LAS 6292: Data Collection and Management

Emilio M. Bruna

November 15, 2026

Table of Contents

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 the Github documentation.

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:

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.

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.

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.

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'`

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`.

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

Best Practices & Conventions for Naming Files

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

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
- 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:**

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?

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

3.1. What is 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

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***

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)

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:

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

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)

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).

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.

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.

Store data and documentation 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.

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.

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.

3.2. Tools & Resources

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

Overview of 'Checksums' and Checksum Calculators

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

File Formats and Extensions

- UF library recommended Data Formats for Preservation Purposes in the Florida Digital Archive
- 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

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

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

Tools for File Syncing

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

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.

Tools for Automating Data Uploading During Data Collection

- IFTTT and others

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

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

3.3. Reading and 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/bestpractices/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

4. Data Organization in Spreadsheets

4.1. Things to remember when organzing Data in Spreadsheets

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.

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.

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.

[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”.

- Do *not* edit raw data after you have entered it in your spreadsheet!

4.2. Tools & Tutorials

1. Data Validation in Google Sheets: blog post and video tutorial. A pdf version is available for download here.
2. 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.
3. 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.
4. Chris Prener, Trevor Burrows (Eds.). Data Carpentry: Data Organization in Spreadsheets for Social Scientists.
5. 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

4.3. Reading and Sources

1. ***Essential reading:*** Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. The American Statistician, 72(1), 2-10. [read online]
2. 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
3. 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

5. Reproducibile Data Cleanup with R

5.1. Set up an RStudio Project and install the relevant packages

1. File -> New Project
2. Name the Project as follows: `las_demo`
3. Create the following three folders:

   ```
   data_raw
   data_clean
   code
   ```

   You can create these in the folder using your operating systems "create folder" option ***or*** you can create within R studio using the **`files`** tab
4. Install libraries, load libraries, and run verify they work by running a few commands

5.2. Gather the Raw Data to be corrected

Save your original (aka 'raw' or 'dirty') data files in the `data_raw` folder.

> **!** Important
>
> 1. Never make changes to the original file of dirty/raw data. Never. **NEVER**. Always import the uncorrected original data file, make corrections using a script, and then save a new, clean file.
>
> 2. *Annote your code with reminders for future you:* the # symbol (aka pound, hash tag) allows you to make comments in the script explaining what different sections or commands do. Annotate your scripts with lots of details on what different commands do, what is being done in each section, even links to websites you used to figure stuff out. Remember: ABA *(always be annotating)*.

5.3. *Take your data from `messy` to `clean` in these 6 steps:*

1. Familiarize yourself with the data set
2. Check for potential errors. These can be structural errors (e.g., misaligned columns, duplicated rows/columns, missing values), data entry errors, or measurement errors. Decide how you will flag and deal with them.

3. Decide how to deal with missing values
4. Identify ways to simplify data values (i.e., codes, abbreviations) and column headings
5. Write code to load the 'raw' data file file, implement your corrections/changes, and save the 'clean' version of the data.

Things to look out for: Make a list of the what you think needs to be corrected and the steps necessary to identify and implement each correction. Some of the things to look out for include:

- Numeric values stored as character data types
- Factors stred as characters
- Duplicate rows
- Spelling mistakes
- inconsistent formatting (eg., codes, capitalizations)
- White spaces
- Missing data
- Zeros instead of null values
- Special characters (e.g. commas in numeric values instead of decimals)
- column headings with spaces between words or that start with numerals

Remember, *the characteristics of clean data set include:*

- Free of duplicate rows/values
- Error-free (correct misspellings, eliminate special characters)
- correct data type for analysis
- outliers identified and dealt in the correct way
- "tidy" data structure

When working with multiple files, should you correct *before* or *after* combining?

We often need to combine multiple files with the same kind of data (i.e., surveys conducted in Year 1 and in Year 2, each of which are recorded in their own .csv file). Is it more efficient to correct each file first, then combine them, or combine first and *then* correct?

It depends and can vary from one project to another. Make an outline of the different steps and corrections to be made to each file and see if you can decide which is more efficient. Note that there might be different ways to do the same thing, this outline will help figure out which is best. For instance you could:

> ***Option 1***
> 1. Import table 1
> 2. Correct column headings in Table 1
> 3. Import table 2
> 4. Correct column headings in Table 2
> 5. Bind Table 1 and Table 2 Together
>
> ***but this is less efficient than...***
>
> ***Option 2***
> 1. Import table 1

2. Import table 2
3. Bind Table 1 and Table 2 Together
4. Correct the column headings in the Table

5.4. Tools & Resources

1. These introductions to R and R Studio were made by Professor Ethan White (UF-WEC). They are a good overview of some R basics.

    - Intro to R and RStudio.
    - Navigate R and RStudio web page
    - Intro to R Packages
    - Expressions and Variables in R

2. The Carpentries' R workshops (self-paced or taught in-person) are excellent, I use many of their materials in class:

    - R for Social Scientists
    - Data Analysis and Visualization in R for Ecologists

3. Software Carpentry lesson on Project Management with R Studio

4. 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.

5. RStudio 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.

6. 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.

7. Ten simple rules for biologists learning to program

8. Lot's more on the course's 'Resources' page

6. Additional (interesting) Reading

1. Lewis, Keith P., Eric Vander Wal, and David A. Fifield. 2018. Wildlife biology, big data, and reproducible research. Wildlife Society Bulletin 42(1): 172-179.
2. White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR. 2013. Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution. 6(2):1-10.
3. The humanities have a 'reproducibility' problem
4. The humanities do not need a replication drive
5. Reproducible Research: A primer for the social sciences
6. Replicability and replication in the humanities
7. Towards reproducible science in the digital humanities
8. The possibility and desirability of replication in the humanities
9. Reproducible Research: A Retrospective

For when you feel more comfortable with R and programming

1. Bryan, J. (2018). Excuse me, do you have a moment to talk about version control? The American Statistician, 72(1), 20-27.
2. Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. Good enough practices in scientific computing. PLoS Computational Biology 13, no. 6 (2017): e1005510.

7. QA/QC: Reducing Error Entry, Finding and Correcting Errors

7.1. Reducing errors when entering data

7.1.1. Using 'Data Validation' Rules

1. Setting up Data Validation in Microsoft Excel: [link to Microsoft post] and a tutorial video.

2. Setting up Data Validation in Google Sheets: [link to post by Google] and a tutorial video

Data entry with 'Text-to-Speech'

MS Excel (& other MS programs)

1. Converting text to speech in Excel for Microsoft 365, Excel 2010-2021 [link].
2. Microsoft for Mac: Hear selected text read aloud from Excel, Word, Power Point, and Outlook [link]. See also *'System Preferences->Accessibility'*
3. Video Tutorials

- Tutorial Video 1. This will allow you to select rows or columns of Excel and have them read back to you.
- Tutorial Video 2: it's not as thorough, but it is a bit easier to see the menu
- Tutorial Video 3. Because why not a third one?

Text-to-Speech in Google Docs

1. If you prefer working in Google Docs you can do the same thing. This article will show you how. You can also watch this tutorial video.

Better still: 'speak-on-enter' to confirm data values as entered

1. Speak-on-enter [tutorial:]

Speech-to-Text for keyboardless data entry

1. Google Sheets: Typing with your voice
2. MS Word: How to dictate documents

Overview of Microsoft Accessibility Tools

1. 'Narrator' function link

7.2. R Packages for Data QA/QC

1. janitor: simple functions for examining and cleaning dirty data. It was built with beginning and intermediate R users in mind and is optimized for user-friendliness. Advanced R users can already do everything covered here, but with janitor they can do it faster. (see this mini-tutorial)

    - clean column names
    - remove empty rows and columns
    - remove duplicated rows

2. cleanr. small R package for cleaning and checking data columns in a fast and easy way.
3. unheadr: used to wrangle spreadhseets with embedded subheaders or values wrapped accross several rows (highlighting, merged cells, etc).

For more advanced R users:

1. Richard's Iannone `pointblank` R package for Data Validation and Organization of Metadata [link
2. validate: designed to test data against a reusable set of data validation rules, investigate, summarize, and visualize data validation results, among other things
3. Data Curator is a simple desktop data editor to help describe, validate and share usable open data.
4. Kim, A. Y., Herrmann, V., Barreto, R., Calkins, B., Gonzalez-Akre, E., Johnson, D. J., Jordan, J. A., Magee, L., McGregor, I. R., Montero, N., Novak, K., Rogers, T., Shue, J., & Anderson-Teixeira, K. J. (2022). Implementing GitHub Actions continuous integration to reduce error rates in ecological data collection. Methods in Ecology and Evolution, 13, 2572–2585. https://doi.org/10.1111/2041-210X.13982

7.3. Readings & Sources

1. Campbell, J. L. *et al.* 2013. Quantity is nothing without quality: automated QA/QC for streaming environmental sensor data. BioScience, 63(7): 574-585. link

2. ***For more advanced users of R/Github:*** Kim, A. Y. *et al.* 2022. Implementing GitHub Actions continuous integration to reduce error rates in ecological data collection. Methods in Ecology and Evolution, 13, 2572– 2585. https://doi.org/10.1111/2041-210X.13982

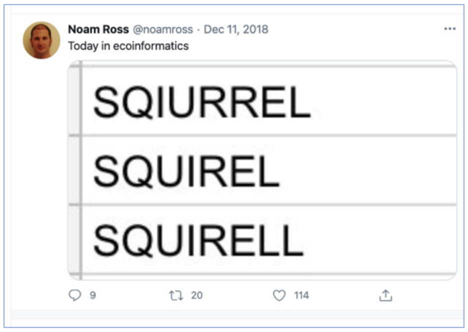
3. Barchard, K. A., & Pace, L. A. (2011). Preventing human error: The impact of data entry methods on data accuracy and statistical results. Computers in Human Behavior, 27(5), 1834-1839. link

4. Atkinson, I. (2012). Accuracy of data transfer: double data entry and estimating levels of error. Journal of Clinical Nursing, 21(19pt20), 2730-2735.link

5. Goldberg, S. I., Niemierko, A., & Turchin, A. (2008). Analysis of data errors in clinical research databases. AMIA Annual Symposium proceedings. AMIA Symposium, 2008, 242–246. link

6. DataONE Education Module: Data Quality Control and Assurance. DataONE. Retrieved Nov12, 2012. From http://www.dataone.org/sites/all/documents/L05_DataQualityControlAssurance.pptx

8. Class Outline: QA/QC 2 - Open Refine



Noam Ross @noamross · Dec 11, 2018

Today in ecoinformatics

SQIURREL

SQUIREL

SQUIRELL

Figure 8.1.: How many ways can you spell....

8.1. OpenRefine

Open Refine is a powerful, free, and open source tool that is used to work with and clean messy data. The website includes a user's manual and links to more tutorials.

8.2. OpenRefine Tutorials

- Data Carpentry: Data Cleaning with OpenRefine for Social Scientists.

- Data Carpentry: Data Cleaning with OpenRefine for Ecologists
- Environmental Data Initiative OpenRefine Tutorial
- Cleaning Data with OpenRefine Video Tutorials:
  - Video Tutorial No. 1
  - Video Tutorial No. 2
- JHU Library: Cleaning Data with OpenRefine
- The Programming Historian: Cleaning Data with OpenRefine.

8.3. GREL Cheatsheets

- Belinda Weaver's GREL Cheatsheet with examples
- OpenRefine GREL Manual
- A really good GREL Guide from the Univ Illinois

- Even better: code4lib Toronoto’s OpenRefine cheatsheets, including for GREL commands.
- Datenschule’s OR Cheatsheets

8.4. R Tools for OpenRefine

- The rrefine package allows you to do some OpenRefine tasks from within R.

9. Resources for Preparing Metadata

9.1. What are Metdatadata?

1. E. Bruna Video for LAS6292: Why Metadata?
2. ICPSR: What is a Codebook?

9.2. Best practices for preparing metadata

1. "Best Practices in Creating Social Science Metadata." p.32 in the ICPSR *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (6th Edition).*
2. Michener, W.K., et al . 1997. Non-geospatial metadata for the ecological sciences. Ecological Applications 7: 330–342.
   [read online]
3. Pp 446-450 in Bernard, H.R. and Bernard, H.R., 2013. Social research methods: Qualitative and quantitative approaches. Sage.
4. ICPSR *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (6th Edition)*
5. DataONE Community Engagement & Outreach Working Group (2017) "Metadata Management". Accessed through the Data Management Skillbuilding Hub at https://dataoneorg.github.io/Education/lessons/07_metadata/index on Aug 31, 2020

9.3. Discipline-specific metadata standards

It is worth looking in these catalogs to see if you can find metadata standards for your discipline. They will provide suggestions on not only what to include, but the standard vocabulary for your discipline.

1. RDA Catalog of metadata standards for different disciplines
2. UK Digital Curation Center Directory of metadata standards for different disciplines
3. Ecological data: Ecological Metadata Language
4. Museum Specimens: Darwin Core

5. Geography Markup Language (GML): Emphasis on geographic features (roads, highways, bridges)

6. Humanities: UF Digital Collections (UFDC) key metadata fields used for non-published items such as posters, archival materials, artists' files, field notebooks, etc. Includes a link to a template you can download. See also Template No.1 from UF's Samuel Proctor Oral History Project, Template #2, which is a more general one from the UF Humanities Archives, and the metadata required by the Qualitative Data Repository.

9.4. Metadata templates

I have created metadata templates based on information from ICSPR (for social sciences) and Michener *et al.* 1997 (for biophysical sciences) that can be downloaded and edited; you can add more fields or delete any that are not relevant. Note that Table 1 in Michener *et al.* is much more comprehensive and provides additional guidance on how to make sure the metadata are useful. The templates are available in `.txt`, `.Rmd`, and `.qmd` format).

1. **Metadata Template for Social Sciences** based on ICSPR standards: .txt format or .Rmd format.

2. **Metadata template for Biophysical Sciences** based on Table 1 from Michener *et al.* 1997: .txt format or .Rmd format

3. **Metadata for the Humanities or those working primarily with Qualitative Data:** The metadata required often depend on the type of material with which you work (e.g., oral history, photos, digital, printed). If your data is in this domain, you can use this general template from the UF Humanities Archives: Template #2. You can also review the metadata required by the Qualitative Data Repository.

9.5. Tools for creating machine-readable metadata

I include these here in case you want to try using them. It's not required, but it could definitely make your life easier if there is a standard tool for your discipline (e.g., MORPHO if you are working with ecological data).

1. Morpho: desktop application that allows researchers to create metadata and then (if they wish) upload to KNB. No longer maintained but open source and can be very useful.

2. giant list from the RDA of tools for creating standardized metadata for different disciplines

3. USGS Metadata Wizard

4. TKME

5. CatMDEdit

6. GRIIDC

9.6. Metadata Dictionaries

1. USGS
2. Global Change Master Dictionary
3. USGS Geographic Names
4. Getty Thesaurus of Geographic Names

9.7. Organizations developing metadata standards and schema

1. The Research Data Alliance (RDA)"has the goal goal of building the social and technical infrastructure to enable open sharing and re-use of data."
2. DDI Alliance: "Established in 2003, the Data Documentation Initiative Alliance (DDI Alliance) is an international collaboration dedicated to establishing metadata standards and semantic products for describing social science data, data covering human activity, and other data based on observational methods."
3. The Dublin Core Metadata Initiative is "an organization supporting innovation in metadata design and best practices across the metadata ecology".

10. Data Management Plans

11. Reminder: What is a DMP?

- Formal document that lays out your plan for managing the data you will collect (or have collected).
- Outlines what you will do with your data during and after you complete your research
- Ensures your data is safe for the present and the future
- Video overview by RWTH Aachen University (12 min).

11.1. Why do one?

- Save time
- Less reorganization later
- Increase research efficiency
- Makes it easier to archive data downt he road (some data centers and repositories have specific requirements for data documentation, and knowing these requirements in advance saves time and effort that might be spent trying to correct data after they are collected).
- Increasingly: you don't have a choice. Many funding agencies (e.g., NSF, NIH, NEH) now require a DMP for proposal submission.

11.2. What are the components of a DMP?

1. Information about data & data format
2. Metadata content and format
3. Policies for access, sharing and re-use
4. Long-term storage and data management
5. Roles and responsibilities
6. Budget

For details on what information should be included in each component, see "DMP Details" below.

Example: DMP for NSF Proposals

"Plans for data management and sharing of the products of research. Proposals must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results, and may include: the types of data, samples, physical collections, software, curriculum materials, and

> other materials to be produced in the course of the project the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies) policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements policies and provisions for re-use, re-distribution, and the production of derivatives plans for archiving data, samples, and other research products, and for preservation of access to them."
> *Some funders or institutions may require specific elements in a data management plan; you should check in advance to make sure you meet all requirements*

11.3. Start Here

1. The DMP Tool (https://dmptool.org/) is a free, open-source, application that helps researchers create data management plans (DMPs). It provides an easy-to-use platform for creating a DMP that complies with the requirements of different private and public funding agencies (e.g., foundations, NSF, NIH). It also has direct links to funder websites, help text for answering questions, and data management best practices resources.

11.4. What should go in my DMP?

1. Qualitative Data have some unique issues regarding their description and storage, so the Qualitative Data Archive has put together a great checklist of things to consider.
2. Short Checklist & Questions to Consider when putting together a DMP
3. Very in-depth Checklist & Questions to Consider when putting together a DMP
4. UF Library Guidelines for preparing a DMP

11.5. Sample DMPs

NSF DMPs

1. NSF General: Mauna Loa example
2. NSF General: Rio Grande example
3. NSF DEB: Emilio Bruna Heliconia example

Sample DMPs on the 'DMP Tool' website

1. Public DMPs (***note: "Public DMPs are plans created using the dmptool service and shared publicly by their owners. They are not vetted for quality, completeness, or adherence to funder guidelines."***)

Humanities DMPs

1. National Endowment for the Humanities: Data Management Plans From Successful Grant Applications (2015-2018) and (2011 - 2014) can be found as downloadable zip files on this page
2. A *really* nice DMP for a Digital Humanities Project by researchers at the University of Maryland, College Park.

11.6. DMP Details

1. Information About Data & Data Format

1.1 Description of data to be produced

- Experimental
- Observational
- Raw or derived
- Physical collections
- Models and their outputs
- Simulation outputs
- Curriculum materials
- Software
- Images
- Other

1.2 How data will be acquired?

- When? Where? Method?

1.3 How data will be processed

- Software used
- Algorithms
- Workflows
- Data transformations/formats needed (Consider archive policies)
- *This step is important to consider before the project since it may affect how data are organized, what formats are used, and how much should be budgeted for hardware and software. Things to consider are what software may be used, what algorithms will be employed, how these fit into the overall workflow of the project.*

1.4 File formats

- Justification
- Naming conventions

1.5 Quality assurance & control during sample collection, analysis, and processing

1.6 Existing data

- If existing data are used, what are their origins?
- Will your data be combined with existing data?
- What is the relationship between your data and existing data?

1.7 How data will be managed in short-term

- Version control
- Backing up
- Security & protection
- Who will be responsible

2. Metadata Content & Format

- Metadata defined:
- Documentation and reporting of data
- Contextual details: Critical information about the dataset
- Information important for using the data
- Descriptions of temporal and spatial details, instruments, parameters, units, files, etc.

2.1 What metadata are needed

- Details that make data meaningful

2.2 How metadata will be created and/or captured

- Lab notebooks? GPS units? Auto-saved on instrument?

2.3 What format will be used for the metadata

- Standards for community
- Justification for format chosen

3. Policies for Access, Sharing, Reuse

3.1 Obligations for sharing

- Funding agency
- Institution
- Other organization
- Legal

3.2 Details of data sharing

- How long?
- When?
- How access can be gained?

- Data collector rights

3.2 Ethical/privacy issues with data sharing

3.3 Intellectual property & copyright issues

- Who owns the copyright? (? Institutions and/or funding agencies often have a policy for intellectual property and copyright. There may be other considerations such as embargos on data due to patents, politics, or journal requirements.)
- Institutional policies
- Funding agency policies
- Embargos for political/commercial reasons

3.4 Intended future uses/users for data

- This helps determine the most appropriate data center to archive your data.

3.5 Citation

- How should data be cited when used?
- Do you need a persistent citation (e.g. DOI)?

4. Long-term Storage & Data Management

4.1 What data will be preserved 4.2 Where will it be archived

- Most appropriate archive for data
- Community standards

4.4 Who will be responsible

- Contact person for archive

5. Roles and responsibilities

5.1 Outline the roles and responsibilities for implementing this data management plan.

- Who will be responsible for data management and for monitoring the data management plan? How will adherence to this data management plan be checked or demonstrated? What process is in place for transferring responsibility for the data? Who will have responsibility over time for decisions about the data once the original personnel are no longer available?

6. Budget

6.1 Anticipated costs

- Time for data preparation & documentation
- Hardware/software for data preparation & documentation
- Personnel
- Archive costs

6.2 How costs will be paid

12. Transcription & Translation

12.1. Transcription

Transcription tools & software

The UF International Ethnography Lab has access to Trint Transcription Software, which has detailed tutorials

Reviews of transcription tools

1. Comparison Video
2. NY Times' Wirecutter review of transcription services

Tools: Automated Transcription via file upload

Top-Tier {.unnumbered}

1. Otter.ai. English only (US & UK with regional accents); free (600 min/mo) and paid plans. Can transcribe zoom meetings (e.g., if you do interviews via zoom) and can integrate with Dropbox. Phone app is available for iOS. There is an online tutorial. ***English-only***
2. Happy Scribe - see descriptionabove.
3. Temi. ***English Only***, less accurate than Otter.ai and Happy Scribe.
4. Online (free demo) version of Watson; note the "detect multiple speakers" option.
5. Zoom via UF; see also this Link and this announcement about the addition of new languages for UF users

Second-Tier

1. Spext. Speech-audio-text editing; free and paid plans.
2. Amazon Transcribe

Tools: Automated transcription via API

1. `googleLanguageR`. Uses the Google Cloud machine learning APIs for text and speech tasks. The Cloud Speech API transcribes sound files to text and the Cloud Text-to-Speech AP
2. `AWS transcribe` uses the Amazon Web Services to do transcription via API. Has program for NGO/non-profits.

Tools: Automated transcription via dictation

1. Google Docs. See also Google Docs Voice Recognition and helpful how-to blog post on transcription using google docs and voice recognition
2. SpeechNotes
3. Apple Dictation and Dictation Commands
4. How to use Windoes Voice Dictation and Windows Speech Recognition Commands

Tools: "Manual" transcription

1. oTranscribe Incredibly easy, web based, with time stamps, variable speed.
2. Express Scribe and a tutorial. There is a free version.
3. Transcribe
4. Nvivo
5. Listen N Write
6. InScribe
7. Pearnote for Mac: make text annotations that connect to locations in audio and video recordings (30 day free trial)
8. Dragon Speech Recognition Software
9. Sonix
10. Adobe Premiere: Speech-to-Text Tutorial and Workaround if you are using a version later than 7.2.2. UF Students can use Adobe products on UFApps

Transcription Style & Coding Guides

1. Samuel Proctor Project Style Guide
2. Library of Congress Guide: How to Transcribe
3. Smithsonian Archives of American Art Style Guide
4. Smithsonian Volunteer Transcription Program's Guide and Expert Tips
5. Guide to Transcription Coding
6. Hepburn, A., & Bolden, G. B. (2017). Transcribing for social research. Sage. linkto book descrption on Sage website
7. Manchester University's very useful Transcription Toolkit and Guide
8. Manchester University's Quick calculator for estimating transcribing time (.xls file)
9. Programming Historian Guide to Text Automation

Examples of Transcription Notes

1. Sample 1
2. Sample 2

12.1.1. Setting up Text Shortcuts

1. Google Docs
2. MS Word

Useful Websites and Tutorials

Dictionaries, Slang, and Colloquialisms

1. Linguee Good for slang and expressions;searches different ways in which one word has been translated. Can also enter several words or even a sentence to find different translations.
2. Thesaurus.com; equivalent in foreign language.

Text comparison

1. Side-by-Side Text Comparison Website

Tutorials and Reviews

1. Jennifer Marie's YouTube channel on Transcription Tools & Tips

Equipment

Pedals

1. Infinity
2. Kinesis

Misc. Software & Apps

1. SayMore: tools for common Language Documentation tasks
2. Lameta tool for metadata associated with files made in the course of documenting language, music, and other cultural expressions.
3. voice recorder app for iPad
4. App for recording phone interviews (note: be aware of relevant local and national laws regarding recording phone calls).

Post-Transcription Text Analytics

1. Planning for text analytics (extraction and organization) in R

2. Silge and Robinson's **outstanding** book 'Text Analysis in R'

12.2. Translation

Tools: Automated Translation of Text

1. Google Translate and help documents.
2. Google Docs: in-doc translation. See also how to translate using google sheets
3. MS Word: in-doc translation
4. DeepL

Tools: Automated Translation of Audio

1. Happy Scribe. 40+ languages. 30 min. free then paid; student pricing available. Can integrate with Zapier to up/download files from Google Drive and Dropbox ***Multiple languages***
2. Google Audio Translation in Beta ### Text or Audio: API or Command Line
3. `googleLanguageR`. Uses the Google Cloud machine learning APIs for text and speech tasks. The Cloud Translation API does detection and translation of text.
4. `aws.translate` uses the Amazon Web Services to do transcription via API. Has program for NGO/non-profits.

1. Watson (IBM) 500 minutes free
2. DeepL
3. Translate Shell via API

12.2.1. Tools: Manual Translation of Text with Computer Assisted Technology (CAT)

1. Video overview of different CAT Tools from 'Freelanceverse'.
2. OmegaT CAT Software. (Free)
3. WordFast Anywhere (free online; desktop version are paid)
4. Across.net (basic version free)

12.3. Sources & Additional Readings

1. Young, JC, Rose, DC, Mumby, HS, et al. 2018. A methodological guide to using and reporting on interviews in conservation science research. *Methods Ecol Evol* 9:10-19 [link]
2. Michael Henry Heim & Andrzej W. Tymowski. Guidelines for the Translation of Social Science Texts. *American Council of Learned Societies*
3. Sophie Chabeda, Jane Kahindi, Manya Van Ryneveld. Preparing data: the not-so-simple stage of transcription and translation
4. Squires, A. 2009. Methodological challenges in cross-language qualitative research: A research review.. *International Journal of Nursing Studies* 46(2):277-287.
5. Hepburn, A., & Bolden, G. B. (2017). Transcribing for social research. Sage.
6. McLellan E, MacQueen KM, Neidig JL. 2003. Beyond the Qualitative Interview: Data Preparation and Transcription. *Field Methods*. 15(1):63-84. [link]
7. MacLean LM, Meyer M, Estable A. 2004. Improving Accuracy of Transcripts in Qualitative Research. *Qualitative Health Research*.14(1):113-123. [link]

8. Hughes, M. Are Online Transcription Services Safe and Private?. *How-to-Geek*.

9. Duca, Daniela. Who's disrupting transcription in academia?. *Sage Ocean*.

10. Oliver, D. G., Serovich, J. M., & Mason, T. L. (2005). Constraints and opportunities with interview transcription: Towards reflection in qualitative research. Social forces, 84(2), 1273-1289. [link]

11. Moore, E., & Llompart, J. (2017). Collecting, Transcribing, Analyzing and Presenting Plurilingual Interactional Data. Research-publishing.net. [link]

13. Paperless Data Collection

13.1. Tools & Resources

1. EpiCollect user's guide

1. Lists of useful apps for data collection in the field: Start with the Bruna Lab Super-List *(if you want to suggest any additions to the list please do so here)*. You can also check [list 2], [list 3], [list 4], [list 5], [list 6], and [list 7]

2. Paid platforms for paperless data collection

- Fulcrum; link to post describing how it was used by an NGO.
- ODK

13.2. Sources & Additional Reading

1. Streamlining Field Data Collection with Mobile Apps

2. Research at Home: Using a Smartphone for Data Collection

3. Keep calm and go paperless: Electronic lab notebooks can improve your research

4. NOAA Pathways to paperless Data Collection

5. Thieler ER, Zeigler SL, Winslow LA, Hines MK, Read JS, Walker JI (2016) Smartphone-Based Distributed Data Collection Enables Rapid Assessment of Shorebird Habitat Suitability. PLoS ONE 11(11): e0164979. link

6. Teacher, A. G., Griffiths, D. J., Hodgson, D. J., & Inger, R. (2013). Smartphones in ecology and evolution: a guide for the app-rehensive. Ecology and evolution, 3(16), 5268–5278. link

7. Palumbo, M.J., Johnson, S.A., Mundim, F.M., Lau, A., Wolf, A.C., Arunachalam, S., Gonzalez, O., Ulrich, J.L., Washuta, A. and Bruna, E.M. (2012), Harnessing Smartphones for Ecological Education, Research, and Outreach. The Bulletin of the Ecological Society of America, 93: 390-393. link

8. Rayhan, R. U., Zheng, Y., Uddin, E., Timbol, C., Adewuyi, O., & Baraniuk, J. N. (2013). Administer and collect medical questionnaires with Google documents: a simple, safe, and free system. Applied Medical Informatics, 33(3), 12-21. link

9. Moylan, C. A., Derr, A. S., & Lindhorst, T. (2013). Increasingly mobile: How new technologies can enhance qualitative research. Qualitative social work: research and practice, 14(1), 36-47. link

10. Aanensen DM, Huntley DM, Feil EJ, al-Own F, Spratt BG (2009) EpiCollect: Linking Smartphones to Web Applications for Epidemiology, Ecology and Community Data Collection. PLoS ONE 4(9): e6968. link

14. Automated Data Collection & Extraction

14.1. Optical Character Recognition

1. Video Primer: What is OCR?

Online OCR Tools (text & data from `.pdf` to `.csv`, `.txt`, etc.)

1. Google Drive - Video Primer: OCR with Google Drive
2. Free online sites for small batches (can upgrade for larger numbers of files)
   - Free Online OCR 1
   - New OCR
   - pdf to excel
   - OnlineOCR
   - PDFTables will convert PDF to .csv, and has an API so you can do your conversions in bulk with R. You can do ~25 pages free; large numbers are reasonably priced.
3. Amazon TextExtract
4. Mathpix Snip digitizes handwritten or printed text, and copies outputs to the clipboard that can be pasted into LaTeX editors like Overleaf, Markdown editors like Typora, Microsoft Word, and more.

OCR with R

1. R package `pdftools`
2. R package `tabulapdf`
3. written tutorial 1
4. written tutorial 2
5. written tutorial 3
6. Video Tutorial 1
7. Video Tutorial 2
8. Detailed Blog Post / Tutorial
9. Convert PDF to text in R OCR pdftools

10. PDFtools in R

11. More advanced but more powerful from the Programming Historian: OCR with Google Vision API and Tesseract

14.2. Extracting tables from images with R

1. R package `magick` *(this package actually includes several very powerful tools for image processing; this is just one of the things you can do with it)*
2. Detailed Blog Post / Tutorial

14.3. Extracting Data from Published Figures

1. Ankit Rohagni's Web Plot Digitizer

   - WPD Video Tutorial
   - WPD Tutorial Blog Post

2. Alternative 1: R package `magick`

3. Alternative 2: GetData extracts data automatically from scanned images (~$30).

4. Alternative 3: R package `digitize` will extract data from scatterplots within the R environment. This article will walk you through the process.

14.4. Text Mining

1. Text Mining with R by Julia Silge and David Robinson

2. `gutenbergr`: Download and Process Public Domain Works from Project Gutenberg. Tutorial can be found here

Useful reading on text mining

1. Atanassova I, Bertin M and Mayr P (2019) Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics. Front. Res. Metr. Anal. 4:2. doi: 10.3389/frma.2019.00002

2. Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S (2018) A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput Biol 14(2): e1005962. https://doi.org/10.1371/journal.pcbi.1005962

3. Salloum, S.A., Al-Emran, M., Monem, A.A., Shaalan, K. (2018). Using Text Mining Techniques for Extracting Information from Research Articles. In: Shaalan, K., Hassanien, A., Tolba, F. (eds) Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence, vol 740. Springer, Cham. https://doi.org/10.1007/978-3-319-67056-0_18

4. Simon, C., Davidsen, K., Hansen, C. et al. BioReader: a text mining tool for performing classification of biomedical literature. BMC Bioinformatics 19, 57 (2019). https://doi.org/10.1186/s12859-019-2607-x Extracting Body Text from Academic PDF Documents for Text Mining

5. Benchimol, J., Kazinnik, S., & Saadon, Y. (2022). Text mining methodologies with R: An application to central bank texts. Machine Learning with Applications, 8, 100286.link

6. Yu, C., Zhang, C., & Wang, J. (2020). Extracting Body Text from Academic PDF Documents for Text Mining. arXiv preprint arXiv:2010.12647.

7. Gulo, C. A., & Rúbio, T. R. (2015, January). Text Mining Scientific Articles using the R. In Doctoral Symposium in Informatics Engineering. linl

14.5. Collecting & Processing data from the Web of Science and Scopus

1. R package `refsplitr`

2. R package `bibliometrix`

3. jstor: An R package for Analysing Scientific Articles

14.6. Scraping websites

1. Library Carpentry Lesson on Webscraping
2. Start Here: Introduction to webscraping
3. Video: Scraping WebData in R with rvest
4. Video: Practical Introduction to Web Scraping using R
5. Very nice written tutorial…
6. ….and another one, this time from the UC Business Analytics R Programming Guide
7. scraping HTML text and scraping HTML tables
8. SelectorGadget is useful to id CSS selectors.
9. Noortje Marres & Esther Weltevrede (2013) Scraping the Social?, Journal of Cultural Economy, 6:3, 313-335, DOI: 10.1080/17530350.2013.772070

14.7. Cell Phone Data

1. Exploratory analyses Part 1 and Part 2

14.8. Social Media Data

1. How to extract Biodiversity Data from Facebook

2. Fox, Nathan, Tom August, Francesca Mancini, Katherine E. Parks, Felix Eigenbrod, James M. Bullock, Louis Sutter, and Laura J. Graham. ““photosearcher” package in R: An accessible and reproducible method for harvesting large datasets from Flickr.” SoftwareX 12 (2020): 100624. https://www.sciencedirect.com/science/article/pii/S235271102030337X

14.9. Automated Image Analysis

1. Pennekamp, F. and Schtickzelle, N. (2013), Implementing image analysis in laboratory-based experimental systems for ecology and evolution: a hands-on guide. Methods Ecol Evol, 4: 483-492. https://doi.org/10.1111/2041-210X.12036

2. How to build your own image recognition app with R! Part 1 and Part 2

14.10. Wearable Devices & RFID tags

1. What is an RFID tag?

2. Rafiq, K., Appleby, R. G., Edgar, J. P., Radford, C., Smith, B. P., Jordan, N. R., Dexter, C. E., Jones, D. N., Blacker, A. R. F., & Cochrane, M. (2021). WildWID: An open-source active RFID system for wildlife research. Methods in Ecology and Evolution, 12, 1580– 1587. https://doi.org/10.1111/2041-210X.13651

3. Build your own RFID device

4. Izmailova, E.S., Wagner, J.A. and Perakslis, E.D. (2018), Wearable Devices in Clinical Trials: Hype and Hypothesis. Clin. Pharmacol. Ther., 104: 42-52. https://doi.org/10.1002/cpt.966

5. Loncar-Turukalo T, Zdravevski E, Machado da Silva J, Chouvarda I, Trajkovik V. Literature on Wearable Technology for Connected Health: Scoping Review of Research Trends, Advances, and Barriers J Med Internet Res 2019;21(9):e14017 doi: 10.2196/14017

6. Why Should Sociologists Care about Wearable Tech?

7. Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. Perspectives on Psychological Science 11(6), 838-854 link

8. Seifert Alexander, Hofer Matthias, Allemand Mathias. 2018. Mobile Data Collection: Smart, but Not (Yet) Smart Enough. 12. Frontiers in Neuroscience https://www.frontiersin.org/article/10.3389/fnins.2018.00971

14.11. Buildimg automated data collectors

1. Calipers that dump data directly to Excel link
2. PiSpy: An Affordable, Accessible, and Flexible Imaging Platform for the Automated Observation of Organismal Biology and Behavior
3. Jolles, J. W. (2021). Broad-scale applications of the Raspberry Pi: A review and guide for biologists. Methods in Ecology and Evolution, 12, 1562– 1579. https://doi.org/10.1111/2041-210X.13652

14.12. Online data archives

Overview: Correia, R.A., Ladle, R., Jarić, I., Malhado, A.C.M., Mittermeier, J.C., Roll, U., Soriano-Redondo, A., Veríssimo, D., Fink, C., Hausmann, A., Guedes-Santos, J., Vardi, R. and Di Minin, E. (2021), Digital data sources and methods for conservation culturomics. Conservation Biology, 35: 398-411. https://doi.org/10.1111/cobi.13706

Government data

1. Data.gov (the open data portal of the US Government) and Using Data.gov APIs in R
2. the rOpengov Project
3. Open Fiscal Data Package
4. `educationdata`: Retrieve data from the Urban Institute's Education Data API as a data.frame for easy analysis. See also here
5. a huge list of data sources for social scientists available with R tools
6. accessing World bank Data with R

US & World Census Data

1. A Guide to Working with US Census Data in R
2. R Package `tidycensus`
3. Tutorial 1
4. Tutorial 2
5. R package `ipumsr`: The ipumsr package helps import IPUMS extracts from the IPUMS website into R. IPUMS provides census and survey data from around the world integrated across time and space.

Education Data

1. `edbuildr`: import EdBuild's master dataset of school district finance, student demographics, and community economic indicators for every school district in the United States.
2. Building R and Stata packages for the Education Data Portal

Other Online Data Portals

1. Giant compendium of open datasets #1
2. Data on Amazonia

3. R package `bdc`: toolkit for gathering & cleaning biodiversity data

Software for gathering data from online archives

1. EcoRetriever: automates the tasks of finding, downloading, and cleaning up publicly available ecological data, and then stores them in a local database or csv files.
2. litsearcher an R package to facilitate quasi-automatic search strategy development for systematic review

15. Ethics & Legal Compliance

15.1. Data/Research Ethics & Compliance for UF Researchers

1. UF Research Misconduct Homepage: Policy, Training, Reporting, and Best Practices
2. UF Office of Research Integrity and Compliance (Research Integrity & Compliance)
3. Info for UF Researchers regarding Risk Assesment for International Collaborations and Export Control
4. UF IACUC
5. UF IRB
6. UF Privacy Office

15.2. Privacy Laws

1. CDC: what is HIPPA?
2. Global Privacy Laws, including the EU General Data Protection Regulation (GDPR)? (link 1), (link 2), (link 3 - UF)
3. What is FERPA?

15.3. Anonymizing Data

Overview

1. UK Data Service Webinar: How to anonymise qualitative and quantitative data
2. Intro to Anonymization
3. Overview by Martin Monkman
4. EDUCASE Guidelines for Data De-Identification or Anonymization
5. UF "HOW TO: DE-IDENTIFY DATA", with very useful **checklists**

Tools for Anonymizing Data Overview

Tip with Title

Anymizing data is difficult and the consequences for failing to do so can be severe. Be ***very*** careful - or better still, seek guidance from experts on campus.

Anonymizing with R

1. Martin Monkman's link to R packages for Anonymizing Data
2. Blog Post / Tutorial
3. R package `Anonymizer`
4. Anonymizing in R: part 1 and part 2

Stand-alone software and online tools for anonymizing

1. list of free tools
2. ARX (open source, free)
3. Amnesia (open source, free)
4. De-identification Tools from NIST
5. AirCloak (paid)

15.4. Working with Indigenous Communities

1. Indigenous peoples and responsible data: an introductory reading list
2. CARE Principles for Indigenous Data Governance
3. Carroll, S.R., Herczog, E., Hudson, M. et al. Operationalizing the CARE and FAIR Principles for Indigenous data futures. Sci Data 8, 108 (2021). https://doi.org/10.1038/s41597-021-00892-0

15.5. Data Collection, Privacy, and Gender

1. How to Ethically and Responsibly Identify Gender in Large Datasets
2. Clair A Kronk et al. 2022. Transgender data collection in the electronic health record: Current concepts and issues. *Journal of the American Medical Informatics Association* 29(2):271–284 https://doi.org/10.1093/jamia/ocab136
3. Heidari, S. et al. 2016. Sex and Gender Equity in Research: rationale for the SAGER guidelines and recommended use. *Res Integr Peer Rev* 1,2 (2016). https://doi.org/10.1186/s41073-016-0007-6
4. The European Institute for Gender Equality's GEAR Tool (Gender Equality in Academia and Research)
5. UN "Methods for gender data collection and estimation" Training Materials.
6. Colaço, R., and Watson-Grant, S. (2021). A Global Call to Action for Gender-Inclusive Data Collection and Use. RTI Press Publication No. PB-0026-2112. Research Triangle Park, NC: RTI Press. https://doi.org/10.3768/rtipress.2021.pb.0026.2112

15.6. Emerging issues resulting from technological advancements

1. Automated Data Collection: The Big Picture and Legal Issues
2. Are transcription services HIPPA compliant?

3. Di Minin, E. et al. 2021. How to address data privacy concerns when using social media data in conservation science. *Conservation Biology*, 35:437-446. https://doi.org/10.1111/cobi.13708

15.7. Text mining, copyright, and the law

1. Text and Data Mining of In-Copyright Works: Is It Legal?
2. The plan to mine the world's research papers
3. Copyright and the Progress of Science: Why Text and Data Mining Is Lawful
4. Copyright's impact on data mining in academic research
5. Text and Data Mining at Springer Nature
6. UF Library Guide: Copyright on Campus: Essentials
7. UF Library Guide: Data Sharing & Copyright

15.8. Ethics of Data Sharing

1. Commit to transparent COVID data until the WHO declares the pandemic is over
2. NIH issues a seismic mandate: share data publicly

Part II.

Class Notes & Activities

16. Introduction & The Research Data Life Cycle: Instructor Notes

Objectives and Competencies

By the end of this lesson students will:

- Gain an appreciation for why I think this subject matter is important
- Understand how different disciplines define and use 'data'
- Learn the 'Research Data Life-Cycle'
- Overview of the Course Structure, Syllabus

16.1. Pre-Class Preparation & Materials Needed (Instructor):

- email students: confirmation of room and zoom link, pls bring own mug or water bottle, welcome to bring food / snacks
- Ask if any students need laptop, arrange for a loan if needed
- Copy of syllabus and Course Roster
- Flip charts, markers, tent cards for names, dry-write markers

16.2. Pre-class Preparation (Students):

- Readings or Online Lectures: None

16.3. Class: 'Data' across disciplines & the Data Life Cycle

Breakout 1: Introductions

Get to know one another! Get in pairs, introduce yourselves, and record responses to the following questions. We will then come back to introduce your partner to the group.

Tell me about yourself

1. Name
2. In what city were you born?
3. What you consider your "hometown"?
4. Program and Degree?

1. Hobbies or what you do to relax / have fun. 1. What is your motivation for taking this class? 1. Any concerns / worries / fears about this class (in particular) and this semester (in general)? 1. Now discuss with your partner the following question: ***What are 'Data?***

Report Back {.unnumbered}

1. What are Data?
    1. Students take turns introducing themselves
    2. Quick summary of the different ways they defined "data"

Intro to Breakout 2: Motivation for Teaching This Course & Data Across Disciplines

1. The defintion of "Data"

2. EB Motivation for teaching the class:

- Reduce Student Stress
- Research integrity: identifies fraud, a shield to protect you when you are right
- Because it really matters to society, both when we get it wrong and when we get it right
- Because your data are a potential gold mine for future researchers…but only if you take a few important steps now.

2. Prompt for the next breakout (to start after break); Breakout into larger groups (n = 4)

BREAK

Breakout 2 Groups: 'Data' across disciplines

Now that data have been defined, we will break into small groups to discuss the following questions. We'll come back and discuss what you came up with for as a group.

1. Identify different kinds of data collected in different disciplines
2. How are these data gathered (tools, techniques) and recorded (media)?

Breakout 2 Discussion

1. What are the different kinds of data collected in different disciplines?
2. How are these data gathered (tools, techniques) and recorded (media)?

After each group reports back, discuss the following questions as a group:

1. Are data types and data recording methods that are unique (or at least much more common) in the biophysical sciences, social sciences, humanities, or other disciplines?

2. It is likely that across disciplines the issues and types of data being used are often very similar, but there might be some differences (e.g., text corpora from humanities).

3. One key is to emphasize that an important part of 'data' likely not brought up by the groups is information about how they were recorded, decisions about how to code them, corrections, etc.

4. **Important:** look to see if the definitions include the word "PLAN": managing data requires a PLAN. This will let you segway into…

Wrap-up & In-class Assignment: Instructor presentation: Research Data Cycle & Syllabus Overview

1. Typical vs. New approach to Data
2. Most courses start with the most boring part: DMP. I take an approach different from many other data manag classes.
3. Introduce Assignments and Format of Weekly Sessions. This is a workshop!
4. *This course is driven by student needs and interests!* The syllabus may change as I learn more about those.

16.3.1. **In-class Assignment: Survey {.unnumbered}

1. Please complete the survey in Canvas under "Assignments".

16.4. After class:

- Be sure you complete and submit the assignment by deadline and prepare for next session by doing reading and watching videos.

17. File names, formats, & organization; Data Storage & Backup: Instructor Notes

Objectives and Competencies

- Describe and implement conventions for proper naming of files
- Explain the difference between proprietary and open formats
- Learn how to efficiently organize their research data files
- Learn the preferred format for storing and archiving different types of data files
- Become familiar with different options for cloud data storage and backup
- Develop and implement a plan for short- and long-term data storage, back-up, and archiving
- Learn rules and policies for data security
- Become familiar with tools for such tasks as batch renaming of files, cloud data storage, and automated data backup.
- Explain options for a long-term sustainable preservation strategy/policy for your data (e.g., discipline specific, institutional, departmental, individual).
- Address the need for conversion to standard formats needed for re-use
- Perform basic archival processes: checksum, auditing, format migration, etc.
- Understand costs & time lines for data storage, management tools and services

17.1. Pre-Class Preparation & Materials Needed (Instructor):

Send in an email to students:

- Confirmation of room and zoom link
- Remind students to bring their computers
- Make sure you know if everyone has R and instlled; level of R fluency
- Snacks
- Copies of the syllabus
- Copy of Course Roster
- Flip charts and markers
- Dry write markers
- Tent cards for student names

17.2. Pre-class Preparation (Students):

- **Readings:**

1. Jan Čurn. 2014. How a bug in Dropbox permanently deleted my 8000 photos. [read online] [download pdf]
2. PSA: Scrivener, Data Integrity and You. Or, How To Avoid Data Loss Heartbreak. [read online] [download pdf]
3. Hart EM et al. (2016) Ten Simple Rules for Digital Data Storage. PLoS Comput Biol 12(10): e1005097. [read online] [download pdf].
4. Panzarino, M. 2012. How Pixar's Toy Story 2 was deleted twice, once by technology and again for its own good. TNW. [read online] [[download pdf]]

- **Online Lectures:**

  1. Video on Project Organization
  2. Video on File Names
  3. Video on Storage and Backup

17.3. Class Outline

17.3.1. Quick Intro

1. Address any questions from last week

2. Overview of today's activities.

17.3.2. File Format Competition

17.3.3. Snack Break

17.3.4. Breakout: Discussion of Data Security and Backup in the field

Prompt: Robin is a graduate student studying Malaria in Tanzania. The research project requires visiting communities, collecting mosquitoes from different sources of standing water (for later identification at the FLMNH and statistical analyses of mosquito diversity and abundance) and conducting semi–structured interviews in the community, where people are asked such things as their age, family income, their access to health care, if they have ever had malaria, their family income, use of mosquito nets. Answers are recorded on data sheets, but the interviews are also recorded with the cell phone for later transcription. Robin also takes pictures of houses to document standing water sources. Throughout the day Robin records observations in a notebook. Robin is staying in a house where there is electricity for a laptop. Once every two weeks Robin goes into town, where there is usually good internet access at the university.

What steps should Robin take in the field to safeguard the data collected in the field?

Breakout Discussion

1. Group discussion of breakout
2. *Key message:* assume the worst case scenario. become paranoid. embrace neurosis. then relax because the plan is in place and all possibilities have been acounted for.

Breakout 2: Backup Procedures - UF Research Group

Prompt: Please ask each other these questions and then briefly discuss and summarize the responses. Keep the answers anonymous. The goal is to document the range of answers to these questions, so you can summarize or give more detailed answers about individuals as needed.

1. Does your advisor / lab / research group have formal policies that govern your data storage, protection, and backup?
2. Are you currently backing up your research data?
3. Do you have a 'backup plan' document?
4. How frequently are you backing up your research data?
5. Is it Partial (incremental) or full backup?
6. Where are files backed up?
7. What metadata accompanies these backups?
8. How do you verify that a backup has been successfully performed?
9. Have you ever attempted to read data from older backups?
10. Have you ever had to restore a file from a backup version?
11. Are you working with data requiring additional protection due to privacy or security concerns?
12. If so, what are the additional safeguards you have implemented?
13. Where and how are your original data collection instruments or samples stored/protected?

Collect answers and submit with Assignment 2

17.4. Messages to instill:

```
a. a sense of paranoia that everything that could go wrong with notebooks, datasheets, s
b. 'Focused Laziness': we want to do this in a way that is as automated (and automatic)
```

17.5. Assignment for Submission

This week's assignment is designed to: (A) get you thinking critically about your own data backup and security procedures, and based on that reflection **(B)** prepare a data backup plan (if you already have one you can answer the questions below based on that plan). **The assignment comes in four parts:**

1. Sign up for a UF Dropbox allocation: https://cloud.it.ufl.edu/collaboration-tools/dropbox/. If you are not automatically eligible to sign up, your advisor can request that you be granted an allocation: https://cloud.it.ufl.edu/media/clouditufledu/How-to-Obtain-Access-to-UF-Dropbox-for-Education.pdf.
2. Sign up for a UF Google Suite Allocation: https://cloud.it.ufl.edu/collaboration-tools/g-suite/
3. Briefly describe how you are adhering to the 3-2-1 backup rule for data related to your thesis. If you are not, describe how you will do so moving forward.
4. Prepare a brief (~1 page max) Data Backup Plan for your thesis research. Include the following information; you can respond with bullet points unless more detailed answers are needed:

   a. what needs to be prepared and for how long;

   b. where backups are located;

   c. who can access backups and how contacted;

   d. how often data should be backed up;

   e. what kind of backups are performed;

   f. who is responsible for performing the backups;

   g. hardware and software used for performing backups;

   h. how / how often to check if backups is successful;

   i. the media are used to backup data;

   j. a list of any data that are not archived or backed up.

Submission and Grading Rubric

1) Submit the Data Backup Plan document (in either `.txt` or `.pdf` format) via Canvas.
2) Grading Rubric

   - Assignment completed with thorough answers: 50
   - Most questions answered completely; some require instructor follow-up: 40
   - Many questions missing answers or answers are cursory: 30
   - Instructor follow-up required for homework submission: 20

17.6. After class:

- Be sure you complete and submit the assignment by deadline
- Prepare for next session (assigned reading, videos, etc).

18. Data Organization in Spreadsheets: Instructor Notes

Objectives and Competencies

- Be able to identify different categories of data
- Learn best practices for data entry
- Recognize and avoid common problems with data entry and formatting in spreadsheets
- Learn and be able to implement 'Tidy' format for data tables in spreadsheets
- Identify problems with and approaches for proper handling of dates in spreadsheets
- Learn how to export data from spreadsheets in open format

18.1. Pre-Class Preparation (Instructor):

Send in an email to students:

- Remind everyone about the computer with spreadsheet software prior to class

Bring to Class:

- Snacks
- Flip charts and markers
- Dry write markers
- Tent cards for student names

18.2. Pre-class Preparation (Students):

Online Lectures: None

Readings

1. Tesi, W. 2020. An Outdated Version of Excel Led the U.K. to Undercount COVID-19 Cases. Slate. [read online]
2. Stolberg et al. 2020. CDC Test Counting Error Leaves Epidemiologists 'Really Baffled'. NY Times. [read online]

3. Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. The American Statistician, 72(1), 2-10. [read online]

4. Johnson, B. D., Dunlap, E., & Benoit, E. (2010). Organizing “mountains of words” for data analysis, both qualitative and quantitative. Substance Use & Misuse, 45(5), 648-70. [read online]

Computer Resources

1. **Please bring your computer**. If you need to borrow a laptop or get access to a computer let me know.

2. You will need a **spreadsheet program** such as MS Excel or LibreOffice installed on your computer. You can use the program of your choice, but the exercises are optimized for MS Excel. UF students can download the Microsoft Office suite free of charge here; if you don’t want to install Excel you can use the online version for free (see this tutorial video on how to do so). LibreOffice is a free and open source package similar to (and compatible with) MS Office. It can be downloaded here.

3. The paper by Broman and Woo on how to organize data in spreadsheets is *especially* important; it may well be one of the more helpful papers you read while a student. Really.

4. It is also time to install two other pieces of software. We won’t use them until Week 4, but it is worth installing them now to make sure they are working smoothly:

   (i) **the R programming language**: You can download the version of R for your computer operating system here; it’s free.

   - ***For Mac computers*** click on ‘Download R for (Mac) OS X’, then chose the version of R for your operating system, i.e., macOS 10.13 (High Sierra), OS X 10.11 (El Capitan), Mac OS X 10.9 (Mavericks), etc.

   - ***For PC computers***: click on Download R for Windows, then click ‘base’ or ‘install R for the first time’.

   - ***For Linux***: click on ‘Download R for Linux’; after which you are on your own. Then again you use Linux, so you probably don’t need my help anyway.

   - ***If you need help:*** watch this video tutorial or contact me. *Note that the tutorial requires you be a UF affiliate and either on campus or logged to the UF network via VPN.*

   (ii) **RStudio**: the software used to work with R. There are other ‘environments’ one can use for R programming, but RStudio is by far the most widely used and useful.

   - We use the ‘Free Open-Source Desktop Version’, which you can download here. Choose the version for your computer operating system and install as you would any other software.

- **If you need help**: watch this video tutorial or contact me. *Note that the tutorial requires you be a UF affiliate and either on campus or logged to the UF network via VPN.*

(iii) **Verify the installations** worked by opening RStudio to see if it opens properly. If you are really motivated, you can also install the `Tidyverse` library by starting RStudio and at the console typing `install.packages("tidyverse")`.

18.3. Session Introduction

Types of Data

Down the road when doing data correction, organizing data, and doing analyses it will be essential to classify data according to their 'type'. It can also help with data entry, which is why we will introduce some of these types here:

1. **Nominal aka Factor:** categories or groups, such as [`apple`, `orange`], [`trumpet`, `flute`, `violin`]
2. **Ordinal aka Ordered Factor:** groups where there is an order: [`first`>`second`>`third`], [`small`<`medium`<`large`]. Note that this order doesn't imply quantitative value, e.g., One is not stating that `medium` is twice the size of `small` or that `large` is twice the size of `medium`.
3. **Character**: [`a`, `gnv`, `mexico`, `Inigo Montoya`]
4. **Numeric (real or decimal):** `2`, `15.5`
5. **Integer:** [`1`,`2`,`3`]
6. **Logical:** [`True`, `False`]
7. **Complex:** `1+4*i`
8. **Interesting case:** what category are [`red`, `orange`, `green`, `blue`]? We usually treat it as Nominal, but it is actually Ordinal - the colors represent wavelengths on the visible light spectrum (650, 600, 550, and 450 nm respectively). If you were recording the wavelength itself, it would be Numeric.
9. *For more on different categories of data you can watch this video on LinkedIn Learning.* More information on how to log on to LinkedIn Learning as a UF Affiliate is found here

18.3.1. Spreadsheets

Spreadsheets are ok for data entry, but they have some features that make it easy to do terrible, terrible things. People often use spreadsheets for much more, including calculations, statistical analyses, and creating tables or figures for publications and presentations.

After you have entered data in a spreadsheet:

> **Don't do calculations or data correction**
>
> **I implore you.**
>
> **Don't. Please**

There are several reasons why.

(A) The 'drag-and-drop', menu-driven nature of spreadsheet programs makes it very difficult (or impossible) to replicate your steps, much less those of another person. This means you can't easily find where mistakes were made, and if you have to reconstruct an analysis or figure you have to start from from the very beginning This is extremely tedious.

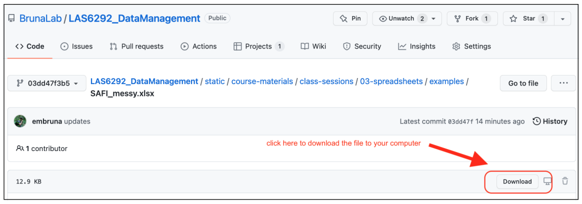
(B) Furthermore, when doing calculations in a spreadsheet it is easy to accidentally apply a slightly different formula to multiple adjacent cells. It is easy to introduce mistakes.

(C) Finally, at some point during your data correction or analyses - probably without even realizing it - you will make a mistake either 'sorting' or trying to fill in cells with 'copy-drag-drop-paste'. This will potentially ruin several days of your life (or more) while you try to fix it (assuming you realize you made this mistake, which people often don't).

18.4. Breakout 1: Group Discussion

Much of your future as a researcher will be spent cleaning and correcting data, but you can reduce the time spent on this task (and the associated stress) considerably by implementing some good practices from the start. To start developing these good habits we will to take a look at some spreadsheets, identify the things that people should ***not*** be doing with them, and then determining what they should be doing instead.

1. Download the following three spreadsheets. To download the files, click the links and then the `download` button (shown below) on the right-hand side.



BrunaLab / LAS6292_DataManagement

LAS6292_DataManagement / static / course-materials / class-sessions / 03-spreadsheets / examples / SAFI_messy.xlsx

embruna updates

Latest commit 03dd47f 14 minutes ago

1 contributor

click here to download the file to your computer

12.9 KB

Figure 18.1.: Download Files by following the link and clicking this button.

- `SAFI_messy.xlsx`: download link.
- `unity-portal-data.xlsx`: download link.
- `dates.xlsx`: {download link](https://github.com/BrunaLab/LAS6292_DataManagement/blob/03dd47f3b52a9bf32be643cf34bafcce6566e555/content/instructor-materials/class-sessions/03-spreadsheets/examples/dates.xlsx)

2. Open `SAFI_messy.xlsx` and look at how the data are have been entered and organized. Now discuss the following questions. Keep in mind the `tidy` principles about which you read in Broman and Woo (2018).

a. What problems can you identify with the way these data are entered/organized?

b. How would you correct each of these issues? Could these data easily be imported into a programming language or a database in its current form?

3. Do the same with `unity-portal-data.xlsx`: review the data and discuss questions a & b.

4. Dates, or things that look like dates, are especially problematic in Excel. Open the file `dates.xlsx` and enter the following dates into the column labeled `date_1`. Be sure to type them in exactly as they are written:

   - `7-2-21`
   - `2 july 2021`
   - `july 2, 2021`
   - `july 2,2021` [no space between the comma and 2021]
   - `07-02-21`
   - `7/2/21`
   - `Jan 5, 1900`
   - `Dec 5, 1899`

   a. Is the value in the cell the same as what you typed in?
   b. Why would these issues be a problem for data organization and analysis?

5. Next enter the dates above into the column labeled `date_2`. Again, be sure to type them in exactly as they are written.

   a. what was different about the way the data are recorded?
   b. can you figure out why?

6. What would you do to enter dates into Excel in a way that avoids the issues observed above?

7. Export the `SAFI_messy.xlsx` as a .csv file with the name `SAFI_messy.csv`; you'll have to click the "OK" when warning box pops up. Now reopen it. What happened? You can find a guide to saving your file in .csv format and why that is a good idea on this website.

18.5. Breakout 1: Returning results

Alternating between groups, guide groups to the following best practices:

- 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.
- Use consistent names, abbreviations/codes, and capitalization.
- Use good null values (not -999, blanks ok, some prefer `NA` or similar but this can be language specific).

- Write dates as YYYYMMDD. Better still have separate columns for Year, Month, and Day.
- 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).
- Avoid using multiple tables within one spreadsheet.
- Avoid spreading data across multiple tabs (but do use a new tab to record data cleaning or manipulations).
- Record zeros as zeros.
- Use an appropriate null value to record missing data.
- Don't use formatting to convey information or to make your spreadsheet look pretty.
- Excel is unable to parse dates from before 1899-12-31. Be careful if your data include a mix of pre/post….you'll have mixed data types.
- 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.
- **NB:** 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 calclulate "days from a given date" for accounting purposes (like invoicing) by adding "date+XX days". * Furthermore, Excel is unable to parse dates from before 1899-12-31. Be careful if your data include a mix of pre/post….you'll have mixed data types.

18.6. Take-home Messages

1. **Once you are done with data entry, save it as 'read only' and make *all* corrections using scripting!**
2. **Entering data in tidy format will make it much easier to analyze.**
3. **Collecting data in tidy format makes it easier to enter data in tidy format.**

18.7. Break

18.8. In-class Group Assignment

The goal of this breakout is to learn some ways to minimize the number of mistakes when entering data. **First**, watch the following video (11 min) on 'Data Validation in Excel'. **Second**, open this web page on 'Quality Assurance and Control in Excel'. It covers the same material, so it's a handy reference to have open during the exercise. (*Note: while* *we are using Excel for this exercise, see "Tools" below for how to do the same in Google Sheets*).

Exercise: Set up a `tidy` sheet for data entry for the Portal data from Breakout 1

1. Create a spreadsheet in Excel for data entry. It should have five columns, in which you will be recording (1) the date of observations, (2) the site in which the observations were conducted, (3) the species captured, (4) the mass of each animal, and (5) the length of each animal.

2. Set the following data validation criteria to prevent invalid data from getting entered:

   a. The Date column should be set so that it does *not* convert dates to other formats.
   b. Use data validation so that Site can only be one of the following A1, A2, B1, B2.
   c. Set the error message on this validation criteria to provide information on what the valid values are.
   d. Use data validation so that Species can only be one of the following: *Dipodomys spectabilis*, *Dipodomys ordii*, *Dipodomys merriami*.
   e. Set the error message on this validation criteria to provide information on what the valid values are.
   f. Use data validation so that Mass can only be a decimal greater than or equal to zero but less than or equal to 500.
   g. Set the error message on this validation criteria to provide information on what the valid values are.
   h. Length should be an integer (i.e., a whole number) between 1 and 10.
   i. Set the error message on this validation criteria to provide information on what the valid values are.

3. Check that the validation rules and data formatting are working by entering some data in the cells

4. Save this file as `data_entry_form.xlsx` and submit it via the Canvas website as 'homework-wk3'.

Grading Rubric:

Assignment completed with data validation correctly programmed with useful error messages: 50 Most data validation properly programmed; some require instructor follow-up: 40 Many of the validation parameters need corrections, error messages not useful: 30 Incorrect data are able to be entered in all categories; Instructor follow-up required: 20

18.9. Time remaining

Any time remaining can be used for:

1. Data Project Hand-out and overview (`r proj_overview` min)

2. Using R to go from ‘dirty’ to ‘clean’ data. (`r cleaning_demo` min). A live coding example to lay the foundation for what we will be doing moving forward, and how scripting makes it easy to go from dirty to clean data in a reproducible fashion.

3. Going over R installations, meeting with students about their data sets for the semester projects, etc.

18.10. Sources for this lesson

1. 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

2. 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

3. Philip Woodhouse, Gert Jan Veldwisch, Daniel Brockington, Hans C. Komakech, Angela Manjichi, Jean-Philippe Venot. 2018. SAFI Survey Results. https://figshare.com/articles/dataset/SAFI_Survey_Results/6262019 doi:10.6084/m9.figshare.6262019.v4

4. Chris Prener, Trevor Burrows (Eds.). Data Carpentry: Data Organization in Spreadsheets for Social Scientists. https://datacarpentry.org/spreadsheets-socialsci/

5. 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

6. Ernest, Morgan; Brown, James; Valone, Thomas; White, Ethan P. (2017): Portal Project Teaching Database. figshare. https://doi.org/10.6084/m9.figshare.1314459.v6

19. Data Organization in Spreadsheets: Instructor Notes

Objectives and Competencies

- Be able to identify different categories of data
- Learn best practices for data entry
- Recognize and avoid common problems with data entry and formatting in spreadsheets
- Learn and be able to implement 'Tidy' format for data tables in spreadsheets
- Identify problems with and approaches for proper handling of dates in spreadsheets
- Learn how to export data from spreadsheets in open format

19.1. Pre-Class Preparation (Instructor):

Send in an email to students:

- Remind everyone about the computer with spreadsheet software prior to class

Bring to Class:

- Snacks
- Flip charts and markers
- Dry write markers
- Tent cards for student names

19.2. Pre-class Preparation (Students):

Online Lectures: None

Readings

1. Tesi, W. 2020. An Outdated Version of Excel Led the U.K. to Undercount COVID-19 Cases. Slate. [read online]
2. Stolberg et al. 2020. CDC Test Counting Error Leaves Epidemiologists 'Really Baffled'. NY Times.
   [read online]

3. Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. The American Statistician, 72(1), 2-10. [read online]

4. Johnson, B. D., Dunlap, E., & Benoit, E. (2010). Organizing "mountains of words" for data analysis, both qualitative and quantitative. Substance Use & Misuse, 45(5), 648-70. [read online]

Computer Resources

1. **Please bring your computer**. If you need to borrow a laptop or get access to a computer let me know.

2. You will need a **spreadsheet program** such as MS Excel or LibreOffice installed on your computer. You can use the program of your choice, but the exercises are optimized for MS Excel. UF students can download the Microsoft Office suite free of charge here; if you don't want to install Excel you can use the online version for free (see this tutorial video on how to do so). LibreOffice is a free and open source package similar to (and compatible with) MS Office. It can be downloaded here.

3. The paper by Broman and Woo on how to organize data in spreadsheets is *especially* important; it may well be one of the more helpful papers you read while a student. Really.

4. It is also time to install two other pieces of software. We won't use them until Week 4, but it is worth installing them now to make sure they are working smoothly:

   (i) **the R programming language**: You can download the version of R for your computer operating system here; it's free.

   - ***For Mac computers*** click on 'Download R for (Mac) OS X', then chose the version of R for your operating system, i.e., macOS 10.13 (High Sierra), OS X 10.11 (El Capitan), Mac OS X 10.9 (Mavericks), etc.

   - ***For PC computers***: click on Download R for Windows, then click 'base' or 'install R for the first time'.

   - ***For Linux***: click on 'Download R for Linux'; after which you are on your own. Then again you use Linux, so you probably don't need my help anyway.

   - ***If you need help:*** watch this video tutorial or contact me. *Note that the tutorial requires you be a UF affiliate and either on campus or logged to the UF network via VPN.*

   (ii) **RStudio**: the software used to work with R. There are other 'environments' one can use for R programming, but RStudio is by far the most widely used and useful.

   - We use the 'Free Open-Source Desktop Version', which you can download here. Choose the version for your computer operating system and install as you would any other software.

- **If you need help**: watch this video tutorial or contact me. *Note that the tutorial requires you be a UF affiliate and either on campus or logged to the UF network via VPN.*

(iii) **Verify the installations** worked by opening RStudio to see if it opens properly. If you are really motivated, you can also install the `Tidyverse` library by starting RStudio and at the console typing `install.packages("tidyverse")`.

19.3. Session Introduction

Types of Data

Down the road when doing data correction, organizing data, and doing analyses it will be essential to classify data according to their 'type'. It can also help with data entry, which is why we will introduce some of these types here:

1. **Nominal aka Factor:** categories or groups, such as [`apple`, `orange`], [`trumpet`, `flute`, `violin`]
2. **Ordinal aka Ordered Factor:** groups where there is an order: [`first>second>third`], [`small<medium<large`]. Note that this order doesn't imply quantitative value, e.g., One is not stating that `medium` is twice the size of `small` or that `large` is twice the size of `medium`.
3. **Character**: [`a`, `gnv`, `mexico`, `Inigo Montoya`]
4. **Numeric (real or decimal):** `2`, `15.5`
5. **Integer:** [`1`,`2`,`3`]
6. **Logical:** [`True`, `False`]
7. **Complex:** `1+4*i`
8. **Interesting case:** what category are [`red`, `orange`, `green`, `blue`]? We usually treat it as Nominal, but it is actually Ordinal - the colors represent wavelengths on the visible light spectrum (650, 600, 550, and 450 nm respectively). If you were recording the wavelength itself, it would be Numeric.
9. *For more on different categories of data you can watch this video on LinkedIn Learning.* More information on how to log on to LinkedIn Learning as a UF Affiliate is found here

19.3.1. Spreadsheets

Spreadsheets are ok for data entry, but they have some features that make it easy to do terrible, terrible things. People often use spreadsheets for much more, including calculations, statistical analyses, and creating tables or figures for publications and presentations.

After you have entered data in a spreadsheet:

> **Don't do calculations or data correction**
>
> **I implore you.**
>
> **Don't. Please**

There are several reasons why.

(A) The ‘drag-and-drop’, menu-driven nature of spreadsheet programs makes it very difficult (or impossible) to replicate your steps, much less those of another person. This means you can’t easily find where mistakes were made, and if you have to reconstruct an analysis or figure you have to start from from the very beginning This is extremely tedious.

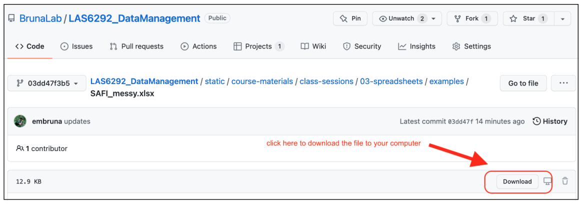
(B) Furthermore, when doing calculations in a spreadsheet it is easy to accidentally apply a slightly different formula to multiple adjacent cells. It is easy to introduce mistakes.

(C) Finally, at some point during your data correction or analyses - probably without even realizing it - you will make a mistake either ‘sorting’ or trying to fill in cells with ‘copy-drag-drop-paste’. This will potentially ruin several days of your life (or more) while you try to fix it (assuming you realize you made this mistake, which people often don’t).

19.4. Breakout 1: Group Discussion

Much of your future as a researcher will be spent cleaning and correcting data, but you can reduce the time spent on this task (and the associated stress) considerably by implementing some good practices from the start. To start developing these good habits we will to take a look at some spreadsheets, identify the things that people should ***not*** be doing with them, and then determining what they should be doing instead.

1. Download the following three spreadsheets. To download the files, click the links and then the `download` button (shown below) on the right-hand side.



BrunaLab / LAS6292_DataManagement

LAS6292_DataManagement / static / course-materials / class-sessions / 03-spreadsheets / examples / SAFI_messy.xlsx

embruna updates

Latest commit 03dd47f 14 minutes ago

1 contributor

click here to download the file to your computer

12.9 KB

Figure 19.1.: Download Files by following the link and clicking this button.

- `SAFI_messy.xlsx`: download link.
- `unity-portal-data.xlsx`: download link.
- `dates.xlsx`: {download link](https://github.com/BrunaLab/LAS6292_DataManagement/blob/03dd47f3b52a9bf32be643cf34bafcce6566e555/content/instructor-materials/class-sessions/03-spreadsheets/examples/dates.xlsx)

2. Open `SAFI_messy.xlsx` and look at how the data are have been entered and organized. Now discuss the following questions. Keep in mind the `tidy` principles about which you read in Broman and Woo (2018).

a. What problems can you identify with the way these data are entered/organized?

b. How would you correct each of these issues? Could these data easily be imported into a programming language or a database in its current form?

3. Do the same with `unity-portal-data.xlsx`: review the data and discuss questions a & b.
4. Dates, or things that look like dates, are especially problematic in Excel. Open the file `dates.xlsx` and enter the following dates into the column labeled `date_1`. Be sure to type them in exactly as they are written:

   - `7-2-21`
   - `2 july 2021`
   - `july 2, 2021`
   - `july 2,2021` [no space between the comma and 2021]
   - `07-02-21`
   - `7/2/21`
   - `Jan 5, 1900`
   - `Dec 5, 1899`

   a. Is the value in the cell the same as what you typed in?
   b. Why would these issues be a problem for data organization and analysis?

5. Next enter the dates above into the column labeled `date_2`. Again, be sure to type them in exactly as they are written.

   a. what was different about the way the data are recorded?
   b. can you figure out why?

6. What would you do to enter dates into Excel in a way that avoids the issues observed above?
7. Export the `SAFI_messy.xlsx` as a .csv file with the name `SAFI_messy.csv`; you'll have to click the "OK" when warning box pops up. Now reopen it. What happened? You can find a guide to saving your file in .csv format and why that is a good idea on this website.

19.5. Breakout 1: Returning results

Alternating between groups, guide groups to the following best practices:

- 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.
- Use consistent names, abbreviations/codes, and capitalization.
- Use good null values (not -999, blanks ok, some prefer `NA` or similar but this can be language specific).

- Write dates as YYYYMMDD. Better still have separate columns for Year, Month, and Day.
- 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).
- Avoid using multiple tables within one spreadsheet.
- Avoid spreading data across multiple tabs (but do use a new tab to record data cleaning or manipulations).
- Record zeros as zeros.
- Use an appropriate null value to record missing data.
- Don't use formatting to convey information or to make your spreadsheet look pretty.
- Excel is unable to parse dates from before 1899-12-31. Be careful if your data include a mix of pre/post….you'll have mixed data types.
- 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.
- **NB:** 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 calclulate "days from a given date" for accounting purposes (like invoicing) by adding "date+XX days". * Furthermore, Excel is unable to parse dates from before 1899-12-31. Be careful if your data include a mix of pre/post….you'll have mixed data types.

19.6. Take-home Messages

1. **Once you are done with data entry, save it as 'read only' and make *all* corrections using scripting!**
2. **Entering data in tidy format will make it much easier to analyze.**
3. **Collecting data in tidy format makes it easier to enter data in tidy format.**

19.7. Break

19.8. In-class Group Assignment

The goal of this breakout is to learn some ways to minimize the number of mistakes when entering data. **First**, watch the following video (11 min) on 'Data Validation in Excel'. **Second**, open this web page on 'Quality Assurance and Control in Excel'. It covers the same material, so it's a handy reference to have open during the exercise. (*Note: while* *we are using Excel for this exercise, see "Tools" below for how to do the same in Google Sheets*).

Exercise: Set up a `tidy` sheet for data entry for the Portal data from Breakout 1

1. Create a spreadsheet in Excel for data entry. It should have five columns, in which you will be recording (1) the date of observations, (2) the site in which the observations were conducted, (3) the species captured, (4) the mass of each animal, and (5) the length of each animal.

2. Set the following data validation criteria to prevent invalid data from getting entered:

    a. The Date column should be set so that it does *not* convert dates to other formats.
    b. Use data validation so that Site can only be one of the following A1, A2, B1, B2.
    c. Set the error message on this validation criteria to provide information on what the valid values are.
    d. Use data validation so that Species can only be one of the following: *Dipodomys spectabilis*, *Dipodomys ordii*, *Dipodomys merriami.*
    e. Set the error message on this validation criteria to provide information on what the valid values are.
    f. Use data validation so that Mass can only be a decimal greater than or equal to zero but less than or equal to 500.
    g. Set the error message on this validation criteria to provide information on what the valid values are.
    h. Length should be an integer (i.e., a whole number) between 1 and 10.
    i. Set the error message on this validation criteria to provide information on what the valid values are.

3. Check that the validation rules and data formatting are working by entering some data in the cells

4. Save this file as `data_entry_form.xlsx` and submit it via the Canvas website as 'homework-wk3'.

Grading Rubric:

Assignment completed with data validation correctly programmed with useful error messages: 50 Most data validation properly programmed; some require instructor follow-up: 40 Many of the validation parameters need corrections, error messages not useful: 30 Incorrect data are able to be entered in all categories; Instructor follow-up required: 20

19.9. Time remaining

Any time remaining can be used for:

1. Data Project Hand-out and overview (`r proj_overview` min)

2. Using R to go from 'dirty' to 'clean' data. (`r cleaning_demo` min). A live coding example to lay the foundation for what we will be doing moving forward, and how scripting makes it easy to go from dirty to clean data in a reproducible fashion.

3. Going over R installations, meeting with students about their data sets for the semester projects, etc.

19.10. Sources for this lesson

1. 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

2. 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

3. Philip Woodhouse, Gert Jan Veldwisch, Daniel Brockington, Hans C. Komakech, Angela Manjichi, Jean-Philippe Venot. 2018. SAFI Survey Results. https://figshare.com/articles/dataset/SAFI_Survey_Results/6262019 doi:10.6084/m9.figshare.6262019.v4

4. Chris Prener, Trevor Burrows (Eds.). Data Carpentry: Data Organization in Spreadsheets for Social Scientists. https://datacarpentry.org/spreadsheets-socialsci/

5. 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

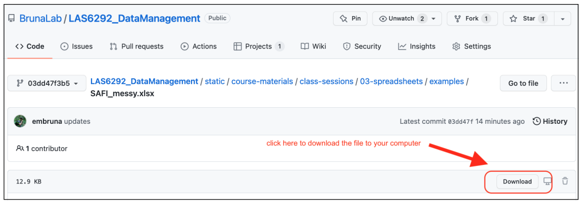
6. Ernest, Morgan; Brown, James; Valone, Thomas; White, Ethan P. (2017): Portal Project Teaching Database. figshare. https://doi.org/10.6084/m9.figshare.1314459.v6

Data Organization in Spreadsheets: In-class Activities

Breakout 1: Group Discussion

Much of your future as a researcher will be spent cleaning and correcting data, but you can reduce the time spent on this task (and the associated stress) considerably by implementing some good practices from the start. To start developing these good habits we will to take a look at some spreadsheets, identify the things that people should ***not*** be doing with them, and then determining what they should be doing instead.

1. Download the following three spreadsheets. To download the files, click the links and then the `download` button (shown below) on the right-hand side.



LAS6292_DataManagement / static / course-materials / class-sessions / 03-spreadsheets / examples / SAFI_messy.xlsx

embruna updates

Latest commit 03dd47f 14 minutes ago

click here to download the file to your computer

12.9 KB

Figure 19.2.: Download Files by following the link and clicking this button.

- `SAFI_messy.xlsx`: download link.
- `untidy-portal-data.xlsx`: download link.
- `dates.xlsx`: {download link](https://github.com/BrunaLab/LAS6292_DataCourse-Book/blob/4aebeb7b92f15b59655cb47e7104c1c83df11887/class_materials/class_sessions/03_spreadsheets/examples/dates.xlsx)

2. Open `SAFI_messy.xlsx` and look at how the data are have been entered and organized. Now discuss the following questions. Keep in mind the `tidy` principles about which you read in Broman and Woo (2018).

a. What problems can you identify with the way these data are entered/organized?

b. How would you correct each of these issues? Could these data easily be imported into a programming language or a database in its current form?

3. Do the same with `unity-portal-data.xlsx`: review the data and discuss questions a & b.

4. Dates, or things that look like dates, are especially problematic in Excel. Open the file `dates.xlsx` and enter the following dates into the column labeled `date_1`. Be sure to type them in exactly as they are written:

   - `7-2-21`
   - `2 july 2021`
   - `july 2, 2021`
   - `july 2,2021` [no space between the comma and 2021]
   - `07-02-21`
   - `7/2/21`
   - `Jan 5, 1900`
   - `Dec 5, 1899`

   a. Is the value in the cell the same as what you typed in?
   b. Why would these issues be a problem for data organization and analysis?

5. Next enter the dates above into the column labeled `date_2`. Again, be sure to type them in exactly as they are written.

   a. what was different about the way the data are recorded?
   b. can you figure out why?

6. What would you do to enter dates into Excel in a way that avoids the issues observed above?

7. Export the `SAFI_messy.xlsx` as a .csv file with the name `SAFI_messy.csv`; you'll have to click the "OK" when warning box pops up. Now reopen it. What happened? You can find a guide to saving your file in .csv format and why that is a good idea on this website.

Breakout 1: Returning results

Alternating between groups, guide groups to the following best practices:

- 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.
- Use consistent names, abbreviations/codes, and capitalization.
- Use good null values (not -999, blanks ok, some prefer `NA` or similar but this can be language specific).
- Write dates as YYYYMMDD. Better still have separate columns for Year, Month, and Day.

- 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).
- Avoid using multiple tables within one spreadsheet.
- Avoid spreading data across multiple tabs (but do use a new tab to record data cleaning or manipulations).
- Record zeros as zeros.
- Use an appropriate null value to record missing data.
- Don't use formatting to convey information or to make your spreadsheet look pretty.
- Excel is unable to parse dates from before 1899-12-31. Be careful if your data include a mix of pre/post….you'll have mixed data types.
- 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.
- **NB:** 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 calclulate "days from a given date" for accounting purposes (like invoicing) by adding "date+XX days". * Furthermore, Excel is unable to parse dates from before 1899-12-31. Be careful if your data include a mix of pre/post….you'll have mixed data types.

Take-home Messages

1. **Once you are done with data entry, save it as 'read only' and make *all* corrections using scripting!**
2. **Entering data in tidy format will make it much easier to analyze.**
3. **Collecting data in tidy format makes it easier to enter data in tidy format.**

In-class Group Assignment

The goal of this breakout is to learn some ways to minimize the number of mistakes when entering data. **First**, watch the following video (11 min) on 'Data Validation in Excel'. **Second**, open this web page on 'Quality Assurance and Control in Excel'. It covers the same material, so it's a handy reference to have open during the exercise. (*Note: while we are using Excel for this exercise, see "Tools" below for how to do the same in Google Sheets*).

Exercise: Set up a `tidy` sheet for data entry for the Portal data from Breakout 1

1. Create a spreadsheet in Excel for data entry. It should have five columns, in which you will be recording (1) the date of observations, (2) the site in which the observations were conducted, (3) the species captured, (4) the mass of each animal, and (5) the length of each animal.

2. Set the following data validation criteria to prevent invalid data from getting entered:

   a. The Date column should be set so that it does *not* convert dates to other formats.
   b. Use data validation so that Site can only be one of the following A1, A2, B1, B2.
   c. Set the error message on this validation criteria to provide information on what the valid values are.
   d. Use data validation so that Species can only be one of the following: *Dipodomys spectabilis*, *Dipodomys ordii*, *Dipodomys merriami*.
   e. Set the error message on this validation criteria to provide information on what the valid values are.
   f. Use data validation so that Mass can only be a decimal greater than or equal to zero but less than or equal to 500.
   g. Set the error message on this validation criteria to provide information on what the valid values are.
   h. Length should be an integer (i.e., a whole number) between 1 and 10.
   i. Set the error message on this validation criteria to provide information on what the valid values are.

3. Check that the validation rules and data formatting are working by entering some data in the cells

4. Save this file as `data_entry_form.xlsx` and submit it via the Canvas website as 'homework-wk3'.

Grading Rubric:

Assignment completed with data validation correctly programmed with useful error messages: 35 Most data validation properly programmed; some require instructor follow-up: 30 Many of the validation parameters need corrections, error messages not useful: 25 Incorrect data are able to be entered in all categories; Instructor follow-up required: 20

20. Reproducible Data Correction with R

20.1. Download the data we will be using in class

1. Open the messy data file `demo_data.csv` by following this link
2. the data will open as a tab in your web browser in `.csv` format; save them to the `data_raw` folder by going to 'File' on the menu bar of your web browser and selecting 'Save page as' from the drop-down menu.
3. save the file to the `data_raw` folder.

20.2. Data Cleaning: Practice

1. Review the `.csv` file

2. What things do you see that need to be corrected?

3. Make a list of the what you think needs to be corrected and the steps necessary to identify and implement each correction. Some of the things to look out for include:

   - Numeric values stored as character data types
   - Factors stred as characters
   - Duplicate rows
   - Spelling mistakes
   - inconsistent formatting (eg., codes, capitalizations)
   - White spaces
   - Missing data
   - Zeros instead of null values
   - Special characters (e.g. commas in numeric values instead of decimals)
   - column headings with spaces between words or that start with numerals

4. Write R code to implement the changes you have identified

5. Save this code as `las_practice.r` and submit it via the Canvas website.

Grading Rubric:

Assignment completed with data validation correctly programmed with useful error messages: 35 Most data validation properly programmed; some require instructor follow-up: 25 Many of the validation parameters need corrections, error messages not useful: 15 Incorrect data are able to be entered in all categories; Instructor follow-up required: 10

21. QA/QC 1: In-class Activities

21.1. Data Entry with Speech-to-Text

MS Excel

1. ‘Speak Cells’ Tutorial Video. This will allow you to select rows or columns of Excel and have them read back to you. *Alternative video: it’s not as thorough, but it is a bit easier to see the menu on this video speak cells on enter Excel https://youtu.be/DSltR90mkgc
2. ‘Speak Cells on Enter’: This function will read back what you entered in a cell when you hit “enter”. You can set it up using this tutorial video.
3. Mac settings are a bit different. Can you figure them out?

Google Docs

If you prefer working in Google Docs you can do the same thing. This article will show you how. You can also watch this video.

21.2. In-Class Activity

1. Enter 10 rows of data in excel that are typical of the different kinds of data you might be collecting (be sure to give each column a properly formatted name). Try the following:
   - strings (e.g., names of towns or people)
   - numbers
   - codes that are both letters (‘AB’) and combinations of rows and numbers (‘A4’, ‘B5’).
   - leave at least one blank cells in each column.
2. Set up “Speak-Cells” in Excel and have it read back the columns to you.
3. Summarize the results: did it read the data back accurately? Were you able to understand what it read? How did it deal with the blank cells?
4. Now set up ‘Speak Cells on Enter’. Enter some of the data you entered above in some new cells. Now have your partner read some of their data to you.

5. Finally, try entering the text below into a cell in your excel document - the way you would normally - and have Excel read the text back to you. Did it catch any mistakes?

   > “I closed the door and put the shop key in its usual place behind Francisco’s Book on Advanced Principles of Astronomy. Poor Chico. No one has wanted his fat gray book for thirty years. Sometimes I wonder why he keeps teaching even if his students all like other subjects. His mom died when he was very young and moved to Mexico City. He thinks it was his destiny to be a soccer player.”

What worked best? What didn’t work? Mac users: Why doesn’t mac read numbers in columns? **Submit the brief answers to these questions via canvas as HW Week 5_**

22. Class Outline: QA/QC 2 - Open Refine

Objectives and Competencies

By the end of this lesson students will:

- Be able to import a data set into OpenRefine, make changes to the data set and its structure, and export the revised data set
- Learn how to automatically track changes made and export the record of changes
- Be able to apply these changes to a different data set

22.1. Pre-Class Preparation (Instructor):

- Remind via email about OpenRefine Installation
- Post Data sets

Bring to Class:

- Snacks
- Tent cards for student names

22.2. Pre-class Preparation (Students):

22.2.1. Online Lectures: None

22.2.2. Readings: None

22.2.3. Computer Resources

1. Install OpenRefine on your computer and verify it works by following the instructions here.
2. *Optional:* Read and watch about how OpenRefine works here. You can also review the basic workflow we will learn.

22.3. In-Class: Using Open Refine to clean data

OpenRefine is a powerful, free, and open source tool that is used to work with and clean messy data. We will be working through some of OpenRefine's basic features, after which you will trying them onm your own on a new data set.

This session uses the Data Carpentry Lesson *OpenRefine for Social Science Data* to introducee students to how OpenRefine can be used to clean datasets. The lesson's objectives are:

1. To describe OpenRefine's uses and applications,
2. Differentiate data cleaning from data organization, and
3. Experiment with OpenRefine's user interface.

22.3.1. Key Points to empghasize:

- OpenRefine provides a set of tools to identify and correct messy data.
- All actions are easily reversed
- If you save your work it will be to a new file. OpenRefine always uses a copy of your data and does not modify your original dataset
- OpenRefine keeps track of all of your actions and allows them to be applied to different datasets
- It is open source with a large user community, making it easy to get
- It works very well with 'large-ish' datasets (100,000 rows), but can adjust memory allocation to accommodate larger datasets.
- OpenRefine always keeps your data private on your own computer until you choose to share it. It works by running a small server on your computer and using your web browser to interact with it. *Your private data never leaves your computer unless you want it to*

22.3.2. Installation and Setup

- Follow the Setup instructions to install OpenRefine.
- If after installation and running OpenRefine, it does not automatically open for you, point your browser at http://127.0.0.1:3333/ or http://localhost:3333 to launch the program.
- The data for this lesson is a part of the 'Data Carpentry Social Sciences' workshop. It is a teaching version of the Studying African Farmer-Led Irrigation (SAFI) database: interviews of farmers in two countries in eastern sub-Saharan Africa (Mozambique and Tanzania). These interviews were conducted between November 2016 and June 2017 and probed household features (e.g. construction materials used, number of household members), agricultural practices (e.g. water usage), and assets (e.g. number and types of livestock). *The data used in this lesson is a subset of the teaching version that has been intentionally 'messed up' for this lesson.*
- Download the data file to your computer to a location where you will be able to find it during the lesson.

22.3.3. Part 1

- Intro to OR
- Working with OR
  -Filtering and Sorting

22.3.4. Break

22.3.5. Part 2

- Examining Numbers
- Using Scripts, Exporting, and Saving
- Wrap-up, Questions

22.3.6. Assignment

Now it's your turn. Download this csv file and use OpenRefine to clean it up. After you create a Project, edit the data as follows:

1. Correct and standardize the names of the countries in which the rodents were captured.
2. The column `scientificName` contains two pieces of information (the genus *and* species of each animal). Split this into two columns, rename them as `genus` and `species`, and then correct and standardize the data in each column as needed. NB: You may run into an obstacle when you try to rename the columns. How can you get around it?
3. Save the clean data as a `.csv` file on your desktop.
4. Extract and save your steps (i.e., 'operation history' as JSON. Save this as a text file.
5. *Bonus Brainteaser:* Many of the cells in the column for the Latin bonomial are blank. How might you go about filling them in based on the column with the abbreviation?
6. **Submission:** Submit your clean `.csv` and the `.txt` file of JSON output on Canvas.

22.3.7. Grading Rubric:

- Data corrected and JSON file can be used on another data set: 35
- Most data correction properly programmed; some require instructor follow-up: 25
- Many of the corrections missing, JSON file unable to process new data : 20
- Instructor follow-up required to implement most changes: 15

The materials in this lesson are derived from the Data Carpentry Social Sciences workshop by The Carpentries.

Part III.

Individual Projects

23. Assignments Overview

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

23.1. 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*.

23.2. 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

23.3. 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.

Useful Resources

24. Useful resources for Data Collection & Managment

24.1. R Programming

Essential Resources

1. ***R for Data Science* book:** 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. **RStudio Keyboard Shortcuts**: A list of the keyboard shortcuts for Mac, Windows, and Linux can be found here.
4. **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.

Tutorials

1. Paul van der Laken's List of books, tutorials, and other resources on topics ranging from data manipulation to data validation to data visualization.
2. R Essential Training: Wrangling and Visualizing Data. (requires a UF email address to access LinkedIn Learning).
3. Software Carpentry: Using RStudio for Project Organization & Management
4. Swirl: learn R programming interactively, at your own pace, and in the R console.
5. R Bootcamp by Ted Laderas and Jessica Minnier. Learn R in your browser.
6. How to clean messy data in R

7. The Ultimate Guide to Data Cleaning is written for Python users but the pricniples apply regardless of language.

24.2. Specific Data Cleaning and Management Problems

Dates & Times

1. **Handling dates and times in R**

Text & Text Mining

2. **Text Mining:** `tidytext` package

Qualtrics

3. **Working with Qualtrics** survey data: `qualtRics` package

Text Extraction

4. **Optical Character Recognition (OCR):** extract text from images: `tesseract` package
5. **Extract text & metadata from pdf files:** `pdftools` package

Images & Image Processing

6. **Image processing**: the `magick` package

24.3. Advanced R Packages for Data Management

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. ROpenSci: tools for accessing, manipulating, and visualizing open data

24.4. Data Visualization

1. *Data Visualization: a practical introduction* by Kieran Healy 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.

24.5. Slide & Presentations

1. Make slide presentations with R

24.6. Documents & Reports

1. `knitr` overview: reproducible documents with R

24.7. Discipline-specific R Resources

History

1. `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.
2. *The Programming Historian* Website: wide range of topics, from text analysis to OpenRefine

Psychology

1. 'Programming for Psychologists: Data Creation and Analysis' by Matthew J. C. Crump

24.8. Data Archives

1. Qualitative Data Repository: dedicated archive for storing and sharing digital data (and accompanying documentation) generated or collected through qualitative and multi-method research in the social sciences and related disciplines.
2. Data Dryad: open data publishing platform and a community committed to the open availability and routine re-use of all research data.
3. ICPSR: data access, curation, and analytical methods for social science.

25. Class Slides

The following slides are available in HTML format You can export them as PDFs by clicking on the three bars in the bottom left corner and selecting “PDF Export Mode” under the “Tools” Option.

1. Introduction
2. Data Organization and Backups
3. ‘Reproducible’ Data Management
4. QA/QC 1 - Error Minimization