**Data Management for Psychological Science:**

**A Crowdsourced Syllabus**

**Note**: This open-source syllabus was collaboratively created during the Data Management Hackathon at SIPS 2021 (held virtually). A comprehensive list of contributors can be found at the end of this document.

**Using the Data Management Syllabus**

Data management - including data preparation, cleaning, storage, and sharing - is critical to psychological research. Despite its importance, data management is rarely formally taught to students. This syllabus provides detailed descriptions of data management topics, resources, and activities that can be used to create a course or workshop on data management. The syllabus is formatted as a series of modules that motivate the importance of high-quality data management and provide information on best practices at various stages - (1) What is data management and why should we care about it? , (2) Data setup and collection, (3) Data storage, (4) Data cleaning and analysis, (5) Data sharing, (6) Locating existing data, and (7) Writing a data management plan. Each module raises key questions and common errors, as well as resources and suggested assignments to help identify and circumvent mistakes and vulnerabilities. The syllabus is extremely comprehensive and should be tailored for individual use, including putting more emphasis on data management practices specific to one’s subfield. There are many different types of resources, including journal articles, blog posts, podcasts, slide decks, etc. and the syllabus can be adapted for graduate seminars, advanced undergraduate courses, or individual study. Because this syllabus links to work from many researchers, please be sure to give appropriate credit to content creators.



| **Module 1: What is Data Management and Why Should We Care About it?** |
| --- |
| **Module Objectives**This module highlights the ways in which data management is an integral aspect of the research process. We make assumptions about data management more explicit, foreshadow good data management practices, and describe how common data management errors can occur. |
| **Key Questions*** What data-related problems have we experienced or heard about from others?
* What are the personal and professional benefits of setting up a workflow that reduces human error and that is usable with collaborators?
* How efficient are the current data management practices that we engage in?
* When were data management practices last evaluated and updated in the spaces where we work? Do we have a process for internally stress testing these practices?
* How can we ensure the quality of the data as we’re collecting it?
* What metadata are useful and/or necessary to describe our data?
* How effectively are data being stored? Are we meeting regulations/recommendations (e.g., ethical or grant-specific considerations)?
 |
| **Common Errors*** Waiting until data are collected to consider how to manage them
* Relying on a system that makes unrealistic assumptions (e.g., humans will not make any errors)
* Having a “way of doing things”, without a standard or documented plan (e.g., written protocols, checklists) integrated into the workflow
* Using data management practices that depend on one person and, therefore, can easily break down over time or with personnel changes
* Unclear data management roles/responsibilities within a research team and insufficient training of data management practices for all team members
 |
| **Readings*** Aczel et al. (2020). [The role of human fallibility in psychological research: A survey of mistakes in data management.](https://psyarxiv.com/xcykz/)
* Strand (2021). [Error Tight: Exercises for Lab Groups to Prevent Research Mistakes.](https://psyarxiv.com/rsn5y/)
 |
| **Additional Resources*** [Short stories](https://notebooks.dataone.org/data-stories/) about data management
* Briney (2021). [Thoughts on data management as housekeeping](http://dataabinitio.com/?p=1022).
* Retraction Watch: [Bad news for a study about bad news](https://retractionwatch.com/2021/01/25/i-dont-think-i-slept-for-a-day-and-a-half-bad-news-for-study-about-bad-news/)
 |
| **Suggested Assignments*** Complete (or create) a data management checklist for a research project. For examples, see the [Data Management Checklist](https://laneguides.stanford.edu/DataManagement/checklist) (Borghi, 2011) and [The Support Your Data Research Data Management](https://riojournal.com/articles.php?id=26439) rubric (Borghi et al., 2018; Table 1).
* Assess and share potential vulnerabilities in current data management practices
 |

| **Module 2: Data Setup and Collection** |
| --- |
| **Module Objectives**This module outlines the data setup and collection process. First, we discuss various ways to collect data, including media (audio/video files), paper, digital surveys (Qualtrics, QuestionPro, etc.), psychophysiology data (HRV, EMG, HR, EEG, fMRI). Then, we outline the best practices for data collection using these diverse methods. In particular, we discuss the following topics:1. Understand how to document design-to-collection pipeline
	1. Creating a plan for data storage before collection
	2. Establishing folder and project structure
	3. Eliminating data redundancy
2. Learn how to set up collection process for easier analysis
	1. Setting up metadata and creating a codebook/data dictionary (e.g., what are the variable names and labels, how will missing values be labeled)
	2. Assigning identifiers to connect across measures
	3. Transferring data from paper to digital (e.g., data entry; double entry & comparison)
	4. Converting data format (there’s the possibility of data being modified when changing data from one program to another)
3. Learn how to quality check during collection
	1. Checking data collection instruments (paper & digital collection instruments)
	2. Establishing a protocol for documenting potential outliers as they occur
	3. Establishing global principles (e.g., communication tree, don’t delete anything, document everything)
	4. Piloting data collection process
4. Consider ethical issues
	1. Obtaining consent for data sharing
	2. Keeping personal data separate from other data (e.g., informed consent documents)
	3. Ensuring data collection platform follows the rules set by consent forms
 |
| **Key Questions*** What kind of data will we collect?
* Is there already a data collection framework for the specific type of data (e.g., BIDS for MRI/fMRI/EEG/MEG, OSF for behavioural research)?
* Does our department or lab have rules or guidelines on data collection?
* Does our data collection process follow the plans outlined in our grant and/or IRB protocol?
* How will data be stored immediately after they are collected?
* Is there a coherent system for naming variables, files, and folders (e.g., snake\_case, camelCase, UpperCamelCase)?
* How will data files connect to one another, if at all?
* If data come from multiple sources or are in multiple formats (e.g., .csv, .sav), how will we combine them into one analyzable dataset?
* How will metadata be set up to be easily used for reference and analysis in the future?
* What will the process be for quality checking our data collection instruments?
* What will the process be for quality checking data that is entered by humans?
* Who can/will/should have access to the data?
* Data anonymization: Is it necessary for the type of data collected? What should be anonymized? Will the source data be already anonymized or will an anonymized version be created? Who will have access to non-anonymized data and deanonymization information (e.g., a csv with real names and anonymous codes)?
	+ If working with a specific population (e.g. children, autistic populations), will we create and use synthetic data (Quintana, 2020)?
 |
| **Common Errors*** Not piloting a study prior to launch
* Not checking Qualtrics (or other online survey platforms) for anonymized responses, correct metadata, and survey set up (e.g., skip & display logic) before sending them out
* Not documenting what variables are supposed to capture (e.g., just calling things mean\_1x)
* Not having consistent names for the same variable across datasets
* Not running our sample analyses with trial data before activating the study
* Missing the collection of some data
* Not specifying the format participants should enter data as
* Changing participant IDs
* Not checking if data are corrupted, complete, and high-quality during the collection process
* Not making sure different data flows can be combined
* Storing data in formats that cannot be translated across forms (e.g., storing information via text or cell color in excel)
* Not automating data backups (e.g., using zapier.com)
 |
| **Readings*** Barchard et al. (2020). [Comparing the accuracy and speed of four data-checking methods.](https://link.springer.com/article/10.3758/s13428-019-01207-3)
* Reynolds and Schatschneider (2020). [The basics of data management.](https://figshare.com/articles/preprint/The_Basics_of_Data_Management/13215350/1)
* Gorgolewski et al. (2016) [The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments](https://www.nature.com/articles/sdata201644)
* Pernet et al. (2019). [EEG-BIDS, an extension to the brain imaging data structure for electroencephalography.](https://www.nature.com/articles/s41597-019-0104-8)
 |
| **Additional Resources*** [Project structure & naming conventions](https://slides.djnavarro.net/project-structure/#1) (Slidedeck created by Danielle Navarro)
* [Within & Between podcast](http://www.withinandbetweenpod.com/) (Hosted by Jessica Logan & Sara Hart) - Season 2 episodes 13and 14
* [Introduction to Data Management Principles](https://digitalrepository.chop.edu/researchdatamanagement/1/) (Video created by Joy Payton, 2020)
* [Example codebook](https://psidonline.isr.umich.edu/CDS/td14media_codebook.pdf)(2018)
* [How to Make a Data Dictionary](https://help.osf.io/hc/en-us/articles/360019739054-How-to-Make-a-Data-Dictionary) (OSF Support)
* [Writing a Readme File](https://data.research.cornell.edu/sites/default/files/SciMD_ReadMe_Guidelines_v4_1_0.pdf) (Guidelines created by Wendy Kozlowsk, 2014)
* [R package for generating codebooks](https://rubenarslan.github.io/codebook/articles/codebook_tutorial.html) (Tutorial by Ruben Arslan, 2021)
* For qualitative data, [The ROCK Book: The Reproducible Open Coding Kit](https://sci-ops.gitlab.io/rockbook/) (Zörgő & Peters, 2021)
* For neuroimaging data, [Brain Imaging Data Structure](https://bids.neuroimaging.io/), [*pybids*](https://bids-standard.github.io/pybids/introduction.html), and [psych-DS](https://github.com/psych-ds/psych-DS)
 |
| **Suggested Assignments*** Generate on-paper questionnaires that have different kinds of responses that a participant might make (e.g., some missing data, decimals where there should be whole numbers, circling multiple points on a Likert scale). Then, give these questionnaires to students and have them each enter these data into a spreadsheet. Did they all generate the same spreadsheet or did they make different subjective decisions?
* Create a codebook or variable name book for a set of measures
* Develop a naming scheme for data that have messy variable names
* Write a data collection protocol (e.g., participant eligibility protocol, outlier exclusion protocol)
* File a preregistration with a focus on the data collection and setup
* Write code to automate a step of the data management collection phase (e.g., folder naming)
* Give students a small amount of time (e.g., 15 minutes) to search for important information (e.g., p-values, references) within a messy folder structure
 |

| **Module 3: Data Storage** |
| --- |
| **Module Objectives**This module presents information about thorough and effective data storage. We discuss the pros and cons (e.g., analog vs. digital, storage capacity) of various storage options (e.g., hard drives; cloud services; server database; no-SQL database solutions) and how these relate to project structure. We discuss the basics of documentation to make sure the data will be usable (e.g., storing codebooks, storing code for transforming raw data to analyzable) and how to make a backup plan to prevent data loss (e.g., creating zips for backups). We also emphasize the importance of ethics and security to make sure the stored data are secure access permissions to different types of data (e.g., identifiable vs. de-identified) are properly managed. |
| **Key Questions*** Do we have a backup solution? If so, where are backups stored?
* Do we use file versions to prevent corrupted files from corrupting backups?
* What is the appropriate way to dispose of old backups, if at all?
* In a team, how can we keep track of who has which version of the data? How can we ensure that different versions of the data are not circulating so people won’t get confused about which one to use and where to get it?
* What privacy obligations do we have?
* How can we minimize the risk of leaking potentially sensitive information?
* How open do we want our data to be, and what will be open at what stage? Note that there are often different storage solutions for private/public datasets.
* How can we conduct proper anonymization?
* What resources (storage solutions) are available at our organisation?
* How can we safely and securely encrypt data?
* Where to save/store the data so that colleagues can access them/work on them simultaneously?
* Raw data vs. processed data (e.g., MRI data): which to save/backup considering limited storage locally or online?
* Do we have a process for pulling data regularly off non-standard or non-networked devices (e.g., data collected on an iPad)?
 |
| **Common Errors*** Data are undocumented or poorly documented, including uninformative variable names (which can result in accidentally using wrong variables), forgotten metadata (e.g., when, where, and in which language were data collected?)
* Standard operating procedures and protocols are written for a single person, so nobody else can decipher them
* Poor security on data storage
* Incorrect or unsafe storage of personally identifiable data
* Sensitive data are downloaded onto a shared, unsecure computer
* Original data are destroyed without backups
* Not knowing which version to use and being afraid to lose important data
* Insufficient number or location of backups (rule of 3; at least one geographically removed)
 |
| **Readings*** Rouder (2016). [The what, why, and how of born-open data.](https://link.springer.com/article/10.3758/s13428-015-0630-z#citeas)
 |
| **Additional Resources*** [Research data management toolkit](https://bond.libguides.com/research-data-management/research-data-management-toolkit/storage-backup) (Bond University Library, 2021)
* [What is data storage? Data storage types & attributes](https://www.cdw.com/content/cdw/en/articles/datacenter/2019/03/18/what-is-data-storage.html) (2019)
* [Guidance on research and data protection](https://www.ed.ac.uk/data-protection/data-protection-guidance/specialised-guidance/research-data-protection) (The University of Edinburgh, 2021)
* [Blog posts on encrypting data](https://axcrypt.net/blog) (AxCrypt)
* [Teaching Integrity in Empirical Research](https://www.projecttier.org/) (Project TIER)
* [External management of sensitive data](https://ukdataservice.ac.uk/deposit-data.aspx) (UK Data Service)
 |
| **Suggested Assignments*** Evaluate documentation of open data (e.g., from recent *Psychological Science* articles)
* Assess whether data contain identifiable or personal data (see GDPR - EU data protection laws - differentiation)
* Practice data storage with the [MySQL tutorial](https://www.w3schools.com/MySQL/default.asp)
 |

| **Module 4: Data Cleaning and Analysis** |
| --- |
| **Module Objectives**This module deals with data cleaning and analysis - two integral parts of the research process. Data cleaning is considered everything that transforms the data from the raw format into a format that lets us test our hypotheses. The motivations for the data cleaning steps should be transparent and, ideally, the criteria for data cleaning should be specified *before* collecting the data. Data analysis focuses on testing our hypotheses or conducting exploratory analysis on our data. Here, ensuring that analysis results are correctly presented in the final report is important. In both stages, the data coding pipeline should be reproducible and the motivations for each step transparent. |
| **Key Questions*** Are all decisions documented and commented properly for individuals and publications (e.g., which participants were excluded, were variables recoded)?
* Are there items that need to be recoded (e.g., reverse-coded) and will they be recoded as existing or new variables?
* How can we ensure that we can track the different versions of the analysis script (syntax) and datafiles?
* Is the code (syntax) reproducible?
* If we disappear tomorrow, is there enough information documented so that somebody else could continue the analysis?
* What software (package version) was used to obtain the results?
* Who is doing the data cleaning? How will team members discuss cleaning issues and come to consensus?
* What software (and packages) will be used to clean and analyze the data? We should be sure to document what (and which versions) will be used (Docker; for R - checkpoint, Packrat, Groundhog; for python - Conda environment)
* What are our aims and how do the data (and individual variables) have to be formatted to answer these aims? Are the variables in the appropriate format?
* Do the data look plausible? Plot the data at intermediate time points to check.
* Did something unexpected or implausible happen during data collection or import? Consider whether variables that should be numbers are really numbers, whether the range of data make sense (e.g., age should be a positive number, age > 100 is quite rare), whether there are outliers including multivariate outliers (e.g., there might be 12 year olds in our data and people who drink alcohol 5 days a week, but 12 year olds drinking alcohol 5 days a week would be weird).
* How will we handle missing data? Are missing variable codes correctly defined and read in by the analysis software?
	+ Also, are the missing data: missing completely at random, missing at random and missing not at random? What method will be used to handle missing data (e.g. multiple imputation)?
* For which variables will we have to compute aggregates (means, sum scores or similar)?
* Are there participants that should be dropped due to data quality issues/exclusion issues (consider dummy variables to indicate whether participants should be included or excluded from analysis, this way we retain the raw data)?
* Is the dataset in the right shape or format for analysis (wide vs. long)?
* What code was used to achieve the results reported in the paper? How can we ensure that our paper is reproducible?
* In what order will we execute steps (run analysis scripts)? Do we have a master script that executes all files in the right order? Should we consider workflow management systems (e.g., targets, Snakemake)?
 |
| **Common Errors*** Not keeping the raw data file
* Not specifying and transparently recording data processing (cleaning and analysis) steps
* Only making data processing considerations/decisions after the data is collected
* Opening and editing the data in different programs, potentially changing formats and variables
* Not double checking variable values, especially if we are using Qualtrics or other online survey platforms (e.g., we think the range of our variable is 0 to 10 but Qualtrics coded it as 1 to 11)
* Manually editing data
* Not using clear naming conventions
* Not knowing which files have been updated and which have not
* Not looking at our data/checking whether after executing step X, they look like they should look
* Not documenting which software version was used to create the results
* Always declaring a variable locally by hand instead of once centrally (against the Don’t Repeat Yourself - DRY - principle)
* Not having the same results after running the same data cleaning
* Not having good version control to roll back if an error occurs or accidentally overwriting variables, objects, and/or files (Git could help)
* Unintentional coding or programmatic errors (factor → numeric in R)
* Not setting (or documenting) a seed when random values are sampled
* Overwriting built-in functions by choosing those as a variable name
 |
| **Readings*** Wickham and Grolemund (2016). [R for data science: import, tidy, transform, visualize, and model data.](https://r4ds.had.co.nz/)
* Peikert and Brandmaier (2019). [A reproducible data analysis workflow with R Markdown, Git, Make, and Docker.](https://psyarxiv.com/8xzqy/)
* Gentzkow and Shapiro (2014). [Code and data for the social sciences: A practitioner’s guide.](https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf)
* Eglen et al. (2017). [Toward standard practices for sharing computer code and programs in neuroscience.](https://www.nature.com/articles/nn.4550)
* van Vliet (2020). [Seven quick tips for analysis scripts in neuroimaging.](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007358)
* Blischak et al. (2016). [A quick introduction to version control with Git and GitHub.](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004668)
* Boettiger and Eddelbuettel (2017). [An introduction to rocker: Docker containers for R.](https://journal.r-project.org/archive/2017/RJ-2017-065/index.html)
* Müller and Freytag (2005). [Problems, methods, and challenges in comprehensive data cleansing.](http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/techreports/2003-hub_ib_164-mueller.pdf)
* Rahm and Do (2000). [Data cleaning: Problems and current approaches.](https://web.archive.org/web/20170809153257/http%3A//lips.informatik.uni-leipzig.de/files/2000-45.pdf)
* Wilson and Collins (2019). [Ten simple rules for the computational modeling of behavioral data.](https://elifesciences.org/articles/49547)
* Haines et al. (2020). [Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox.](https://psyarxiv.com/xr7y3/)
 |
| **Additional Resources*** [Quantitude podcast](https://quantitudepod.org/) (Hosted by Patrick Curran & Gregory Hancock)
* [Stack overflow](https://stackoverflow.com/)
* [The ultimate guide to data cleaning](https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4) (Blog post by Omar Elgabry, 2019)
* [From chaos to order](https://openscience-nijmegen.nl/assets/slides/2021_05_26_From_Chaos_to_Order_Efficient_file_management.pdf) (Slidedeck by Johannes Algermissen, Hannah Peetz, & Eva Poort, 2021)
* [Naming things](https://speakerdeck.com/jennybc/how-to-name-files) (Slidedeck by Jenny Bryan, 2015)
* [File naming conventions](https://datamanagement.hms.harvard.edu/collect/file-naming-conventions) (Harvard University)
* [The tidyverse style guide](https://style.tidyverse.org/) (Hadley Wickham)
* [Matlab programing style guidelines](http://www.datatool.com/prod02.htm) (Datatool)
* [Code guidelines](https://www.fieldtriptoolbox.org/development/guideline/code/) (Field Trip)
* [Learning Git without tears](https://www.youtube.com/watch?v=pTJWeuQB1Yc&ab_channel=SIOSInitiative) (Video by Julia Haaf)
* [Happy Git and GitHub for the useR](https://happygitwithr.com/index.html) (Created by Jenny Bryan & Jim Hester)
* [Git - fast version control](https://git-scm.com/docs/gittutorial)
 |
| **Suggested Assignments*** Follow a list of instructions to recode and edit a raw dataset. Discuss with peers whether everyone ended up with the same dataset. If not, what was done differently?
* Swap data cleaning scripts with another student. How well can students follow and reproduce their data cleaning steps? What difficulties did they encounter? How can we avoid those problems?
* Create a list of problems in a “bad” data set and consider how to avoid them in the future.
* Write an analysis plan and well-commented code documenting the steps we will take when working with a dataset. Write the final report of our research project as a Jupyter notebook/RMarkdown script.
* Do a [Red Team](http://www.the100.ci/2020/06/29/red-team-part-1/) exercise: Try to find mistakes in each others’ data, code, and manuscripts.
 |

| **Module 5: Data Sharing** |
| --- |
| **Module Objectives**In this module, we will think critically about who our data can and should be shared with. The aim of this module is to appreciate how the format of data and associated documentation influence their usability, and to develop an awareness of how to safely share data. We will also begin to develop an ability to make principled decisions about who, what, where, when, why, and how of our data sharing plan. |
| **Key Questions*** Before data collection: What should be included in ethics documents (e.g., consent forms, protocols) to allow for the desired data sharing?
* What are we allowed to share?
	+ Are these data coming from a protected source (i.e., electronic health record data)?
	+ Are there relevant data use agreements to consider?
	+ Are there risks of identifying participants (e.g., data on IP addresses, prolific IDs)?
		- How will people with specific conditions be anonymised?
		- If data are indirectly identifiable, what strategies will we use to appropriately de-identify data?
		- How will additional data sharing affect risk of re-identification (e.g., sharing class summary data from students with teachers)?
	+ Do we have the necessary ethics approval to share?
	+ If unable to directly share the data, could we share synthetic datasets/restricted access models?
* Which information should be shared?
	+ Do we want to provide coded or uncoded data (e.g., only raw data but not accuracy derived from it)?
* When will data be shared (e.g., upon collection, publication, completion of the entire project)?
	+ Will there be an embargo period for the data?
* What format will the data be stored in (e.g, csv, json, BIDS)?
	+ How large is the raw data file?
	+ Are the data accessible in a spreadsheet or a relational database?
	+ Is the format proprietary or open source?
	+ Is the format human-readable?
* Where are the data stored? (open or restricted curated repository, on a personal website)
	+ If in a data repository, how should we choose one?
	+ If access is restricted, what will the data sharing agreement be?
	+ How long will the data be accessible?
* How will licenses affect data sharing?
	+ Who really “owns” the data (i.e., is there a custodian)?
	+ Which license will we apply?
	+ If the funder and university contradict each other on how to share data, who owns the data?
* Who is the data for? Who might want to use/see it?
	+ How will people find our data?
	+ How can we make our data easy to find?
	+ How can our data sharing be comprehensive enough to accommodate unexpected uses of our data as well as facilitate common ways to explore the data in our field?
 |
| **Common Errors*** Sharing data with sensitive or potentially identifiable combination of information
* Providing only cleaned/processed data (limiting the ability of others to use the data)
* Not including necessary documentation for the data to be interpretable (e.g., poor/no metadata or codebooks)
* Forgetting to un-embargo our dataset (or project)
* Sharing data in a proprietary format that is difficult to access (e.g., .sav, .sas, .sps)
* Not providing information on how to contact the data curator
 |
| **Readings*** Alter and Gonzalez (2018). [Responsible practices for data sharing.](https://psycnet.apa.org/doiLanding?doi=10.1037%2Famp0000258)
* Buchanan et al. (2021). [Getting started creating data dictionaries: How to create a shareable data set.](https://journals.sagepub.com/doi/full/10.1177/2515245920928007)
* David et al. (2020). [FAIRness literacy: the Achilles’ heel of applying FAIR principles.](https://datascience.codata.org/articles/10.5334/dsj-2020-032/)
* Korkiakangas et al. (2014). [Challenges in archiving and sharing video data: Considering moral, pragmatic, and substantial arguments.](https://discovery.ucl.ac.uk/id/eprint/10019201/1/Challenges_in_Video_Data_PRE_PRINT.pdf)
* Logan et al. (2021). [Data sharing in education science.](https://journals.sagepub.com/doi/10.1177/23328584211006475)
* Mannheimer et al. (2019). [Qualitative data sharing: Data repositories and academic libraries as key partners in addressing challenges.](https://journals.sagepub.com/doi/10.1177/0002764218784991)
* Meyer (2018). [Practical tips for ethical data sharing.](https://journals.sagepub.com/doi/10.1177/2515245917747656)
* Tsai et al. (2016). [Promises and pitfalls of data sharing in qualitative research.](https://linkinghub.elsevier.com/retrieve/pii/S0277953616304269)
 |
| **Additional Resources*** [Open Research Resources Browser](https://ukrn-orr.netlify.app/) (UK Reproducibility Network)
* [Changing the Culture of Data Management and Sharing: A Workshop](https://www.youtube.com/playlist?list=PLGTMA6QkejfgsxahcRn0HwFuF8m8lOnOQ) (YouTube playlist with 39,8-60 minutes, videos by the National Academies of Sciences, Engineering and Medicine Health and Medicine Division)
* [Data de-identification procedures for data sharing workshop](https://osf.io/em3da/) (Center for Open Science)
* [Anonymization: Managing data protection risk code of practice](https://ico.org.uk/media/1061/anonymisation-code.pdf) (UK Information Commissioner’s Office)
* [Ultimate Consent Form](https://open-brain-consent.readthedocs.io/en/stable/ultimate.html) (Open Brain Consent working group)
 |
| **Suggested Assignments*** Consult with a librarian (e.g., ask what resources they can connect us with).
* Write a data sharing plan. Include any information about requirements for ethics documents, what data will be shared, where, with whom, format of data, and contact info as well as what needs to be included in a data processing record and/or coding scheme.
 |

| **Module 6: Locating Existing Data** |
| --- |
| **Module objectives**Collecting primary data can be challenging and resource intensive. As a result, secondary analysis of existing data has become an efficient means to address knowledge gaps and to make use of data already collected. There are an ever increasing number of readily available datasets that can be accessed and used in this regard. However, finding and making good use of such datasets can be challenging. This module will discuss where to begin, what to do once we find the data, and what sorts of practices/problems we may encounter when working with existing data. |
| **Key Questions*** What data do we need and why? In particular, what specific variable(s) are we interested in? Do we need data specific to a certain geography, population, and/or time period? Do we need a certain number of assessments (i.e., longitudinal data)?
* Where might we find data to answer our research questions?
* What state are the data in that we are interested in accessing (i.e., raw, cleaned, etc.)?
* Has the research question already been answered using this dataset?
* If the data are longitudinal, have the constructs (and how they are measured) changed over time, limiting their comparability?
* Can we access the data online or will we need to go into a facility?
* Is there a data dictionary available to guide us through using the dataset?
* How will we request and navigate any challenges when the data are “Available upon request”?
* Do we need ethics approval to use the data for our purposes?
 |
| **Common Errors*** Collecting data that already exist
* Not reading data dictionaries, readme files, or associated documentation prior to using the data
* Not reaching out to the data management team if we have questions
* Not understanding how the data were collected
* Not checking for updates to the dataset
* Not checking whether the scale used was abbreviated, shortened, and/or modified
 |
| **Readings*** Gregory et al. (2018). [Eleven quick tips for finding research data.](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006038)
* Cheng and Phillips (2014). [Secondary analysis of existing data: opportunities and implementation.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4311114/)
* Gregory et al. (2019). [Lost or found? Discovering data needed for research.](https://arxiv.org/ftp/arxiv/papers/1909/1909.00464.pdf)
 |
| **Additional Resources*** [Data Sharing and Management Snafu in 3 Short Acts](https://www.youtube.com/watch?v=66oNv_DJuPc) (NYU Health Sciences Library)
* [Research Data Management: Six ways to discover existing data](https://www.youtube.com/watch?v=AZMUKgM8X-A) (Utrecht University)
* Where to look for existing data:
	+ Projects shared on the [Open Science Framework (OSF](https://osf.io/dashboard))
	+ [Tidy Tuesday](https://github.com/rfordatascience/tidytuesday)
	+ Government data
	+ [Inter-university Consortium for Political and Social Research (ICPSR](https://www.icpsr.umich.edu/web/pages/ICPSR/index.html))
	+ [UK Data Service](https://www.ukdataservice.ac.uk/)
	+ [National Institutes of Health (NIH](https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html))
	+ [United Nations Educational, Scientific, and Cultural Organization (UNESCO) Institute of Statistics](http://uis.unesco.org/)
	+ [World Bank Data](https://data.worldbank.org/)
	+ Data repositories from university libraries
	+ [The Dataverse Project](https://dataverse.org/)
	+ [Google Dataset Search](https://datasetsearch.research.google.com/)
	+ [Kaggle Datasets](https://www.kaggle.com/datasets)
	+ [Zenodo](https://zenodo.org/)
	+ [Figshare](https://figshare.com/)
 |
| **Suggested Assignments*** Identify a potential dataset to answer one of our research questions and write a report identifying the variables of interest, any specific considerations that need to be addressed when using the dataset (i.e., sample weights, items/constructs that change over time), and limitations related to our use of the dataset.
* Locate a dataset used in a previous publication (e.g., from recent *Psychological Science* articles) and assess documentation of the data. Can we understand the data without input from the research team? What is well-documented and what is confusing?
 |

| **Module 7: Writing a data management plan** |
| --- |
| **Module objectives**Many of the topics covered in this module we have encountered in previous modules. Writing a data management plan is where we will bring all the concepts we have learned thus far together to demonstrate and document that we have a plan for our data and have considered potential risks associated with collecting and managing data and how to mitigate them. Indeed, most ethics review boards, and funding agencies require data management plans; however a template that applies to *all* research situations does not exist because different research projects have different data management needs. Fortunately, there are components that apply to *most* data management plans. Thus, the goal of this module is to address key questions common to most data management plans to generate a basic data management plan that can be adapted to fit our specific needs in the future.  |
| **Key Questions*** Are we collecting primary data or analyzing previously collected data? If we are collecting new data, why do we need to collect it (e.g., existing data do not answer our research question)?
* What do we plan to do with the data in the short term and long term? For example, we may plan to write manuscripts with the data in the short term and to share the data publicly in the long term).
* How will we store and keep the data secure (including against unexpected data loss or data breach)?
* How will data be shared beyond the research team? Do we need to obtain consent from research participants to do so?
* What could go wrong during the life cycle of our data? How will we guard against unfavorable events happening?
* What are the ethical considerations at each step of the data life cycle?
* Have we consulted with a librarian or our ethics specialist about our data management plan?
* Will our data management plan be available for others to review?
* How will we transparently document any updates or changes made to the data management plan?
 |
| **Topics That Will be Addressed in Most Data Management Plans**These topics are adapted from the [USGS Data Management Plan Checklist](https://prd-wret.s3-us-west-2.amazonaws.com/assets/palladium/production/atoms/files/data-management-checklist_508-compliant.pdf): 1. **Plan:**
	1. Provide basic information about project (e.g., title, purpose, review of research question(s), names and contact information of people involved).
	2. Information about the time frame of the project (e.g., start and end date of data collection).
	3. Identify plans for data sharing, and personnel responsible for sharing data, including if there are any limitations for data sharing.
2. **Acquire:**
	1. If the data already exist, when and how were they (or will they be) obtained?
	2. If the data do not yet exist, how will they be collected?
	3. What various formats will the raw data be in?
3. **Process and analysis:**
	1. Indicate if the data have been manipulated or processed in any way and if so, describe the technologies used to process or transform the data (e.g., software, models).
4. **Preserve (data backup):**
	1. Indicate how the data will be stored including information about where, how often, and whose responsibility it is to establish and maintain the data backup.
	2. Describe the various formats and types of data that will be preserved.
5. **Metadata**:
	1. Indicate what metadata and documentation will accompany the datasets.
	2. Identify who is responsible for creating and maintaining metadata files and any accompanying documentation.
6. **Publish and share:**
	1. Describe the anticipated format for publication (e.g., data release, manuscript publication).
	2. Describe the plan to maintain or update shared data (if applicable).
	3. Describe any data sharing restrictions or limitations.
 |
| **Common Errors*** Not creating a data management plan prior to data collection
* Not sharing the data management plan with the broader research team
* Not updating the plan as changes occur or as new members of the research team join/leave the team
* Data are not organized in an accessible way or structured in line with common guidelines/standards
* Considering data management plans as final and not allowing for flexibility
 |
| **Readings*** Klein et al. (2020).[A practical guide to transparency in psychological science](https://psych-transparency-guide.uni-koeln.de/)
* Dabrowski (2018). [Productivity through data management](https://oaktrust.library.tamu.edu/bitstream/handle/1969.1/166284/productivity-through-data-management-slides.pdf?sequence=1)
 |
| **Additional Resources*** [Writing a Data Management Plan](https://www.ucl.ac.uk/library/research-support/research-data-management/policies/writing-data-management-plan) (UCL)
* [Data Management and Data Management Plans](https://figshare.com/articles/presentation/Data_Management_and_Data_Management_Plans/7890827) (presentation by Jessica Logan, 2019)
* [Write a Data Management Plan](https://libraries.mit.edu/data-management/plan/write/) (tips on writing data management plans, MIT)
* Example data management plans:
	+ [Political Protest and Generational Change in New Democracies](https://osf.io/zv4e9/wiki/home/) (Joly, 2017)
	+ [Digital Humanities and Secondary Data](https://zenodo.org/record/4019309#.YPiVRi2ZM19) (Gray & Cooper, 2020)
	+ [Pathways to Research Data Management](https://osf.io/u3rqg/) (University of Victoria)
 |
| **Suggested Assignments*** As a final, cumulative project, create a data management plan using the skills learned throughout the course
* Find a publicly posted data management plan and identify strengths and weaknesses
* Design a data management template for a specific design
 |

**SIPS 2021 Hackathon Facilitated By:** Michaela DeBolt, Arianne Herrera-Bennett, Katherine M. Lawson, Sarah Schiavone, Anna Wysocki

**Contributors (listed alphabetically by last name)**: Johannes Algermissen, Sharon Armstrong, Nick Augustat, Annouk Bakker, Hilmar Brohmer, Denver Brown, Debora I. Burin, Sau-Chin Chen, Clare Conroy-Murray, Li-Ching Chuang, Michaela DeBolt, Lisa DeBruine, Mahmoud Elsherif, Fabienne Ennigkeit, Rose Franzen, Daria Gerasimova, Amélie Gourdon-Kanhukamwe, James Green, Samuel Guay, Abbey Hammell, Arianne Herrera-Bennett, Lea Hildebrandt, Karoline Huth, Hanz IJzerman, Matt Jaquiery, Evgenii Kalenkovich, Julian Karch, Matthew Kim, Laura-Isabelle Klatt, Kai Krautter, Anton Kunnari, Katherine M. Lawson, Holly Levin-Aspenson, Yih-Shiuan Lin, McKenzie Lockett, Jessica Logan, Jamal Mansour, Effie Marathia, Erin Mazerolle, Sarah McIntyre, Johanne Nedergaard, Kelly Nguyen, Jay Patel, Julia Pauquet, Mahjabeen Raza, Kim Reece, Sarah Schiavone, Lea Schumacher, Shivendra Singh, Julia Strand, Anna Szabelska, Tamer Said, Alex Uzdavines, Anna Wysocki