43
Variable Level	43
Data Tracking	45
Why track data?	45
Build a system	45
Creating participant IDs	45
When to build it, who builds it, tools to build it in	45
Data Collection	47
Why consider data management in data collection?	47
Consents	47
Electronic data collection instruments	47

<i>CONTENTS</i>	5
Paper data collection instruments	47
Interviews/cocus groups	47
Data Capture	49
Electronic data capture	49
Paper data capture	49
Extant data	49
Data Storage and Security	51
Types of data you'll be storing	51
General security rules	51
Participant tracking database	51
Electronic data	51
Detachable media	51
Audio/visual data	51
Paper data	51
Sharing data	51
Data Cleaning	53
Foundational knowledge	53
Data structure	53
Data cleaning plan	53
Data validation	53
Why use code?	53
Data Sharing	55
Why share your data?	55
Considering FAIR principles	55
Best practices	55
Retractions and revisions	55
Wrapping It Up	57
Connecting practices to outcomes	57
Putting in the work	57

Call to Action	59
Last thoughts	59
Training for future researchers	59
Investing in data management and data managers	59
Appendices	61

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.

The results of educational research studies are only as accurate as the data used to produce them. - Aleata Hubbard¹

Introduction

In 2013, without knowing that the term research data management existed, I accepted a job as a Research Associate with a prevention science 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. ECKLS-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.

¹Aleata Hubbard, *Data Cleaning in Mathematics Education Research: The Overlooked Methodological Step*, 2017, <https://eric.ed.gov/?id=ED583982>.

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 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.

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 Gulick³ 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 Logan⁴ 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 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

²Kristin Briney, *Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success*, Research Skills Series (Exeter, UK: Pelagic Publishing, 2015).

³John 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>.

⁴A. 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>.

⁵R. 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>.

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 share data,⁶ 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

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 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 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 In their 2021 study of 488 researchers who had published in a psychology journal between 2010 and 2018, Kovacs et al.¹⁰ 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

⁶“Psych-DS Specification. Google Docs,” accessed September 16, 2022, https://docs.google.com/document/d/1u8o5jnWk0Iqp_J06PTu5NjBFVsd0PbBhstht6W0fFp0/edit?usp=embed_facebook.

⁷John 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>.

⁸Isabel 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>.

⁹Lise 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>.

¹⁰Marton 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>.

their research data, about 10% mentioned “being nervous about mistakes” as a reason for not sharing.¹¹ 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

About this book

While the field as a whole may not have agreed upon guidelines for data management, there are still best practices that are proven to result in more usable, 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.

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.

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

¹¹Open Science Foundation, “COS Engagement with the Education Community” (2022), <https://docs.google.com/presentation/d/1LpyVOj8oJPr3SVkRM2GfCFnl2Qeo10YbbqcqwtwrVUM>.

¹²Monya Baker, “1,500 Scientists Lift the Lid on Reproducibility,” *Nature* 533, no. 7604 (May 1, 2016): 452–54, <https://doi.org/10.1038/533452a>.

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 implement 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.

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 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.

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!

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 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:

Research Data Management

What is research data management?

Research data management (RDM) involves the organization, storage, preservation, and dissemination of research study data.¹³ Research study data includes materials generated or collected throughout a research process.¹⁴ 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.

Standards

Data management standards refer to how data should be stored, organized, and described.¹⁵ While some fields have generally adopted metadata standards developed by organizations such as the Data Documentation Initiative¹⁶ and Dublin Core¹⁷ (we'll talk about this more in documentation []), and some have adopted data structure standards developed by organizations like the Text Encoding Initiative,¹⁸ it is common knowledge that there are no agreed-upon norms for how to structure and share data in the field of education.¹⁹ The rules for

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

¹⁴National Endowment for the Humanities, "Data Management 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.

¹⁵Borghi and Van Gulick, "Promoting Open Science Through Research Data Management."

¹⁶"Welcome to the Data Documentation Initiative | Data Documentation Initiative," accessed October 21, 2022, <https://ddialliance.org/>.

¹⁷"DCMI Metadata Terms," accessed October 21, 2022, <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>.

¹⁸"TEI: Text Encoding Initiative," accessed October 21, 2022, <https://tei-c.org/>.

¹⁹IES, "Frequently Asked Questions about Providing Public Access to Data," accessed October 21, 2022, https://ies.ed.gov/funding/datasharing_faq.asp.

what data should be produced and how it should be documented is often left up to each individual team, as long as external requirements of the IRB and funders are met.²⁰ However, with a growing interest in open science practices and expanding requirements for federally funded research to make data publicly available,²¹ data repositories will most likely begin to play a stronger role in promoting standards around data formats and documentation.²²

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.

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.

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 proposal²³. 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

²⁰Carol 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>.

²¹Office 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/>.

²²Borghi and Van Gulick, “Promoting Open Science Through Research Data Management.”

²³Science {and} Technology Policy, “OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay.”

data sharing policy, comes the incentive to manage your data for the purposes of data sharing.²⁴

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 ONE²⁵ as an example). 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, they will monitor how you manage your data. They care about 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.²⁶
4. **Open science practices:** With a growing interest in open science practices, sharing well managed data, 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).²⁷ Sharing data that has been managed in a reproducible way also allows others to 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.²⁸ 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.²⁹

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

²⁴Borghi and Van Gulick, “Promoting Open Science Through Research Data Management.”

²⁵PLOS One, “Data Availability,” n.d., <https://journals.plos.org/plosone/s/data-availability>.

²⁶Association 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>.

²⁷Jesse 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>.

²⁸Alston and Rick.

²⁹Borghi 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.

standardized way.

1. **Facilitates use of your data:** Every member of your research team being able to find and understand the data and documentation that they need is a huge benefit. It allows for the easy use and re-use of your data, and hastens efforts like the publication process.³⁰ 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 writing and less time playing detective.
2. **Encourages validation:** Implementing reproducible data management practices encourages and allows your team to internally validate and replicate your processes to ensure your outputs are accurate.
3. **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.³¹ Furthermore, good data management enables continuity when handing off projects to collaborators or when picking up your own projects after a long hiatus.³²
4. **Increases efficiency:** Documenting and automating tasks reduces duplication of efforts for repeating tasks, especially in longitudinal studies.
5. **Reduces data curation debt:** Taking the time to implement quality data management through the entire research study reduces data curation debt caused by suboptimal data management practices.³³ Having poorly 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 at the end of your study scrambling to get your data up to acceptable standards.
6. **Upholds research integrity:** Errors come in many forms, from both humans and technology. We've seen evidence of this in the papers cited as being retracted for "unreliable data" in the blog Retraction Watch. 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.

³⁰Florian 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>.

³¹Borghi and Gulick, "Data Management and Sharing"; Cowles, "Research Guides."

³²Markowetz, "Five Selfish Reasons to Work Reproducibly."

³³Oliver 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>.

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.

FAIR

In 2016, the FAIR Principles³⁴ 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.³⁵ The principles are:

F: Findable

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

A: Accessible

Your data is accessible if humans can access your data. This can mean your data is available in a repository or through a request system.

I: Interoperable

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 such as .csv, .txt, .dat, etc. Furthermore, thorough data documentation should accompany data to allow that data to be interoperable.

R: Reusable

Your metadata should provide information on the broad context of your project as well as your data collection to allow for accurate use of your data. You should also have clear licensing for data use.

³⁴“FAIR Principles. GO FAIR,” accessed October 21, 2022, <https://www.go-fair.org/fair-principles/>.

³⁵Mark 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>.

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.³⁶ 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

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, pre-registration, and open access.³⁷ Organizations such as the Center for Open Science,³⁸ have become a well-known proponents of open science, offering the open science framework (OSF)³⁹ 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.

³⁶“Standards for Excellence in Education Research - Standards for Excellence in Education Research,” accessed October 21, 2022, <https://ies.ed.gov/seer/index.asp>.

³⁷Wilhelmina 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>.

³⁸Center for Open Science, “Center for Open Science,” accessed October 21, 2022, <https://www.cos.io>.

³⁹Erin 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>.

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.

Term	Other Terms
Anonymous data	NA
Append	NA
Archive	NA
Attrition	NA
Clean data	processed data
Cohort	NA
Confidential data	pseudonymisation, coded data, indirectly identifiable
Confidentiality	NA
Control	business as usual
Cross-sectional	NA
Data	research data
Database	relational database
Dataset	dataframe, spreadsheet
De-identified data	anonymized data
Derived Data	NA
Directory	file structure
Experimental Data	NA
Extant Data	NA
File formats	NA
Identifiable data	NA
Longitudinal	NA
Merge	join, link
Missing data	NA
Observational Data	NA
Participant database	study roster, demographic file, master list, master key, linking key, code key, key c
Path	file path
PII	NA
Privacy	NA
Qualitative data	NA
Quantitative data	NA
Randomized Controlled Trial	RCT
Raw data	primary, untouched
Replicable	NA
Reproducible	NA
Simulation Data	NA
Standardization	NA
Study	NA

Study ID	participant ID, location ID, site ID, unique identifier (UID), subject ID
Subject	case, participant, site, record
Syntax	code, program
Treatment	experiment
Variable	column, field, question
Variable name	header
Wave	time period, time point, event, session

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 the image below how throughout the project, data management and project coordination 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.

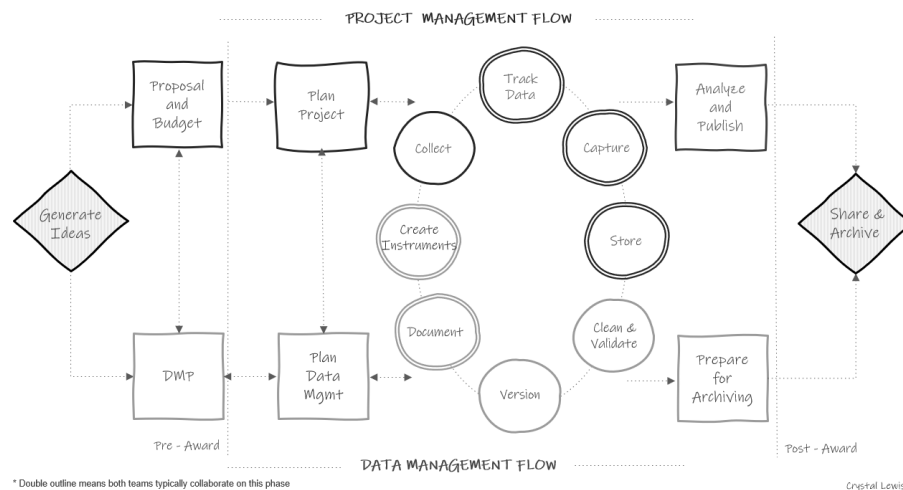


Figure 1: The research project life cycle

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 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 2-5 page 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 project is funded, 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 DMP 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 like data dictionaries, style guides, and protocols.
 - Once documentation is started, both teams begin to create any necessary **data collection instruments**. These instruments will be created with input from the documentation. During this phase the team may also develop their participant tracking database.
 - Next, the project management team moves into the **data collection** phase. This may involve recruitment and consenting, as well as data collection. 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.
 - 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. 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, 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.

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.

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.

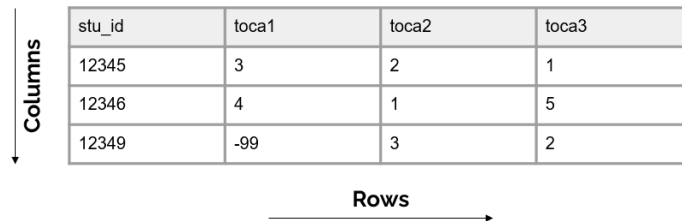
Those data come in many forms (ex: video, transcripts, documents, files), represented as text, numbers, or multimedia.⁴⁰ In the world of quantitative education research, we are often working with digital data in the form of a dataset, a structured collection of data. These datasets are organized in a 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.

Columns

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

⁴⁰USGS, “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>.



stu_id	toca1	toca2	toca3
12345	3	2	1
12346	4	1	5
12349	-99	3	2

Figure 2: Basic format of a dataset

- 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).
- 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 section .

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 ??.
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?”).

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.

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 the fall of 2022. 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).

stu_id	age	score_a	score_b
1234	12	250	150
1235	10	219	176
1236	9	188	160
1237	14	160	119

Figure 3: Representation of a cell value

Dataset organization rules

In order for your dataset to be machine-readable and analyzable, it should adhere to a set of structural rules.⁴¹

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.

⁴¹Karl 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>.

not a rectangle				
	1234	1235	1236	1237
age	12	10	9	14
	1234	1235	1236	1237
score_a	250	219	188	160
	1234	1235	1236	1237
score_b	150	176	158	119

rectangle			
stu_id	age	score_a	score_b
1234	12	250	150
1235	10	219	176
1236	9	188	160
1237	14	160	119

Figure 4: A comparison of non-rectangular and rectangular data

2. 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.

numeric variable	
tch_id	age
12345	22
12346	24
12349	46
12350	36

text variable	
tch_id	age
12345	22
12346	24
12349	49 years old
12350	36..0

Space before 24 makes this entry text

Text added makes this entry text

Double decimal point makes this entry text

Figure 5: A comparison of variables adhering and not adhering to a data type

3. 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)

two things in one variable			two things in two variables			
sch_id	level	incident_rate	sch_id	level	incident	enrollment
235	elementary	55/250	235	elementary	55	250
236	elementary	72/303	236	elementary	72	303
237	middle	140/410	237	middle	140	410
238	high	219/552	238	high	219	552

Figure 6: A comparison of two things being measured in one variable and two things being measured across two variables

4. 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
 - 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
5. 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.

not explicit values				explicit values			
sch_id	year	grade	n_students	sch_id	year	grade	n_students
204	2020	3	100	204	2020	3	100
		4	80	204	2020	4	80
		5	90	204	2020	5	90
205	2020	3	98	205	2020	3	98
		4	88	205	2020	4	88
		5	91	205	2020	5	91

Figure 7: A comparison of variables with empty cells and variables with not empty cells

not explicit values			explicit values			
stu_id	date	test_score	stu_id	date	test_score	treatment
12345	2022-04-13	35	12345	2022-04-13	35	0
12346	2022-04-12	42	12346	2022-04-12	42	1
12349	2022-04-13	50	12349	2022-04-13	50	1
12350	2022-04-11	19	12350	2022-04-11	19	0

Cell color indicates
treatment condition

Figure 8: A comparison of variables with implicit values and variables with explicit values

- 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). 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 9: A comparison of data with duplicate cases and data with no duplicate cases

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.

Database design

A database is “an organized collection of data stored as multiple datasets.”⁴² 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”. And 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 ??.

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. The image below shows what variables were collected from each instrument and how each table can be linked together through a primary key (circled in yellow).

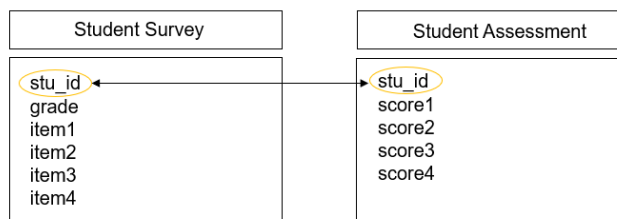


Figure 10: Linking data through primary keys

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. The image below 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).

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. Here is

⁴²USGS, “What Are the Differences Between Data, a Dataset, and a Database? | u.s. Geological Survey.”

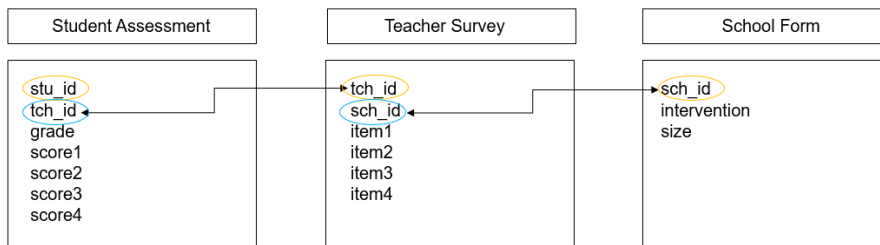


Figure 11: Linking data through foreign keys

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.

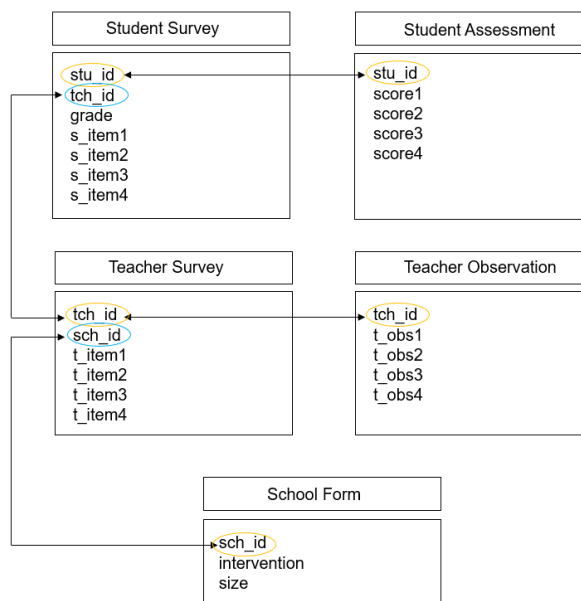


Figure 12: Linking data through primary and foreign keys

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 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.

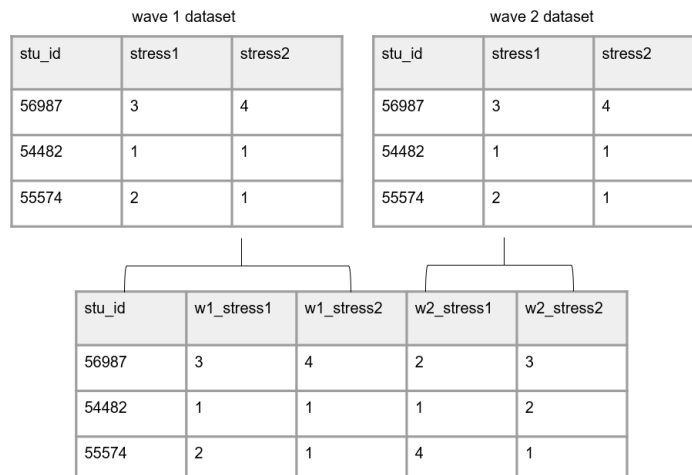


Figure 13: Data structured in wide format

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.

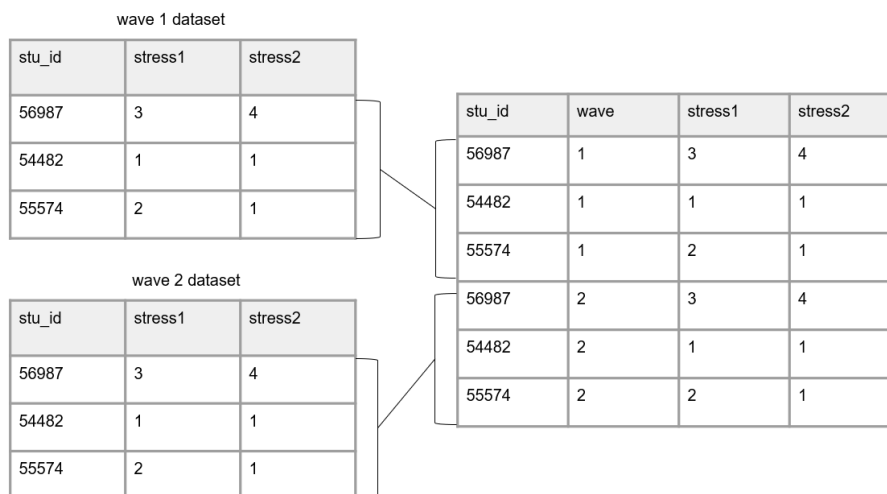


Figure 14: Data structured in long format

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.⁴³ We will further review decision making around data structure in our data cleaning chapter ??.

⁴³Karen 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/>.

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 ??.

Data Management Plan

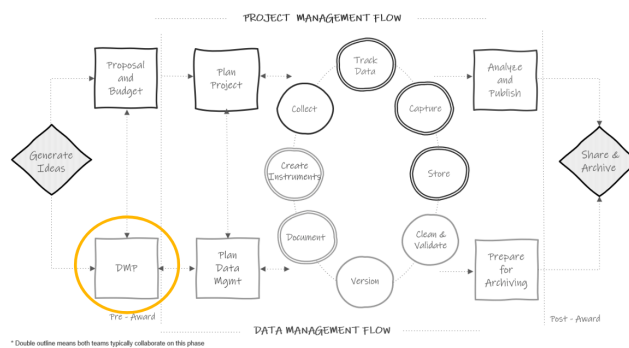


Figure 15: Data management plan in the research project life cycle

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 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.”⁴⁴ 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

⁴⁴The 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->.

federally funded publications and their supporting data accessible to the public with no embargo on their release.⁴⁵

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 world⁴⁶.

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.

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. 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?

⁴⁵Science {and} Technology Policy, "OSTP Issues Guidance to Make Federally Funded Research Freely Available Without Delay."

⁴⁶"Creating a Data Management Plan (DMP) Document - OSF Support," accessed October 28, 2022, <https://help.osf.io/article/144-creating-a-data-management-plan-dmp-document>.

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 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 metadata will you create? (Consider project level, dataset level and variable level metadata)
- What format will your documentation be in? (Ex: xml, csv, pdf)
- What other documentation do you plan to include when sharing data? (Ex: code, data collection instruments, protocols)

5. Standards

- Are there any data or documentation standards being used? (Ex: DDI)

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 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 Sciences⁴⁷
- National Institutes of Health⁴⁸
- National Institute of Justice⁴⁹
- National Science Foundation⁵⁰

Getting help

When constructing your DMP it may be important to enlist help. If you have a data manager or data team, you will most certainly want to consult with them when writing your plan. 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.

⁴⁷Institute of Education Sciences, “Data Sharing,” accessed October 27, 2022, https://ies.ed.gov/funding/datasharing_implementation.asp.

⁴⁸National Institutes of Health, “Writing a Data Management & Sharing Plan | Data Sharing,” accessed October 27, 2022, <https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-DMS/writing-a-data-management-and-sharing-plan>.

⁴⁹National Institute of Justice, “Data Archiving. National Institute of Justice,” accessed October 27, 2022, <https://nij.ojp.gov/funding/data-archiving>.

⁵⁰National Science Foundation, “Data Management for NSF EHR Directorate Proposals and Awards,” 2017, <https://www.nsf.gov/bfa/dias/policy/dmpdocs/ehr.pdf>.

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 from the University of Virginia Library⁵¹. There is also a well-known free online application called the DMPTool⁵² that guides you in constructing a data management plan for many of the large funding agencies you might work with. Their site also has many searchable public DMPs that you can review for inspiration.

Budgeting

As briefly mention above, funding agencies acknowledge that there are costs associated with implementing your data management plan and allow you to list these costs in your budget narrative. Costs associated with the entire data life cycle should be considered and can range from infrastructure costs, such as fees for storage or software, to the salaries required to pay personnel to prepare FAIR datasets that are acceptable for data sharing.

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 Service⁵³, the University of Twente⁵⁴, and Utrecht University⁵⁵ (among others), have put together checklists to help you think through your various potential data management costs.

⁵¹University of Virginia Library Research Data Services, “Data Management Plan Templates | University of Virginia Library Research Data Services + Sciences,” accessed October 27, 2022, <https://data.library.virginia.edu/data-management-plan-templates/>.

⁵²“DMPTool,” accessed October 27, 2022, <https://dmptool.org/>.

⁵³UK Data Service, “Data Management Costing Tool and Checklist,” 2022, <https://ukdataservice.ac.uk/app/uploads/costingtool.pdf>.

⁵⁴University 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>.

⁵⁵Utrecht 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>.

Planning Data Management

Why spend time on planning?

Planning checklists

How to move from a planning checklist to a workflow

Project Roles and Responsibilities

Why it's important to assign roles

Typical roles in a research project

Documentation

What is documentation?

Why is documentation important?

Team Level

Project Level

Dataset Level

Variable Level

Data Tracking

Why track data?

Build a system

Creating participant IDs

When to build it, who builds it, tools to build it
in

Data Collection

Why consider data management in data collection?

Consents

Electronic data collection instruments

Paper data collection instruments

Interviews/focus groups

Data Capture

Electronic data capture

Paper data capture

Extant data

Data Storage and Security

Types of data you'll be storing

General security rules

Participant tracking database

Electronic data

Detachable media

Audio/visual data

Paper data

Sharing data

Data Cleaning

Foundational knowledge

Data structure

Data cleaning plan

Data validation

Why use code?

Data Sharing

Why share your data?

Considering FAIR principles

Best practices

Retractions and revisions

Wrapping It Up

Connecting practices to outcomes

Putting in the work

Call to Action

Last thoughts

Training for future researchers

Investing in data management and data managers

Appendices

- 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>.
- 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>.
- 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>.
- 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>.
- Briney, Kristin. *Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success*. Research Skills Series. Exeter, UK: Pelagic Publishing, 2015.
- 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>.
- 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>.
- 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>.
- 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>.
- Cowles, Wind. "Research Guides: Research Data Management at Princeton: Home." Accessed September 15, 2022. <https://libguides.princeton.edu/c.php?g=102546&p=665862>.
- "Creating a Data Management Plan (DMP) Document - OSF Support." Accessed October 28, 2022. <https://help.osf.io/article/144-creating-a-data-management-plan-dmp-document>.
- "DCMI Metadata Terms." Accessed October 21, 2022. <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>.
- 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." Accessed October 27, 2022. <https://dmptool.org/>.
- 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>.
- 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.
- "FAIR Principles. GO FAIR." Accessed October 21, 2022. <https://www.go-fair.org/fair-principles/>.
- 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>.
- Foundation, Open Science. "COS Engagement with the Education Community," 2022. <https://docs.google.com/presentation/d/1LpyVOj8oJPr3SVkRM2GfCFnl2Qeo10Ybbqcc>.
- 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/>.
- Health, National Institutes of. "Writing a Data Management & Sharing Plan | Data Sharing." 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. White-house.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 Management 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.
- IES. "Frequently Asked Questions about Providing Public Access to Data." Accessed October 21, 2022. https://ies.ed.gov/funding/datasharing_faq.asp.
- Justice, National Institute of. "Data Archiving. National Institute of Justice." Accessed October 27, 2022. <https://nij.ojp.gov/funding/data-archiving>.
- 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>.
- 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>.
- 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>.
- One, PLOS. "Data Availability," n.d. <https://journals.plos.org/plosone/s/data-availability>.
- "Psych-DS Specification. Google Docs." Accessed September 16, 2022. https://docs.google.com/document/d/1u8o5jnWk0Iqp_J06PTu5NjBfVsd0PbBhstht6W0fFp0/edit?usp=embed_facebook.
- 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>.
- 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. "Data Management Costing Tool and Checklist," 2022. <https://ukdataservice.ac.uk/app/uploads/costingtool.pdf>.
- "Standards for Excellence in Education Research - Standards for Excellence in Education Research." Accessed October 21, 2022. <https://ies.ed.gov/seer/index.asp>.
- "TEI: Text Encoding Initiative." Accessed October 21, 2022. <https://tei-c.org/>.
- 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.
- 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>.
- 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>.
- 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>.
- Virginia Library Research Data Services, University of. “Data Management Plan Templates | University of Virginia Library Research Data Services + Sciences.” Accessed October 27, 2022. <https://data.library.virginia.edu/data-management-plan-templates/>.
- “Welcome to the Data Documentation Initiative | Data Documentation Initiative.” Accessed October 21, 2022. <https://ddialliance.org/>.
- Wickham, Hadley. “Tidy Data.” *Journal of Statistical Software* 59 (September 12, 2014): 1–23. <https://doi.org/10.18637/jss.v059.i10>.
- 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>.