LAS 6292: Data Collection and Management

Emilio M. Bruna

November 15, 2026

Table of Contents

List of Figures

List of Tables

Introduction

This book is a manual for students in LAS 6292: Data Collection & Management. The course is designed for graduate students from any discipline – social sciences, humanities, biophysical sciences – and at all stages of their graduate program. It is an introduction to methods for collecting, organizing, managing, and visualizing both qualitative and quantitative data. Students will gain hands-on experience with best practices and tools.

Describe the different types of research data & the research data life-cycle

Explain the need for and benefits of data management and sharing

Describe and implement best practices for data collection, storage, management, and sharing

Find, download, and analyze publicly available data from repositories

Carry out simple and reproducible data corrections and dataset organization

Describe public policies and agency requirements for data management and sharing

Articulate the major legal & ethical issues regarding the data collection, use, and storage

Create and Implement Data Management Plans in funder-specific formats

Identify and properly use tools for more efficient and secure data collection in the field

---

Citation

This guide is a Quarto Book hosted on the Bruna Lab's Github site. Please suggest edits to the text, missing resources, or make suggestions for improvement either by pull request or by posting an issue on the course book's repository. For more information and tutorials for doing so, see **?@sec-manual**.

Part I.

Course Topics

1. Naming Conventions

1.1. Introduction

The first (and easiest) way to get started organizing your data is to use a simple, clear, and consistent system for naming your files (i.e., a "naming convention"). Using a well-thought out naming convention has a number of important benefits:

- Files will be easier to find
- You won't have to open files to see what they are
- Files are easier to sort
- Files are easier to share with collaborators (and for collaborators to use)
- It helps prevent accidentally overwriting or deleting files

1.2. How to name your files

Your file names should:

1. **Tell you about the contents of the file and allow you to uniquely identify it.** For instance (compare `'data 1998'` vs `'survey data'` vs `'survey data 1998'` vs `'small mammal survey data 1998'`). You can use things like the an acronym for the project, abbreviation for the study, location where collected, the name of the investigator, the year or month when collected, the type of data type, etc.
2. **Be as short as possible.** Many database systems (like MS One Drive) have limits on file name length. Use a max of 25 characters, and aim for less than that.
3. **Avoid using special characters and spaces** Don't use characters like `$ % ^ & # | :`. This is really important because software, automated data processing tools, web browsers, and computer operating systems use spaces and special characters for dividing up (parsing) strings of text, as well as other processes.
4. **Start with letters, never numbers.** Using numbers at the start of file names can lead to problems when they are read into stats software and other programs. Instead of `'2020_survey_responses'` use `'survey_responses_2020'`
5. **Help you quickly sort files chronologically or numerically**. Dates are very useful in file names because they can help identify the most recent files easily. You can *not* rely on the file date associated with the file on the computer! However, it is really important to use a clear format for dates, especially if collaborating internationally. I learned the hard way that not everyone reads `'census_data_3-4-2018'` the same way I do, so I started using `'census_data_4march2018'`. However, even this isn't ideal because of

how it sorts things in a folder. Now I now use this format: `'file_name_YYYYMMDD'` (`'census_data_20180403'`). You can also use `file_name_YYYY-MM-DD`. This is based on the international format for dates and times (ISO 8601). In addition to the benefit of sorting more easily, you'll never have to use the the dreaded `'_____.final.docx'` and `'_____.actual_final.docx'` in your file names ever again…you'll always know which is the last one. If the files don't require a date, but you want to maintain a sequence, use leading zeros to maintain sort order. `7` and `701` don't sort the way you expect them to, so use `007` and `701`.

6. **Use a consistent method of dealing with spaces and letter case.** . For a the spreadsheet with responses to a survey by 5 different people, don't use `'survey responses Florida'`. Instead use (i) `'survey_responses_florida'` or (ii) `'SurveyResponsesFlorida'`. Try to avoid mixing the two (e.g., `'survey_responses_Florida'`). By the way, (i) is known as `pothole` and (ii) is known as `camel`. Personally, I prefer pothole name is better because of it is easier to read and you never to remember what words get capitalized (because none of the do). Use underscore `( _ )` or dashes `( - )` to separate meaningful parts of file names.

Once you have a system, write it down, print it out, and put it up somewhere you can easily refer to it. Remember -whenever possible make the names simple, informative, and unique. If you collected survey data and behavioral data in 2017, don't call them both `data_2017` even if they are in different folders. Call one `'behavior_obs_2017'` and the other `'survey_responses_2017'`).

Simple names also help you avoid problems with "system and software portability" - you might be working on a Mac but your collaborators are working on Linux or Windows, and the software made for each can read file names in different ways. File names that are simple, lower-case, and have no special characters are less software- and platform-dependent.

1.3. Tools & Resources

1.3.1. Best Practices & Conventions for Naming Files

- Smithsonian Library
- UF Library *(bottom of the page)*
- Stanford Library

1.3.2. Tutorials and Tools for Bulk Renaming of Files

- Batch Renaming of Files in Windows: Post
- Batch Renaming of Files in MacOS: Post 1, Post 2
- Adobe Bridge *(a powerful "asset manager" used to organize collections and edit metadata)*
- Renamer (macOS)
- Renamer Lite
- PSRenamer
- Bulk Rename (Windows)

- Rename Master (Windows)
- Ant Renamer

2. File Organization

2.1. File Organization - Electronic

There aren't really "rules" about how to organize your files, but I think the best way to organize is to think in terms of "projects". Of course, how you define a "project" is sometimes tough…is the PhD the 'Project' or is each chapter of the dissertation a 'Project'? You'll have to give this some thought and think of a system that works for you. That said…

I like 'Project' based organization for two main reasons:

1. **First**, it pushes you to think in terms of outcomes: 'projects' lead to 'products' (e.g., paper, report, grant proposal). This helps you think more clearly about what data need to be collected and the analyses that need to be done. If the data aren't for a new project, and it's not clear to you if or how the data fit into a pre-existing project, then are you sure you need to spend time, effort, and money collecting them?
2. **Second**, project-based organization focuses on the individual steps and components that take a project from start-to-finish: planning your data collection, data gathering and clean-up, analysis, presentation, and data archiving.

2.2. Example: Project Based Organization

A project-based organization scheme for a master's thesis might look something like this:

1. Top Folder: emb_masters_thesis
2. Sub-folder 1: `Chapter1_lit_review`
3. Nested Folders and Subfolders for:
    - Plans for Data Collection & Management
        - DMP
        - plans for storage and backup
    - data
        - raw data
        - clean data
        - metadata
    - analyses
        - intermediate outputs from analyses
        - final outputs from analyses
    - presentation (e.g., `masters_ch1_text` or `ch1_slides`)

- data archiving

You can have one of these for each thesis chapter. Note that some (DMP, plans for storage and backup) might be better placed one folder up in the `emb_master_thesis` folder

But what if I will be using the same data for multiple chapters or papers? This is a great questions, and there are several options. One would be to include both chapters in the same folder, each with its own subfolder (manuscript 1, manuscript 2). The other would be to have each as a separate project, and pull the data in from the online archive where you deposited the data (or folder where the data are stored). As you can probably guess, I think #2 is better because I think of each paper as a different project.

2.3. File Organization - Physical

Whenever possible I try to organize physical records and archives in the same way as digital ones: in a single box, file folder, cabinet, etc.** try to set it up so that you go to one place for everything related to a project. Of course, you can't always do that because specimens go to museums or stay in an archive…but you can keep the information on how to *find* the originals there. Don't forget to include copies of the survey, data sheets, etc., as well as maps or other such resources.

Correspondence is a bit more complicated. You can keep notes - electronic or paper - in the relevant folder. But one thing that has gotten complicated is email correspondence regarding projects, manuscripts, etc. Printing seems like a huge waste of paper, and yet if critical it might be worth doing. Other wise keep an email folder for each project as well - with the same name as the electronic version so you can easily find email correspondence related to each project.

2.4. Documentation

Write it down. Prepare a file documenting your data organization structure so you can share it with team members (or look at it years later when looking for something). These can be `'README.txt'` files as needed for each folder. The `'README.txt'` file might include such things as explanations of naming conventions and how the structure of the directory relates to the structure of the project.

There are some excellent tools for organizing projects. Use appropriate tools, such as version control tools or electronic lab notebooks, to keep track of the history of the data files (more on this later in the semester).

3. Data Storage & Backup

1. **What is the difference between a backup and an archive?**

    - Original Data
    - Operation Data
    - Backup
    - Archive

2. **Why backup?**

    - Hardware failure
    - Software or media obsolescence
    - Virus infection, malware, or malicious hacking
    - Power failure
    - Human error
    - Future validation of results
    - Data requests
    - Legal Obligations: Lawsuits, FOIA reequests, agency mandates
    - Required by funding agency or foundation
    - Server, hardware, or software failure
    - Financial instability of the projecct or organization

4. The Back-up Rule: 3-2-1

3-2-1: A secure backup requires a *minimum* of **3 copies** of your data on **2 types of storage media** with **one copy off-site**. Having 1 copy off-site protects your data from local risks like theft, lab fires, flooding, or natural disasters. Using 2 storage media improves the likelihood that at least one version will be readable in the future should one media type become obsolete or degrade unexpectedly. Having 3 copies helps ensure that your data will exist somewhere without being overly redundant. ***Be sure to migrate data from physical storage every 3-5 years***

5. Storage Media

Successful preservation depends in great part on storage media that are in good physical and operational condition. There are many different kinds of storage media that can be used to save and back-up digital files; each has pros and cons. They include:

- Desktop and Laptop Computers
- Campus servers
- Commercial Cloud Backup (e.g., Dropbox, Google Drive, Amazon, Carbonite, Spideroak). ***Caution: if Google Drive or Dropbox are set up for file-syncing, then they are not a true backup.***
- External hard drives (range from disks to USB drives)
- Optical Storage (CD, DVD)
- Paper (e.g., printouts of spreadsheets and text files)

All storage media, whether hard drives, discs or data tapes, will wear out over time, rendering your data files inaccessible. To ensure ongoing access to both your active data files and your data archives, it is important to continually monitor the condition of your storage media and track its age. Older storage media and media that show signs of wear should be replaced immediately. **Use the following guidelines to ensure the ongoing integrity and accessibility of your data:**

1. **Test Your Storage Media Regularly:** It is important to routinely perform test retrievals or restorations of data you are storing for extended periods on hard drives, discs or tapes. It is recommended that storage media that is used infrequently be tested at least once a year to ensure the data is accessible.
2. **Beware of Early Hardware Failures:** A certain percentage of storage media will fail early due to manufacturing defects. In particular, hard drives, thumb drives and data tapes that have electronic or moving parts can be susceptible to early failure. When putting a new drive or tape into service, it is advisable to maintain a redundant copy of your data for 30 days until the new device "settles in."
3. **Determine the Life of Your Hard Drives:** When purchasing a new drive unit, note the Mean Time Between Failure (MTBF) of the device, which should be listed on its specifications sheet (device specifications are usually packaged with the unit, or available online). The MTBF is expressed in the number of hours on average that a device can be used before it is expected to fail. Use the MTBF to calculate how long the device can be used before it needs to be replaced, and note that date on your calendar (For example, if the MTBF of a new hard drive is 2,500 hours and you anticipate having the unit powered on for 8 hours a day during the work week, the device should last about 2 years before it needs to be replaced).

4. **Routinely Inspect and Replace Data Discs:** Contemporary CD and DVD discs are generally robust storage media that will fail more often from mishandling and improper storage than from deterioration. However lower quality discs can suffer from delamination (separation of the disc layers) or oxidation. It is advisable to inspect discs every year to detect early signs of wear. Immediately copy the data off of discs that appear to be warping or discolored. Data tapes are susceptible both to physical wear and poor environmental storage conditions. In general, it is advisable to move data stored on discs and tapes to new media every 2-5 years (specific estimates on media longevity are available on the web).

5. **Handle and Store Your Media With Care:** All storage media types are susceptible to damage from dust and dirt exposure, temperature extremes, exposure to intense light, water penetration (more so for tapes and drives than discs), and physical shock. To help prolong its operational life, store your media in a dry environment with a comfortable and stable room temperature. Encapsulate all media in plastic during transportation. Provide cases or plastic sheaths for discs, and avoid handling them excessively.

6. **Multimedia (e.g., photo, video, audio) backup and archiving is a bit more challenging, in part because transcriptions and captioning can be important for interpretation, discovery, and accessibility.** Storage of images solely on local hard drives or servers is not recommended. Unaltered images should be preserved at the highest resolution possible. Store original images in separate locations to limit the chance of overwriting and losing the original image.There are a number of options for metadata for multimedia data, including MPEG standards, PBCore, and the US Library of Congress standards for still images, sound, and moving images.

7. **Finally, consider the short- and long-term security of the originals:** specimens, samples, documents and data sheets, photographs, or other physical items. Options include:

    - Campus Office (long-term security/stability: poor)
    - Lab (long-term security/stability: moderate)
    - Museum/Herbarium (long-term security/stability: high)
    - Departmental Office ((long-term security/stability: poor/moderate)
    - Commercial storage facility (long-term security/stability: depends on funding source)
    - Home ((long-term security/stability: poor)
    - Library (long-term security/stability: high)
    - Dryad, Zenodo, OSF, or other digital repository with long-term archiving via CLOCKSS or similar (long-term security/stability: high)

6. Backup Procedures

Backing up requires discipline and care to ensure nothing is missed and the file back-ups are secure. You can make this easier by doing the following:

1. **Plan ahead, and write down your plan.** A written data back-up and security plan includes such information as:

    - where backups are located
    - who can access backups and how they can be contacted
    - how often data should be backed up
    - what kind of backups are performed
    - who is responsible for performing the backups and their contact information. For large projects or projects with high-volume data streams include a person with secondary responsibility (the back-up back-up) in case the primary person responsible is unavailable
    - what hardware and software are used or recommended for performing backups
    - how and how often to check if backups have been performed successfully
    - the media are used to backup data
    - a list of any data that are *not* archived or backed up

2. **Backup your data at regular frequencies**, with backup strategies (e.g., full, incremental, differential) optimized for the data collection process and data type. For instance, you should (at least):

    - Back up when you complete your data collection activity
    - Back up after you edit / clean data
    - Streaming data (e.g., from data loggers) should be backed up at regularly scheduled points in the collection process
    - High-value data should be backed up at much higher frequencies (daily)

3. **Put procedures in place to make sure you follow your plan.**

    - set up calendar reminders
    - use checklists.
    - Ensure backup copies are identical to the originals (using checksums, file counts, etc.)
    - Be sure of one of the regular calendar items is to verify successful recovery from a backup copy.
    - ***Automate as much as possible.*** Automation simplifies frequent backups. but be careful with accidental file loss and that back-up copies are actually independent (e.g. Dropbox and a synced folder on your laptop are *not* independent copies).

4. **Ensure the Security of Restricted Data.**

- All data collected by UF affiliates falls into one of the classifications defined in the UF Data Policy. Both original and electronic archives must be secured; failure to do so can have serious consequences for both researchers and institutions.

7. File formats for storage

For long term preservation is it necessary to store data in file formats that will be readable in the future and by a variety of operating systems. It is also important to provide descriptive information on these data file types and formats. This will facilitate data retrieval and reuse.

1. Document and store data using file-types that are open (ie., non-proprietary), uncompressed, unencrypted (if at all possible), and stable:

   - `ASCII` formatted files will be readable into the future. For tabular data use comma-separated values, `.csv`, for text use `.txt` (alternatively `.rtf`).
   - Images: `TIFF` (uncompressed)
   - Video: MPEG-4 (`.mp4`) or motion JPEG 2000 (`.jp2`)
   - AudioL Free Lossless Audio Codec (`.flac`)
   - Documentation: `.txt` (preferred), PDF/A (`.pdf`)
   - Structured, highly coded data: `.xml`
   - For geospatial (raster) data the following provide a stable format:
     - `GeoTIFF/TIFF`
     - `ASCII Grid`
     - Binary image files
     - `NetCDF`
     - `HDF` or `HDF-EOS`
   - For image (Vector) data use the following file formats (these are mostly proprietary data formats; please be sure to document the Software Package, Version, Vendor, and native platform):
     - ARCVIEW software: store components of an ArcView shape file (`.shp`, `.sbx`, `.sbn`, `.prj`, and `.dbf` files)
     - ENVI – `.evf` (ENVI vector file)
     - ESRI Arc/Info export file (`.e00`)

   - ***Note:*** Certain file formats, such as shapefiles, can be made up of as many as 7 individual files. If one of those files is absent from the file assembly the shapefile data utility may be lost. Awareness of adherence to a particular file format standard can also be helpful for determining, for example, if a particular software package can read the data file. Awareness of whether that standard or format is open source or proprietary will also influence how and if the data file can be read.

2. If a particular software package required to read and work with the data file, you need to make sure you have a copy, a backup copy, *and a computer with the proper operating system to use the software.*

3. For audio and image files, back up with the highest-quality file-type possible (i.e., the 'lossless' format rather than the 'lossy' format). For example, `jpeg` is lossy, meaning images saved with this file type will have less detail than images saved with the lossless `TIFF` format if at all possible back-up.

8. Tools & Resources

8.1. UF Library Guides

- UF Library: Data File Types and Formats
- UF Library: Data Storage
- UF IRB: Data/Record Storage and Security
- UF IT: Data Classifications for Confidentiality & Security

8.2. Overview of 'Checksums' and Checksum Calculators

- Fisher, Tim. What Is a Checksum? (updated 15 April 2020).

8.3. File Formats and Extensions

- What is XML?
- Common Windows File Extensions (NMU IT)
- Common File Extensions (PC Net)
- List of Open File Formats (Wikipedia)
- Common Windows File Extensions (NMU IT)
- Common File Extensions (PC Net)
- List of Open File Formats (Wikipedia)
- File formats-Best Practices (Stanford Library)
- Library of Congress

8.4. UF-based Storage Options

- Overview of UF Cloud Data Services, aka GatorCloud
- Free UF Dropbox for Education, UF Dropbox FAQs, and information on how your advisor can submit a request for your UF Dropbox account
- GStuite Allocation for UF personnel
- Free MS One Drive (but also read why I suggest NOT using this here)
- for very large data volumes or data data required extra security protection due to privacy restrictions, etc. seek the assistance of UF Research Computing

8.5. Storage and backup for very large volumes of data (multiple TB)

- Cyverse
- Amazon Web Services (AWS) S3
- UF Research Computing
- Open Storage Network

- Argonne Leadership Computing Facility
- Globus
- Uploading large files to Google drive: tutorial blog post

8.6. Tools for File Syncing

- MacDropAny (macOS)
- MultCloud
- Dropbox (use via UF for research, teaching)
- Google Drive (use via UF for research, teaching)

8.7. Portable HDs and Backup Services

- **Remember:** Dropbox and Google drive are not true backups if set up as a file syncing tool.
- NY Times *Wirecutter* Reviews & Recommendations:
- portable hard drives
- automated data backup services; their recommendation was SpiderOak.

8.8. Tools for Automating Data Uploading During Data Collection:

- IFTTT and others

8.9. Online multimedia repositories and tools

- YouTube
- Vimeo
- Flickr
- Google Photos

These services have some advantages:

- are often low-cost or free
- are open to all

- have functions for provide community commenting and tagging
- some provide support for explicit licenses and re-use
- provide some options for valuable metadata such as geolocation
- may allow for large-scale dissemination
- optimize usability and low barrier for participation

However, they are commercial services, and hence have a number of potential drawbacks:

- models for sustainability is profit-based
- may have limits on file size or resolution
- may have unclear access, backup, and reliability policies, be sure to review them carefully

There are specialized photo repositories designed with researchers in mind.

- Tropy
- MorphBank

Some advantages of these are:

- often open-source, free, developed and maintained by academic institutions
- domain-specific metadata fields and controlled vocabularies customized for expert users
- highly discoverable for those in the same domain
- they can provide assistance in curating metadata
- they can optimize scientific use cases such as vouchering, image analysis
- may provide APIs for sharing or re-use for other projects
- they are recognized as high-quality, scientific repositories

But they too have some drawbacks

- rely on research or institutional/federal funding
- may require high-quality multimedia, completeness of metadata, or restrict manipulation
- may not be open to all
- may have restrictions on bandwidth usage

8.10. A step back: what needs to be preserved?

To meet multiple goals for preservation, researchers should think broadly about the digital products that their project generates, preserve as many as possible, preserve all that they are required to preserve for as long as required or longer, and plan the appropriate preservation methods for each. Consider how long and how to preserve the following, taking into account **(1)** what would be necessary to reconstruct a complete data set if files downstream were lost **(2)**, how long each needs to be kept, and **(3)** the cost of and space required for preserving data for different lengths of time:

1. Raw data (written form, electronic form). *Raw data are almost always worth preserving.*
2. Tables, spreadsheets, or databases of raw observation records and measurements

3. Tables, spreadsheets, or databases of clean observation records and measurements. *If clean data can be easily or automatically re-created from raw data, consider not preserving. If quality control or analysis is time-consuming or expensive, then consider preserving the clean version.*
4. Intermediate products: partly summarized or coded data that are the input to the next step in an analysis
5. Documentation of protocols used to clean, summarize, or code data
6. Software or algorithms developed to prepare data (cleaning scripts) or perform analyses. *Algorithms and software source code cost very little to preserve*
7. Results of an analysis, which can themselves be starting points or ingredients in future analyses, e.g. distribution maps, population trends, mean measurements. *These may be particularly valuable for future discovery and also cost very little to preserve.*
8. Any data sets obtained from others that were used in data processing

9. Sources

1. DataONE Community Engagement & Outreach Working Group, DataONE (July 01, 2010) "Best Practice: Plan for effective multimedia management". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/bestpractices/plan-for-effective on Aug 31, 2020
2. DataONE Best Practices Working Group, DataONE (July 01, 2010) "Best Practice: Ensure integrity and accessibility when making backups of data". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/best-practices/ensure-integrity-and on Aug 31, 2020
3. Cindy Parr, Heather Henkel, DataONE (Aug 30, 2011) "Best Practice: Create and document a data backup policy". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/bestpractices/create-and-document on Aug 31, 2020
4. Cindy Parr, Heather Henkel, Keven Comerford, DataONE (May 11, 2011) "Best Practice: Decide what data to preserve". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/bestpractices/decide-what-data on Aug 31, 2020

10. Data Organization in Spreadsheets

10.1. Make your data `tidy`

- Spreadsheets should be a rectangle, with only rows and columns.
- Each column is a different variable (a thing you are measuring, like 'weight' or 'temperature').
- One row per observation. Each cell has only one value.
- Column headers: Use short meaningful column names with no spaces or special characters. Don't start column names with numbers. Record units in column headers.
- Don't enter the same data on multiple spreadsheets: Use one for each category of data to avoid duplicated data and to simplify corrections (e.g., taxonomy).
- Never put multiple tables in a single spreadsheet.
- Avoid spreading data across multiple sheets
- *Collecting* data in tidy format makes it easier to *enter* data in tidy format.

10.2. Use consistent names, abbreviations/codes, and capitalization.

- Write dates as YYYYMMDD. Better still have separate columns for Year, Month, and Day.
- Excel is unable to parse dates from before 1899-12-31. Be careful if your data include a mix of dates before and after this date, then you'll have mixed data types in one column. [1]
- Record zeros with a numeral (`0`), not a blank cells. For missing data use an appropriate null value indicator (e.g., `NA`).
- Don't use formatting to convey information or to make your spreadsheet look pretty.
- *Remember that data format and excel defaults can vary by region.* For example, depending on the part of the world where a user is based, the default value for the decimal and thousands operator could be a `,` (comma) or a `.` (period); some regions use mm-dd for dates while others use dd-mm.

[1] The reason dates in Excel are so weird is that it is *accounting software*. It counts the days from a default of December 31, 1899, and thus stores July 2, 2014 as the serial number 41822. This is so one can can easily calculate "days from a given date" for accounting purposes (like invoicing) by adding "date+XX days".

10.3. DO NOT EDIT OR CORRECT RAW DATA FILES!

- Once you are done with data entry, save your file in 'read only' format and make *all* corrections using scripting.
- Do *not* edit raw data after you have entered it in your spreadsheet!*

10.4. Readings, Tools, & Resources

1. Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. The American Statistician, 72(1), 2-10. [read online]
2. Data Validation in Google Sheets: blog post and video tutorial. A pdf version is available for download here.
3. Why not bypass spreadsheets like Excel and use a csv editor like [Comma Chameleon][https://comma-chameleon.io/] instead? CC and other csv editors allow you to enter data in the same way - into cells, by adding and removing rows - and then export your file. But that's about it, which means you can't do many of the things (e.g., calculations, color in cells) that cause problems down the road.
4. More advanced users comfortable with R can also look into Data Curator, with which you can create and edit tabular data from scratch or from a template, open Microsoft Excel and CSV files, and automatically correct common problems found in these and other file types.
5. DataONE Community Engagement & Outreach Working Group (2017) "Data Quality Control and Assurance". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/lessons/05_qaqc/index on Aug 31, 2020
6. DataONE Community Engagement & Outreach Working Group (2017) "Data Entry and Manipulation". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/lessons/04_entry/index on Aug 31, 2020
7. Chris Prener, Trevor Burrows (Eds.). Data Carpentry: Data Organization in Spreadsheets for Social Scientists.
8. Peter R. Hoyt, Christie Bahlai, Tracy K. Teal (Eds.), Erin Alison Becker, Aleksandra Pawlik, Peter Hoyt, Francois Michonneau, Christie Bahlai, Toby Reiter, et al. (2019, July 5). datacarpentry/spreadsheet-ecology-lesson: Data Carpentry: Data Organization in Spreadsheets for Ecologists, June 2019 (Version v2019.06.2). Zenodo. http://doi.org/10.5281/zenodo.3269869

Part II.

In-Class Activities

11. Course Introduction: In-Class Work

11.0.1. This week's assignment is to briefly answer the following questions. *If you respond to the Google forms survey there is no need to submit anything via Canvas*

This survey has two goals:

1) To learn a bit more about your interests and experience so that I can tailor the modify the course content to match the cohort's interests.

2) To gather some information the data set(s) you might use for the Data Cleanup Project: one of the goals this semester is to take a "messy" data set, clean it, organize it, and – assuming you have permission and if appropriate at this time – archive it in a repository. This assignment will introduce me to your data set so we can evaluate if it is suitable for the assignment (i.e., is it messy enough, is it big enough?). The data can be from previous research, even if it is unpublished or not part of a thesis.

It is important to remember that the 'data set' can be more than one file. For instance, your data set might be four files (or collections of files): (1) A file with the responses to surveys of community members, (2) Census data downloaded from a government website, (3) The transcripts of interviews with key informants, and (4) A collection of photographs for ethnographic research.

Please also remember that you may not have a data set with which to work. That's fine! There are plenty of (useful) data sets out there that are messy and in need of organization. I'll help you find one that matches your interests.

11.1. Survey Questions

1. UFID Number
2. Last Name (as it appears on Canvas)
3. Do you have a name or nickname by which you prefer to be called?

4. Canvas has an option for students to list their preferred pronouns. If your preferred pronouns aren't available as an option on Canvas, please feel free to tell me what they are here.
5. What degree are you pursuing?
    - MA/MS
    - PhD
    - Other
6. What is your department or degree program?
7. Please provide a brief summary - 1 paragraph max - of your research interests.
8. Do you have any questions or concerns about the class you'd like to share with me? Perfectly acceptable answers include: "I'm worried about the workload", "I've never taken a bio class", "I'm worried we might have to move on-line in a few weeks because of covid,"I'm worried classes are inperson instead of on-line because of covid", and "None".

Potential Data Set for Course Project

9. Do you have Messy Data you would like to clean up and organize as your course project? This can be data from your current or prior research, including side projects, Master's thesis, undergraduate thesis, etc. *If you don't have a data set of your own, THAT'S NO PROBLEM - I will help you find one.*
    - Yes
    - No
    - Maybe
10. What research question(s) are being addressed with these data?
11. What is the status of this data set?Leave only the most relevant answer.*
    - I have finished collecting the data (Skip to question 13)
    - I have started data collection but will *not* be done before the semester ends (Skip to question 13)
    - I have started data collection and *will* be done before the semester ends (Skip to question 13)
    - Other (please elaborate) (Skip to question 12)

11.1.1. Data Set Status = Other

12. If you answered "other" to the previous question, briefly elaborate here.

11.1.2. Tell me about the data set(s)

13. What are they and what kind of information do they contain?
14. What tools and techniques did you use to collect it?
15. What format are they in (e.g., sheets of paper, notebooks, spreadsheet, text files, photographs)
16. How much data is it? (estimated number of files, surveys, photographs, etc).
17. Who collected the data?
    - I did / I am / I will
    - A 3rd party (e.g., my advisor, another student, a collaborator)
    - It it is previously published or public data Other:
18. Are the data from research conducted with human subjects?
    - Yes
    - No
19. Are any collaborators currently working with the data?
    - Yes
    - No
20. Will anyone (e.g., collaborators, advisor) be working with the data in the future?
    - Yes
    - No
    - Don't Know
21. What instruments did you use for COLLECTING and RECORDING raw data (check all that apply)
    - paper survey or data recording forms
    - online surveys
    - notebooks/forms in which personal observations, notes, or responses to questions were recorded
    - audio recordings
    - data from automated data loggers
    - photographs
    - video
    - medical instruments (EKG, eye trackers; please elaborate below)
    - tablets/smartphone
    - Other (elaborate in "Data Collection" section below)
22. What were you gathering on or about? Check all that apply.
    - field-based experiments, observations, or data collection with live animals or plants

- museum collections: herbarium specimens, animals (including tissue or blood), fossils
- plant or non-human animal tissue or blood (not from a museum specimen).
- human subjects - behavior, observations, experiments (education, psych, linguistics, etc)
- human subjects - tissue, blood, or similar human subjects - records (medical, educational)
- museum collections - human remains
- museum collections - archeological (pottery, middens mounds)
- museum collections - anthropological
- museum collections - art (painting, sculpture)
- Other (elaborate in "Data Collection" section below)

11.1.3. Data collection Instruments & Subjects

23. If you answered "other" to either of the 2 previous questions, please elaborate on the data collection instruments or subjects.

11.1.4. Instruments

24. What instruments and software do you use for PROCESSING and ORGANIZING data? Check all that apply

    - MS Word
    - Google Docs
    - Other word processor
    - MS Excel
    - Google Sheets
    - Other spreadsheet software
    - Scrivener
    - OSF
    - Evernote
    - Flickr
    - Text editor
    - Other (elaborate below)

25. If you answered "other", please list the other tools you use for PROCESSING and ORGANIZING your research data

26. If appropriate for your research domain, do you use any of the following to analyze your data?

27. If you answered "other", please list the other tools you use for ANALYZING your research data

28. Are you an R user? NOTE: YOU ARE NOT EXPECTED TO HAVE ANY PROGRAMMING EXPERIENCE TO TAKE THIS CLASS.

    - I have never used R

- Yes - a novice R user
- Yes - an intermediate R user Yes - an advanced R user

29. Will you be gathering data this summer or fall?

    - Yes
    - No
    - Not Sure

30. Is there anything you'd like to add? Are there particular topics you'd like to see covered or tools you'd like to learn?

11.2. Grading Rubric:

1. Survey completed with thorough answers: 50

2. Most questions answered completely; some require instructor follow-up: 40

3. Many questions missing answers or answers are cursory: 30

4. Instructor follow-up required for survey submission: 20

Part III.

Assignments

12. Assignments Overview

Grades in the course will be based on the following assignments:

12.1. Weekly In-class activities

- **Overview**: Most of the in-class assignments involve hands-on practice with data collection or manipulation. In some weeks, however, assignment will be the submission of questions for group discussion or brief reflection on the issues from the readings. Most in-class assignments are designed to be completed during the class session, but to ensure students master the concepts rather than rush through them *they can be submitted anytime until 9 am the following Friday.*

12.2. Reproducible Data Organization Project

- **Overview:** This project is an opportunity to put some of the lessons learned into practice with a data set of your own. Your assignment is to **(1)** clean and organize a 'messy' data set and prepare metadata describing the resulting 'clean' data. The complete project requires the submission of these three items via the course Canvas website:

  (1) R code that imports, cleans, and organizes, and saves 'messy' data

  (2) The resulting corrected and organized data in an appropriate format

  (3) Metadata describing the corrected data set

12.3. Thesis Research DMP

- **Overview:** The Data Management Plan (DMP) is a critical document describing the data to be collected for a research project, how it will be stored and managed, and the investigator with primary responsibility for its management. Many funding agencies, including NSF and NIH, now require a DMP with all grant applications. Each student will prepare a Data Management Plan (DMP) for their thesis research.

Part IV.

Appendix 1: Useful Resources

13. Useful resources for Data Collection & Managment

13.1. R Programming

13.1.1. Essential

1. Hadley Wickham wrote a book on using the tidyverse and the online version is FREE. This is a phenomenal resource on using R to import, tidy, and visualize data.
2. Posit Cheat Sheets: help with commands for using the different `tidyverse` packages, RStudio shortcuts and tricks, help with R commands, and more. You definitely want the ones for Data Import, Work with Strings, Factors, Data Transformation, and Base R.
3. Where and How to ask for help
   - Hadley Wickham's advice on how to write a good reproducible example for getting help with R
   - how to post good questions on StackOverflow
   - The UF R-users listserv is *very* user friendly and a great place to post requests for help.

13.1.2. Tutorials and Books

1. R Essential Training: Wrangling and Visualizing Data
2. Software Carpentry: Using RStudio for Project Organization & Management
3. Swirl
4. R Bootcamp
5. Kieran Healy's *Data Visualization: a practical introduction* is my favorite introductory (yet super-comprehensive) book on data visualization with R. If you scroll down to the bottom of the page you can download the datasets and code used to make the figures in the book, which makes life much easier.
6. So. Many. Resources.
7. ROpenSci: tools for accessing, manipulating, and visualizing open data

8. How to clean messy data in R
9. The Ultimate Guide to Data Cleaning
10. Learning R
    Swirl

14. Specific Problems in Data Cleaning and Managemnt

1. Handling dates and times in R
2. Text Mining: `tidytext` package
3. Working with Qualtrics survey data with the `qualtRics` package
4. Optical Character Recognition (OCR): extract text from images: `tesseract` package
5. Extract text & metadata from pdf files: `pdftools` package
6. Image processing in R: the `magick` package

14.0.1. Advanced R Packages

1. `DataCurator` package: 'a simple desktop data editor to help describe, validate and share usable open data'.
2. RegExr: online tool to learn, build, & test Regular Expressions (RegEx / RegExp)
3. janitor (cleanup of file names, etc.)
4. `knitr` overview: reproducible documents with R
5. qualtRics
   ## Discipline-specific Resources
6. `historydata` package: Sample data sets for historians learning R. They include population, institutional, religious, military, and prosopographical data suitable for mapping, quantitative analysis, and network analysis.
7. *The Programming Historian* Website: wide range of topics, from text analysis to OpenRefine

14.1. Slide Presentations in R

1. Make slide presentations with R

14.2. Data Archives

Qualitative Data Repository

14.3. Text Extraction and Organization

Plan for extraction and organization

14.4. Form Design

Best Practice for Form Design

14.5. Data Security

UF Office of Information Security and Compliance
Cyber Safeguards for UF
UF IRB
UF Data Classification Policy
UF Office of Information Security and Compliance

14.6. Platforms for Organization & Collaboration

Open Science Framework

15. Slide decks

Here are the slide decks:

- Introduction — Slides
- File Names — Slides