Curating Chaos - Session Outlines

Emilio M. Bruna

1/7/2021

4. Reproducible data (re)organization

count the steps exercise - make a figure or do an analysis, change a variable, and count the number of steps involved.

- **Intro Material (lecture/videos):**
  - https://dataoneorg.github.io/Education/lessons/09_analysis/index
  - https://dataoneorg.github.io/Education/bp_step/analyze/
  - https://dataoneorg.github.io/Education/bp_step/collect/
  - https://dataoneorg.github.io/Education/bp_step/integrate/
- **In-class Exercise:** Intro to R, Markdown, show / do rearranging of spreadsheets from last week, wide-to-long
  - Data Carpentry R Intros note: *depends on familiarity of students with R*
  - https://datacarpentry.org/r-socialsci/
  - https://datacarpentry.org/R-ecology-lesson/

5. Data validation & correction 1

- **Intro Material (lecture/videos):**

QA/QC: finding and fixing mistakes. Best way is not to make them.

MAKE A VIDEO ABOUT THIS Finding and fixing mistakes: Before R: have it read the data back to you immediately.

A. Reading back https://www.dataquest.io/blog/load-clean-data-r-tidyverse/ INSIDE OF R STUDIO https://rpubs.com/williamsurles/291107

B. Plotting different ways

C. Summary calculations

D. Randomly sampling from digital file to verify find spelling mistakes in factors; add new ones if needed find and remove duplicate rows correct data types find outliers find and replace strings split columns find missing data remove white space at start and end

E. Open Refine (next time)

F. MOST IMPORTANT: DO NOT FIX RAW DATA FILE. Txt file with changes that need to be made, then make them with a scripting language

- **In-class Exercise:**
  - visualizations to help find errors, working on own datasets

6. Data validation & correction 2

- **Intro Material (lecture/videos):**
  - review of all that needs to be done: https://dataoneorg.github.io/Education/bp_step/assure/
- **In-class Exercise:**
  - Data Carpentry Lesson (2:15 h)
  - https://datacarpentry.org/openrefine-socialsci/
  - https://datacarpentry.org/OpenRefine-ecology-lesson/

7. Documentation: Metadata, Codebooks

https://www.youtube.com/watch?v=N2zK3sAtr-4

- Understands the concept of, and rationale for, metadata
- Identify and list the types of information typically included in metadata records
- Identify applicable standards for documenting and capturing metadata in your discipline
- Familiar with tools for creating metadata appropriate for your discipline and data type
- Formulate an approach to creating metadata for a project
- Develops an ability to read and interpret metadata from external disciplinary sources

it is hard to explain to someone how to do stuff, what things mean example exercise of the importance of detailed instructions: definitions

build-a-sandwhich reporducibility exercise (they do instructions, i will make on screen trying to follow! don't tell them I'm going to do it, ask them to do it and then spring it on them),

- **Intro Material (lecture/videos):**
  - https://dataoneorg.github.io/Education/lessons/07_metadata/index
  - https://dataoneorg.github.io/Education/bp_step/describe/
- **In-class Exercise:** Begin Drafting Metadata File, Code Books

8. Data Management Plans

- **Intro Material (lecture/videos):**
  - https://dataoneorg.github.io/Education/lessons/03_planning/index
  - https://dataoneorg.github.io/Education/bp_step/integrate/
- **In-class Exercise:** DMP tool to start buildling their own DMP

9. Efficient data collection

- **In-class Exercise:** build a better form, find error minimization

10. Transcription & Translation

- **In-class Exercise:** compare different transcription / translation tools

11. 'Paperless' data collection

- **In-class Exercise:** build epicollect file and collect data withit

12. Automated data extraction

- **In-class Exercise:** Compare OCR tools

13. Legal and Ethical Issues

- Explain ownership considerations related to data sharing

- Explain and evaluate potential legal issues connected to your data; intellectual property, copyright claims, licenses needed for use, monetary charges for data
- Explain ethical considerations related to data sharing
- Understand privacy levels for research data as required by potential funding agencies
- Recognize the importance of privacy with some forms of research data (HIPAA)
- Understand the importance of removing key personal identifiers to facilitate confidentiality
- **Intro Material (lecture/videos):** https://dataoneorg.github.io/Education/lessons/10_policy/index

14. Data Sharing, Reuse, & Archives

understand advantages of data sharing and archiving discuss concerns and and obstacles related to data sharing *understand restrictions on reuse of data, software, and instrumentation, including different CCO and Public Domain licences)* adress reuse / sgharing requiorements from granting aqgencies or sponsors

understand options for maximizing data reuse understand data identifiers and what they are for * underrstand benefits of using unique researcher id in metadata

- Identify types of available repositories/archives (discipline-based, institutional, etc.)
- Chooses appropriate option for long-term storage of data
- Understand process issues for depositing data in repository
- **Intro Material (lecture/videos):**
  - https://dataoneorg.github.io/Education/lessons/02_datasharing/L02_Exercise.pdf
  - https://dataoneorg.github.io/Education/lessons/08_citation/index
  - https://dataoneorg.github.io/Education/bp_step/discover/
- **In-class Exercise:** Find data, look at data archive and see if they can understand it