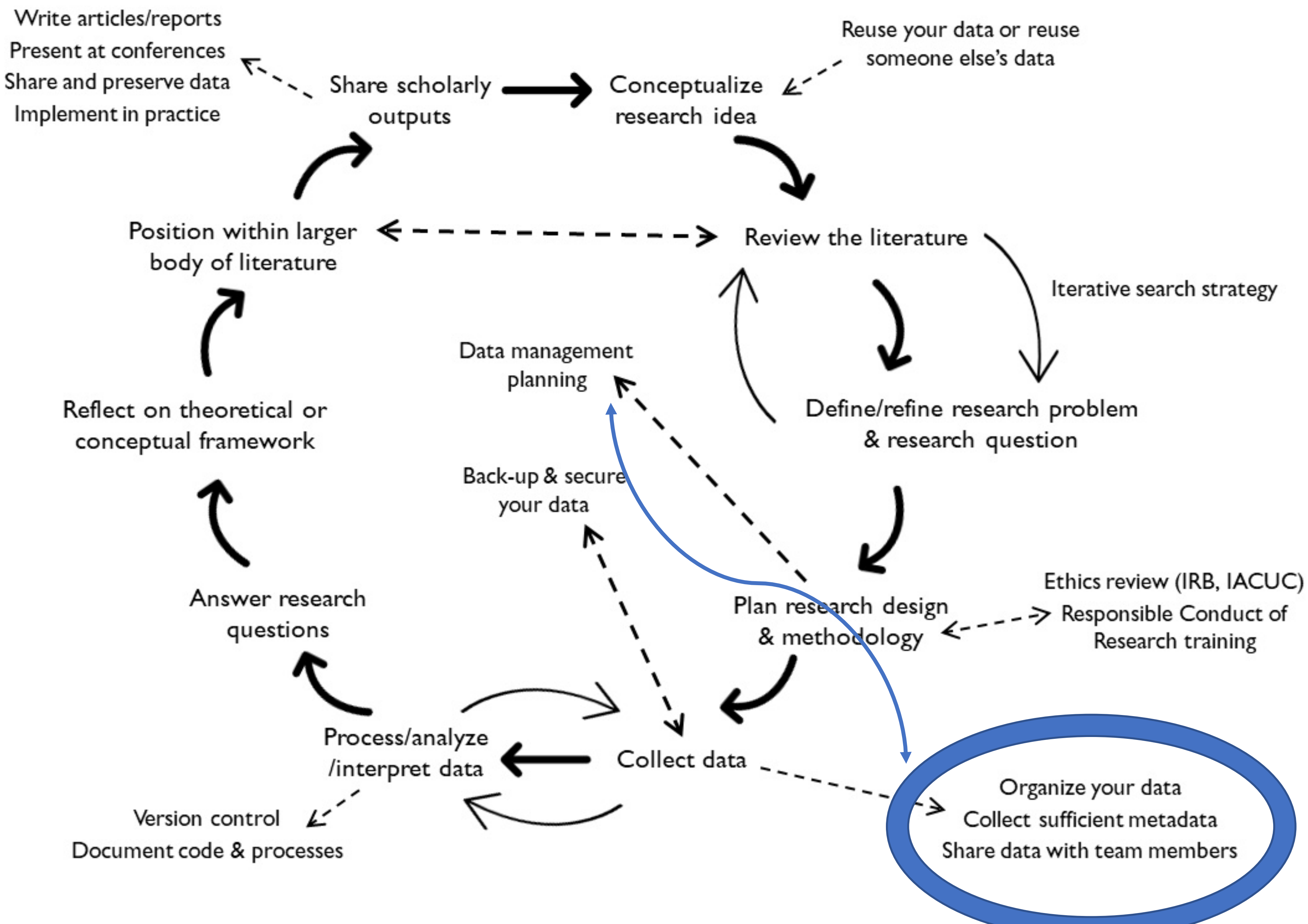
Data Organization & Documentation

DATA 3101

Elizabeth Stregger

October 6, 2022



Write articles/reports
Present at conferences
Share and preserve data
Implement in practice

Share scholarly outputs

Conceptualize research idea

Reuse your data or reuse someone else's data

Review the literature

Iterative search strategy

Define/refine research problem & research question

Plan research design & methodology

Ethics review (IRB, IACUC)
Responsible Conduct of Research training

Data management planning

Back-up & secure your data

Collect data

Organize your data
Collect sufficient metadata
Share data with team members

Process/analyze /interpret data

Version control
Document code & processes

Answer research questions

Reflect on theoretical or conceptual framework

Position within larger body of literature

Organization

- Organizing your data makes it possible for you (and others) to find it
- Organization systems must be logical and based on an understanding of your research tasks and workflow
    - How do you usually search for data (by date? experiment?)
    - Primary arrangement by project, by researcher, by date, by research notebook number, by sample number, by experiment type / instrument, by data type
- Have one main location for all your data (and then back it up!)
- If you have both physical and digital collections:
    - Use same arrangement, or
    - Create an index

Organization: Collaboration

- If possible, agree on shared storage platforms and use the same system to organize your files, name files, control versions
- If not possible, document organizational systems so that others understand it
- Organizing research literature:
  - File naming system: use as much citation information as possible
    - Name_Date_Title
  - Use citation management software for collaborative writing

Organization: File Naming

- File name should include enough information to uniquely identify what is in the file, such as: experiment type, experiment number, researcher name or initials, sample type, sample number, analysis type, date, site name, version number
- Principles:
    - Names should be descriptive
    - Names should be consistent
    - Names should be short, preferably less than 25 characters
    - Use underscores or dashes instead of spaces
    - Avoid special characters
    - Follow the date convention YYYYMMDD

Organization: File naming tips

- Replace automatically generated file names, such as digital camera file names, which may otherwise be overwritten
- Consider scalability
- You may need more than one file naming convention for different types of data
- Use camel case (capitalize the first letter of each word) or pothole case (separate words with underscore)
  - CamelCase
  - pot_hole

Organization: Sample naming systems

1. YYYYMMDD_site (date and site name)
2. YYYYMMDD_ExpmtNum (date, experiment type, experiment number)
3. Species-expmt-num (species name, experiment type, experiment number)
4. Expmt_Sample (experiment type, sample name)
5. IntXX_BY_YYYMMDD (interview number, interviewer initials, date)
6. YYYYMMDD_source_sample (date, sample source, sample name)
7. ChXXvXX (chapter and version)

Organization

- Interesting example: compare data files deposited in Borealis to GitHub structure
- Data for: Early and late cyanobacterial bloomers in a shallow, eutrophic lake
- Borealis: https://doi.org/10.5683/SP3/MP6NXF
- GitHub: https://github.com/biogeochem/buffalopound_blooms

- Without the file folder structure from GitHub, the file names are difficult to navigate.
- Image names are not descriptive for reuse

Organization: File versioning

- When collaborating or working on complex processes, save working files at key points
- Save files with a new version number before editing
- Include a version number at the end of the file name
- Keep notes on what the version contains
- Examples: “PlasmaPaper_v01”, “PlasmaPaper_v02”,…, “PlasmaPaper_FINAL”
- Consider: controlling write access to important versions so they are not accidentally overwritten

Documentation

- Documentation provides context for your data
- Who collected the data? Who was studied?
- What was collected, and for what purpose? What is the content/structure of the data?
- Where was this data collected? What were the experimental conditions?
- When was this data collected? Is it part of a series, or ongoing experiment?
- Why was this experiment performed?

- Having clear documentation helps you and your collaborators understand your data, especially as time passes

Levels of documentation

- Study-level documentation: information and context on research design, data collection and manipulation, and findings.
    - The abstract or summary at the beginning of a data management or in a data deposit
- Data-level documentation: information about individual data files
    - Variable names, data types, coding and classification schemes, codes for missing values

Check-in

- Does your dataset have:
    - Study-level documentation?
    - Data-level documentation?

Documentation: Research notes and Lab notebooks

- Research notes: use headings, always label values and figures with units, correct mistakes
- Laboratory notebooks (print or electronic):
    - Follow practices established by organization or laboratory
- In print: number pages and create an index, always date your pages, use headings, record the context of your data, paste data into book or specify the location, draw a line through errors, put a large X through empty spaces, keep it in a secure location, and scan the finished notebook
- Electronic lab notebooks should have robust note-taking and search capabilities, the ability to embed data and image files, secure log-in and tracking of data entry, audit trail for changes, and the ability to export to .pdf

Documentation: Methods

- Methods: how to acquire, analyze, and interpret the data
  - Physical set up
  - Preparation of data for analysis (details on filtering, cleaning, removing artifacts)
  - Computer code
  - Grouping of data points, information coding, units of measurement
- Track protocol changes using version numbers
- Explicitly list the protocol version in your research notes

Documentation: Interview Transcription

- When converting audio recordings of interviews to text format, check for theoretical and methodological approaches in discipline
- Develop a standard transcription template with written instructions and guidelines
- Include: unique identifier, uniform layout (including date, place, interviewer and interviewee details), speaker tags, line breaks between speakers, page numbers, and conventions for anonymization edits

Documentation: README.txt files

- Create a simple text file called README.txt that outlines the contents and organization of your data files.
- Include:
  - Project name
  - Project summary
  - Previous work on the project and location of that information
  - Funding information
  - Primary contact information
  - Other people working on the project
  - Location of data and supporting information (lab notebooks, methods, etc.)
  - Organization and naming conventions used for the data

Documentation: Data dictionaries

- Create a data dictionary of variables so that the dataset itself can be streamlined and computable
- For each variable, include:
    - Name
    - Variable definition
    - How the variable was measured
    - Data units
    - Data format
    - Minimum and maximum values
    - Coded values and their meanings
    - Representation of null values
    - Precision of measurement
    - Known issues with the data (missing values, bias, etc.)
    - Relationship to other variables

Documentation: Metadata

- Structured, standardized fields for the experimental information
- Each observation can be entered as a record, often in XML
- If a metadata schema, such as Darwin Core, is used widely in a discipline, using this schema will make your data more interoperable for others
- Schema contains:
  - Definition of each element
  - Format for each value
  - Parent and child elements
  - Possible element qualifiers
  - Required and recommended elements
  - Number of times each element type may be repeated

Documentation: Standards

- Uniformity in format or allowable values
- Makes it easier to search, retrieve, and understand data
- ISO 8601 for date formats:
  - By day: YYYY-MM-DD or YYYYMMDD
  - By month: YYYY-MM
  - By date and time: YYYY-MM-DDThh:mmX (ex. “2015-02-04T14:35Z” where X is the offset from Coordinated Universal Time)
- ISO 6709 for latitude and longitude
- Seven SI base units: m, kg, s, A, K, mol, cd
- Discipline-specific standards, classifications, taxonomies

Improving Documentation: Data dictionary and standards example

- UBC Library Circulation of Physical Items: https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP2/DLKGZV
- LANGUAGE is described as “Language of material”
- The data file has three letter codes and we need to know what the codes represent
- Cataloguers use the MARC Code for Languages from the Library of Congress, see: https://www.loc.gov/marc/languages/

Standards-based Data Entry

- Set up validation rules in data entry software
- Consider using data entry screens, such as an SPSS data entry form
- Use controlled vocabularies, code lists, and choice lists to minimize manual data entry errors

StatCan example

- Canadian Community Health Survey, Annual Component 2018:
  - https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=795204
  - Compare to documentation downloaded by electronic file transfer
    - (In Elizabeth Stregger's OneDrive, available under StatCan DLI license)

Documentation: General Tips

- Have someone else review your documentation to make sure it is understandable to an outsider
- Provide training to new team members so that they follow organization and documentation practices
- Post descriptions of filing naming conventions in shared spaces (physical and digital)