Organize Your Research Data Part 1

SFU Library Data Services Team

Website: www.lib.sfu.ca/data
Email: data-services@sfu.ca

SFU LIBRARY

Part 1: Documentation

- Why data documentation is important
- Codebooks, data dictionaries, and ReadMe files
- Levels of documentation

Why document?

- Enable other scholars to:
  - Understand your findings and verify your results
  - Review your submitted publication
  - Replicate your results
  - Design similar studies
  - Find your data in repositories
- Help you understand your own data!

What makes a good dataset?

Imagine you have found a dataset you want to use. What information would you need to interpret and use the data?

- Scope of the study
- Methodology of the study

What makes a good dataset?

Imagine you have found a dataset you want to use. What information would you need to interpret and use the data?

- Scope of the study
- Methodology of the study
- Definitions of terms
- Measurement units
- Definitions of abbreviations
- How data was anonymized
- What instrumentation was used
- Multiple data versions
- Code/script associated with the data
- And more!

Describing data

You need to create contextual information for your data (metadata):

- Final reports, working papers, lab books
- Codebooks
- Data dictionaries
- ReadMe files
- Appropriate filenames

Who?

What?

When?

Where?

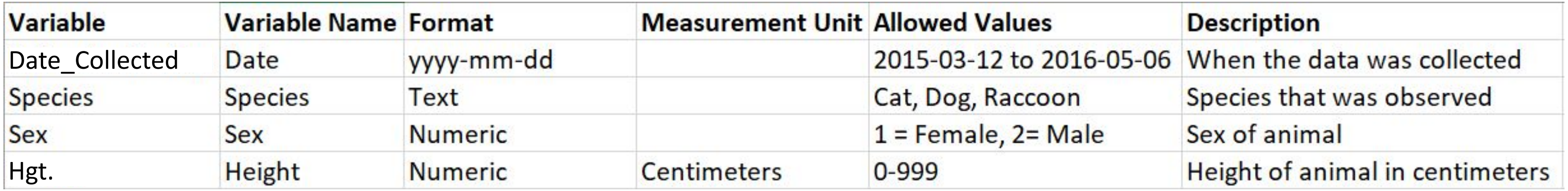
Why?

Codebooks

- Provides information about data from a survey instrument
- Can include:
  - Layout and structure of a data file
  - Response codes for each variable
  - Exact questions used in a survey
  - Missing data codes
  - A copy of the survey questionnaire
  - Information on data collection, data processing, and data quality

Data dictionaries

- Gives info about variables
  - Variable names
  - Format
  - Measurement unit
  - Expected values (nulls, mix/max, list)



Variable	Variable Name	Format	Measurement Unit	Allowed Values	Description
Date_Collected	Date	yyyy-mm-dd		2015-03-12 to 2016-05-06	When the data was collected
Species	Species	Text		Cat, Dog, Raccoon	Species that was observed
Sex	Sex	Numeric		1 = Female, 2= Male	Sex of animal
Hgt.	Height	Numeric	Centimeters	0-999	Height of animal in centimeters

ReadMe Files

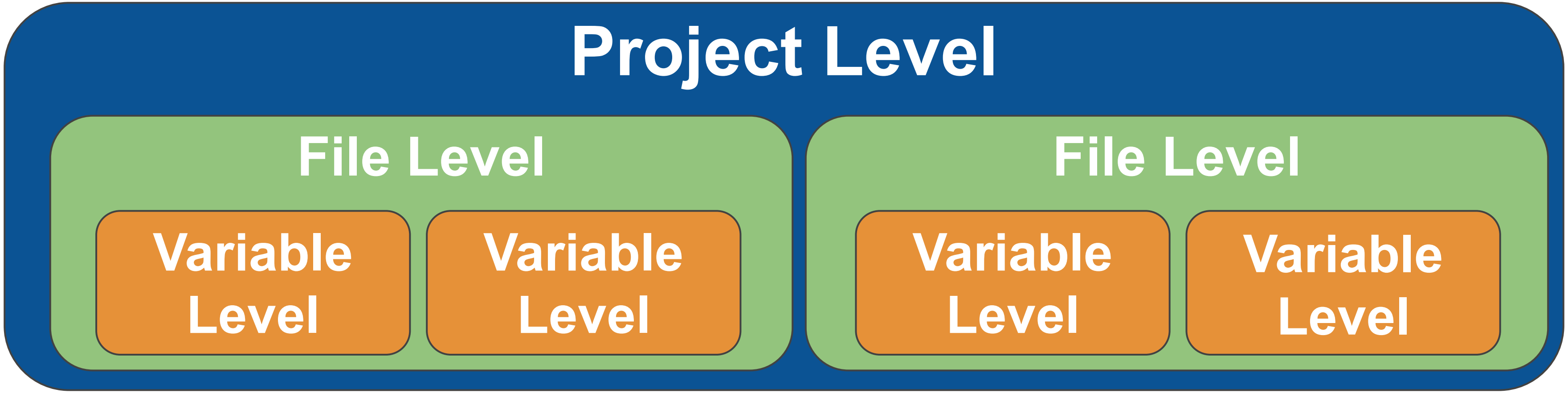
- ReadMe are plain text files that document:
    - General project Info
    - Data and file overview
    - Methodology
    - Sharing and access info

Data documentation resources

- Codebooks:
  - ICPSR Guide to Codebooks
  - Data Documentation Initiative (DDI) list of examples of marked up codebooks
- Data dictionaries:
  - U.S. Geological Survey Data Management: Data Dictionaries
  - Open Science Framework: How to Make a Data Dictionary
- ReadMe files:
  - Cornell University Guide to writing "readme" style metadata

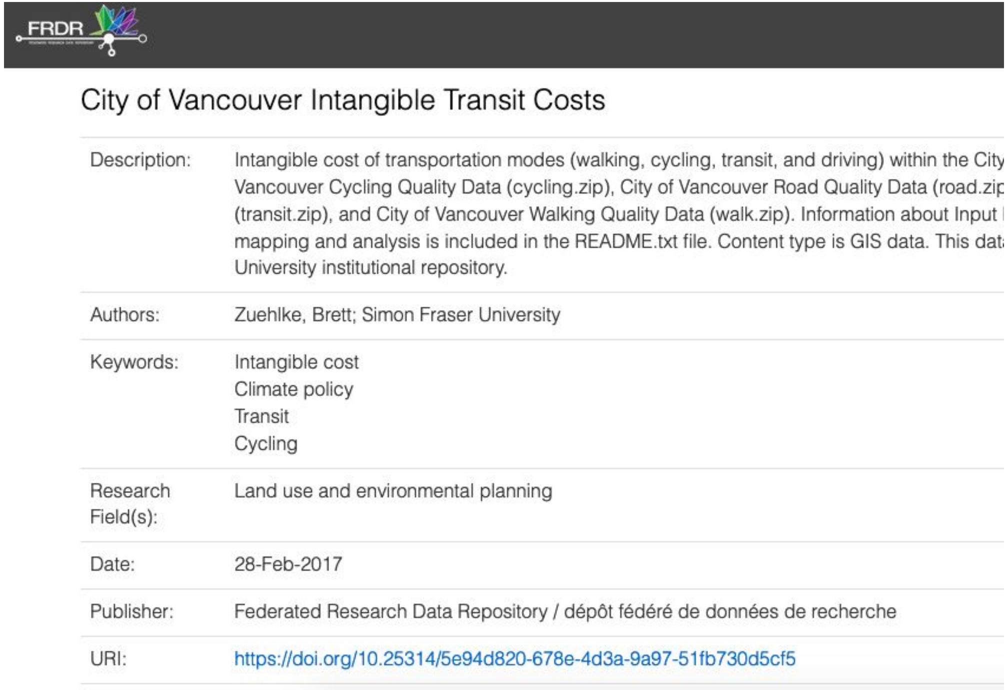
Levels of Documentation

- Project level
- File or database level
- Variable or item level



Project level documentation

- Dataset title
- Authors/contact
- Objectives
- Hypothesis
- Methodology
- Date(s) & location(s)
- Funding
- Licence/copyright
- Persistent identifier (e.g. DOI)



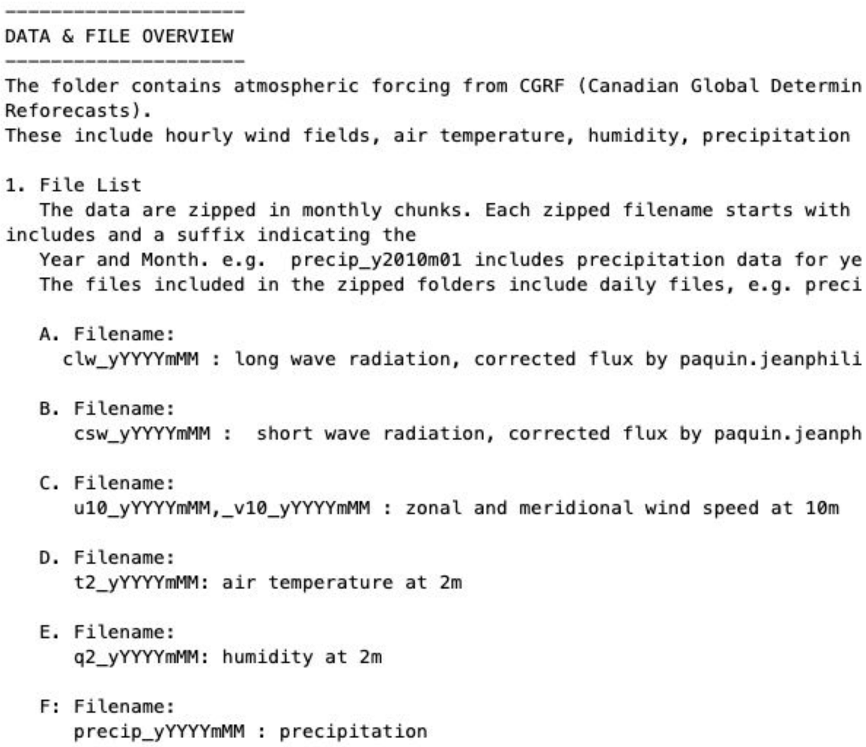
City of Vancouver Intangible Transit Costs

Description:	Intangible cost of transportation modes (walking, cycling, transit, and driving) within the City Vancouver Cycling Quality Data (cycling.zip), City of Vancouver Road Quality Data (road.zip (transit.zip), and City of Vancouver Walking Quality Data (walk.zip). Information about Input mapping and analysis is included in the README.txt file. Content type is GIS data. This dat University institutional repository.
Authors:	Zuehlke, Brett; Simon Fraser University
Keywords:	Intangible cost Climate policy Transit Cycling
Research Field(s):	Land use and environmental planning
Date:	28-Feb-2017
Publisher:	Federated Research Data Repository / dépôt fédéré de données de recherche
URI:	https://doi.org/10.25314/5e94d820-678e-4d3a-9a97-51fb730d5cf5

https://doi.org/10.25314/5e94d820-678e-4d3a-9a97-51fb730d5cf5

File or dataset level documentation

- Filename
- Description
- Format
- Date(s) & location(s)
- Version
- Relationship between files
- Software used

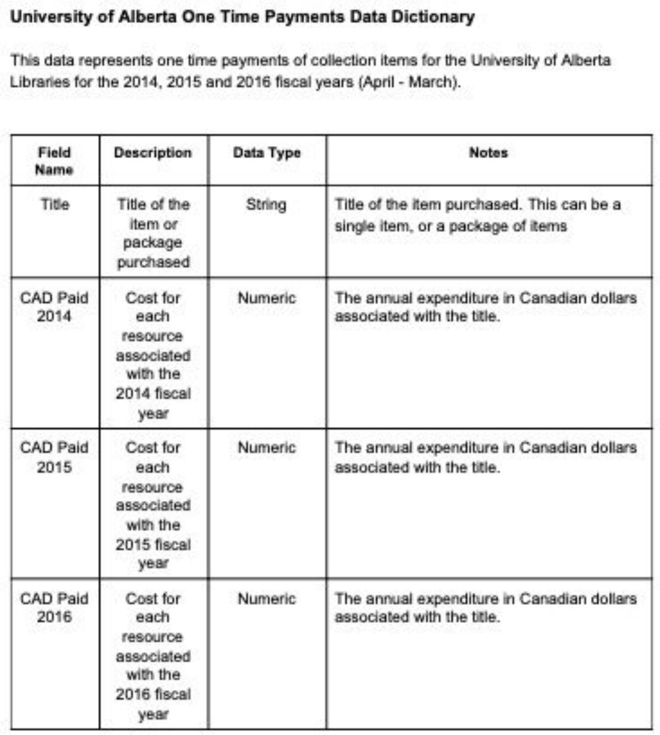


```
---------------------
DATA & FILE OVERVIEW
---------------------
The folder contains atmospheric forcing from CGRF (Canadian Global Determin
Reforecasts).
These include hourly wind fields, air temperature, humidity, precipitation

1. File List
   The data are zipped in monthly chunks. Each zipped filename starts with
includes and a suffix indicating the
   Year and Month. e.g.  precip_y2010m01 includes precipitation data for ye
   The files included in the zipped folders include daily files, e.g. preci

   A. Filename:
     clw_yYYYYmMM : long wave radiation, corrected flux by paquin.jeanphili

   B. Filename:
      csw_yYYYYmMM :  short wave radiation, corrected flux by paquin.jeanph

   C. Filename:
      u10_yYYYYmMM,_v10_yYYYYmMM : zonal and meridional wind speed at 10m

   D. Filename:
      t2_yYYYYmMM: air temperature at 2m

   E. Filename:
      q2_yYYYYmMM: humidity at 2m

   F: Filename:
      precip_yYYYYmMM : precipitation
```

https://doi.org/10.20383/101.023 ; CGRF_README.txt

Variable or item level documentation

- Variable name
- Description
- Data type
- Code for missing values
- Units of measurements
- Notes



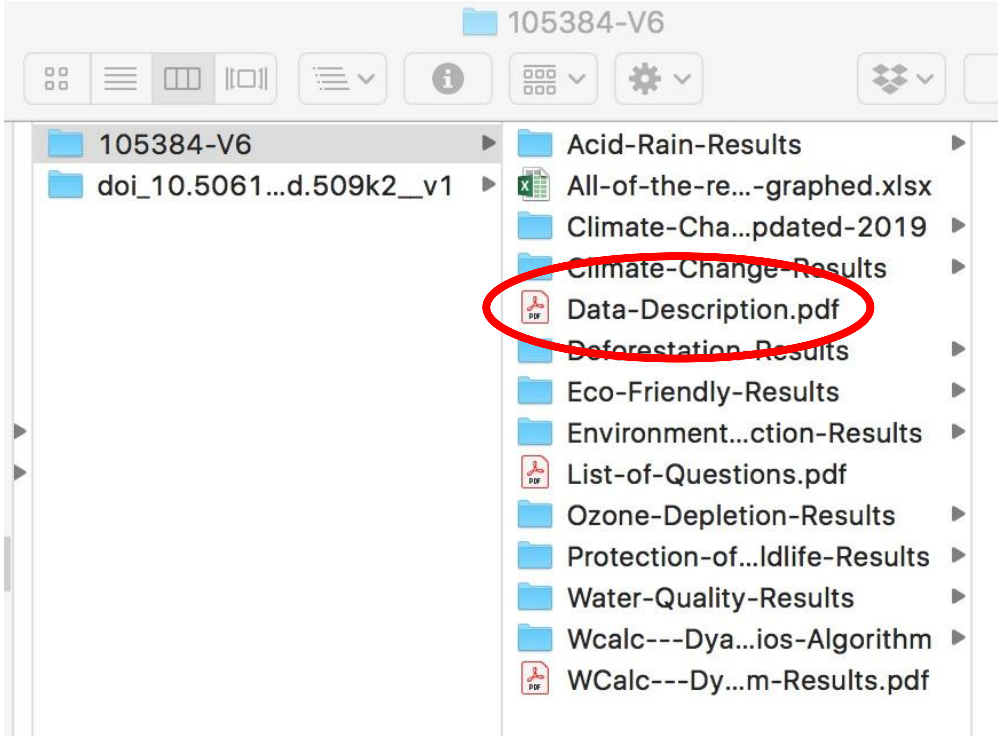
University of Alberta One Time Payments Data Dictionary

This data represents one time payments of collection items for the University of Alberta Libraries for the 2014, 2015 and 2016 fiscal years (April - March).

Field Name	Description	Data Type	Notes
Title	Title of the item or package purchased	String	Title of the item purchased. This can be a single item, or a package of items
CAD Paid 2014	Cost for each resource associated with the 2014 fiscal year	Numeric	The annual expenditure in Canadian dollars associated with the title.
CAD Paid 2015	Cost for each resource associated with the 2015 fiscal year	Numeric	The annual expenditure in Canadian dollars associated with the title.
CAD Paid 2016	Cost for each resource associated with the 2016 fiscal year	Numeric	The annual expenditure in Canadian dollars associated with the title.

https://doi.org/10.7939/DVN/10963 ; One_Time_Expenditures_DataDictionary.pdf

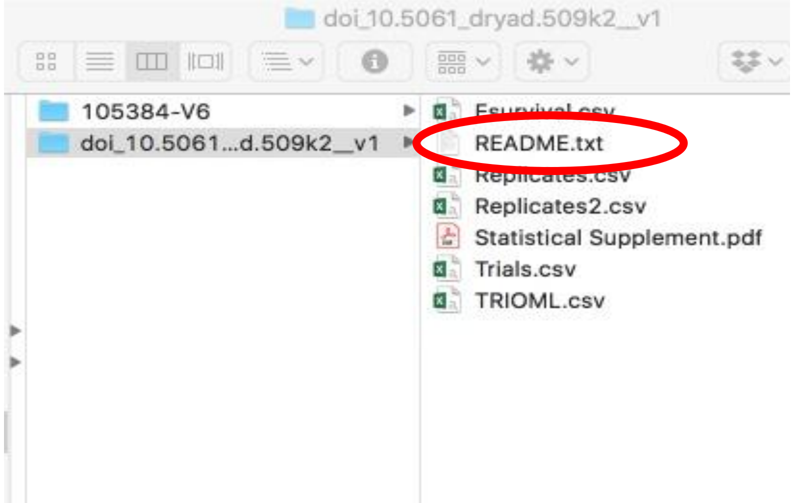
Metadata is essential



105384-V6

105384-V6
doi_10.5061...d.509k2__v1

Acid-Rain-Results
All-of-the-re...-graphed.xlsx
Climate-Cha...pdated-2019
Climate-Change-Results
Data-Description.pdf
Deforestation-Results
Eco-Friendly-Results
Environment...ction-Results
List-of-Questions.pdf
Ozone-Depletion-Results
Protection-of...ldlife-Results
Water-Quality-Results
Wcalc---Dya...ios-Algorithm
WCalc---Dy...m-Results.pdf

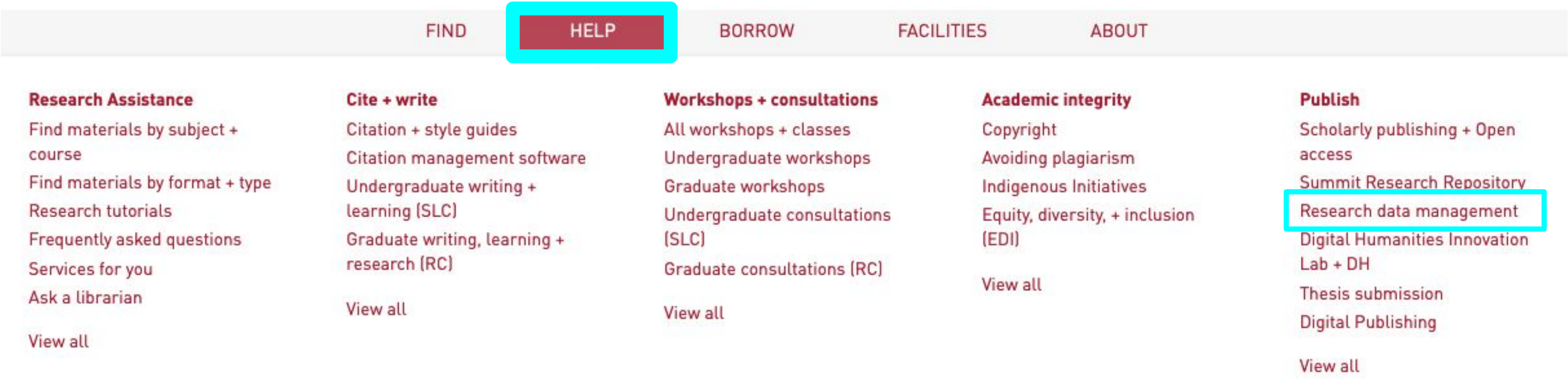


doi_10.5061_dryad.509k2__v1

105384-V6
doi_10.5061...d.509k2__v1

Esurvival.csv
README.txt
Replicates.csv
Replicates2.csv
Statistical Supplement.pdf
Trials.csv
TRIOML.csv

For more information



FIND HELP BORROW FACILITIES ABOUT

Research Assistance

Find materials by subject + course

Find materials by format + type

Research tutorials

Frequently asked questions

Services for you

Ask a librarian

View all

Cite + write

Citation + style guides

Citation management software

Undergraduate writing + learning (SLC)

Graduate writing, learning + research (RC)

View all

Workshops + consultations

All workshops + classes

Undergraduate workshops

Graduate workshops

Undergraduate consultations (SLC)

Graduate consultations (RC)

View all

Academic integrity

Copyright

Avoiding plagiarism

Indigenous Initiatives

Equity, diversity, + inclusion (EDI)

View all

Publish

Scholarly publishing + Open access

Summit Research Repository

Research data management

Digital Humanities Innovation Lab + DH

Thesis submission

Digital Publishing

View all

Contact us at: data-services@sfu.ca

Organize Your Research Data Part 2

SFU Library Data Services Team

Website: www.lib.sfu.ca/data
Email: data-services@sfu.ca

SFU LIBRARY

Part 2: File management

- File and folder naming
- Versioning
- File organization

File and folder names

- Keep names short, but meaningful
- Don’t use spaces!
  - Use camelCase (dateOfCollection)
  - Or underscores (Date_Of_Collection)
- Date format should be YYYYMMDD
- File names should be descriptive outside their folders

Project No.	Create date	Creator	Description	Research team	Publication date	Version

AHRC_TechnicalAppResponse_20120925_v01_02.docx

by [ULeicester, Research Data: Naming files and folders]

File versioning

- Avoid descriptive version labels
- Zero-filled numbers for major version changes (e.g. 01, 02, 03)
- Underscores for minor changes (i.e. 01_01, 01_02)

Smith_interview_July2010_1	→	Smith_interview_201006_V01
lipid analysis rate edited2		LipidAnalysisRate_V02_02
Nov2801_ILB_AB_CS3_6		20011128_ILB_AB_CS3_V06

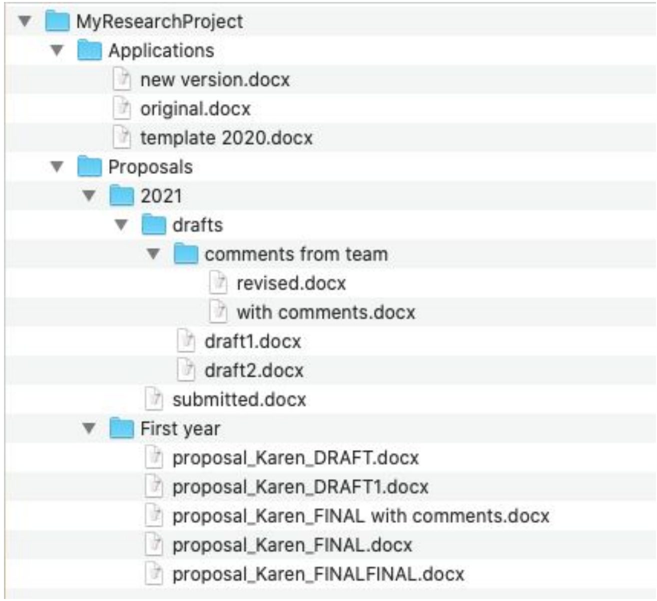
- Consider version control system (e.g., Open Science Framework, Git, Wiki, etc)

File renaming

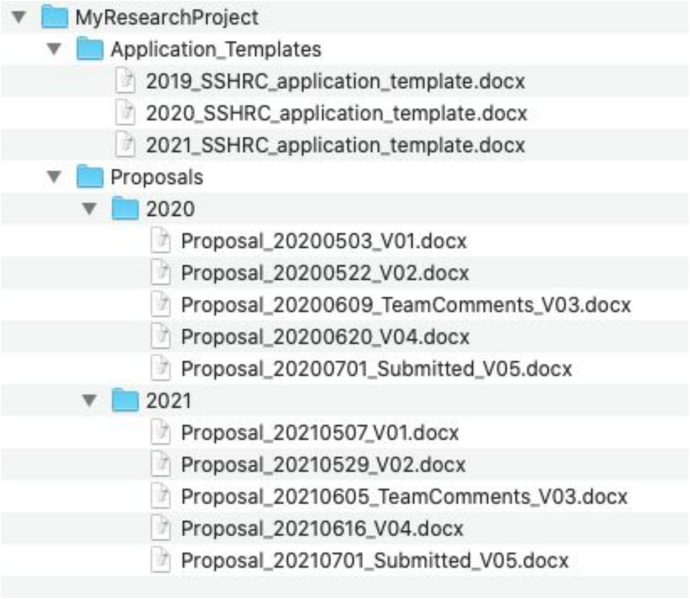
- Tools:
  - Bulk Rename Utility (Windows)
  - WildRename (Windows)
  - Renamer (MacOS)
- Ensure you have back ups before you start renaming!
- Document how filenames changed

Directory Structures

- Use folder hierarchy from general to specific
    - Don’t go too deep - use file names instead



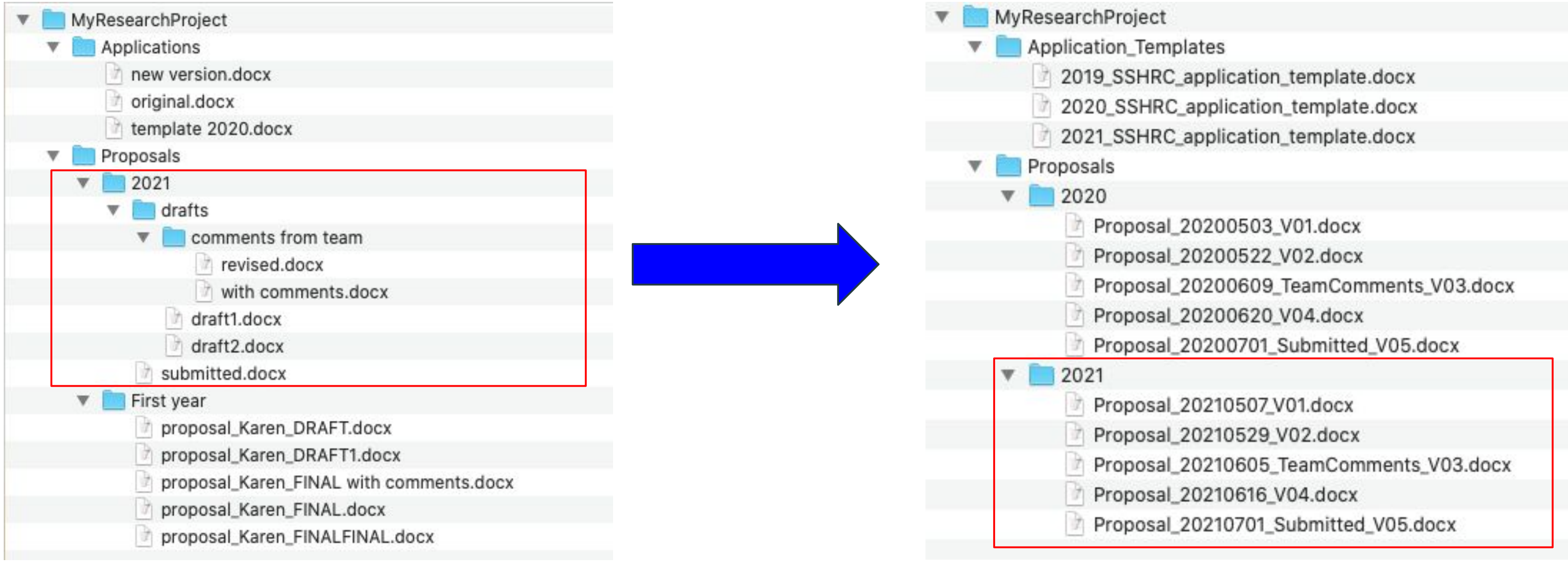
- MyResearchProject
    - Applications
        - new version.docx
        - original.docx
        - template 2020.docx
    - Proposals
        - 2021
            - drafts
                - comments from team
                    - revised.docx
                    - with comments.docx
                - draft1.docx
                - draft2.docx
            - submitted.docx
        - First year
            - proposal_Karen_DRAFT.docx
            - proposal_Karen_DRAFT1.docx
            - proposal_Karen_FINAL with comments.docx
            - proposal_Karen_FINAL.docx
            - proposal_Karen_FINALFINAL.docx



- MyResearchProject
    - Application_Templates
        - 2019_SSHRC_application_template.docx
        - 2020_SSHRC_application_template.docx
        - 2021_SSHRC_application_template.docx
    - Proposals
        - 2020
            - Proposal_20200503_V01.docx
            - Proposal_20200522_V02.docx
            - Proposal_20200609_TeamComments_V03.docx
            - Proposal_20200620_V04.docx
            - Proposal_20200701_Submitted_V05.docx
        - 2021
            - Proposal_20210507_V01.docx
            - Proposal_20210529_V02.docx
            - Proposal_20210605_TeamComments_V03.docx
            - Proposal_20210616_V04.docx
            - Proposal_20210701_Submitted_V05.docx

Directory Structures

- Use folder hierarchy from general to specific
  - Don’t go too deep - use file names instead

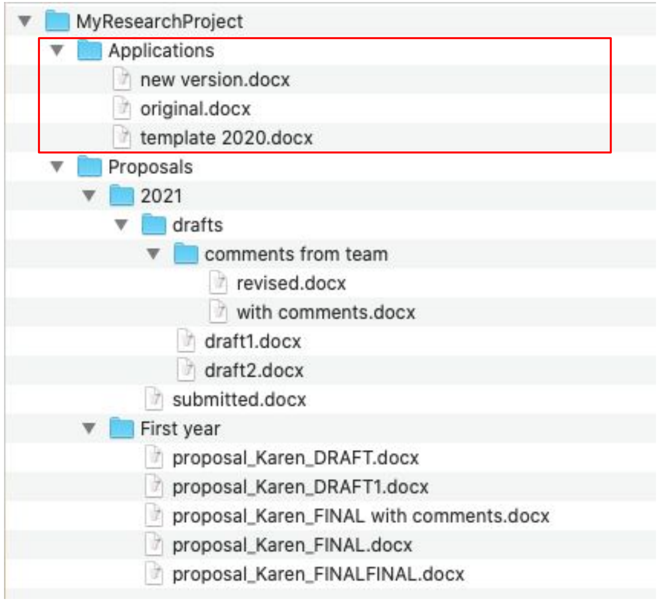


```
▼ MyResearchProject
    ▼ Applications
          new version.docx
          original.docx
          template 2020.docx
    ▼ Proposals
        ▼ 2021
            ▼ drafts
                ▼ comments from team
                      revised.docx
                      with comments.docx
                  draft1.docx
                  draft2.docx
              submitted.docx
        ▼ First year
              proposal_Karen_DRAFT.docx
              proposal_Karen_DRAFT1.docx
              proposal_Karen_FINAL with comments.docx
              proposal_Karen_FINAL.docx
              proposal_Karen_FINALFINAL.docx
```

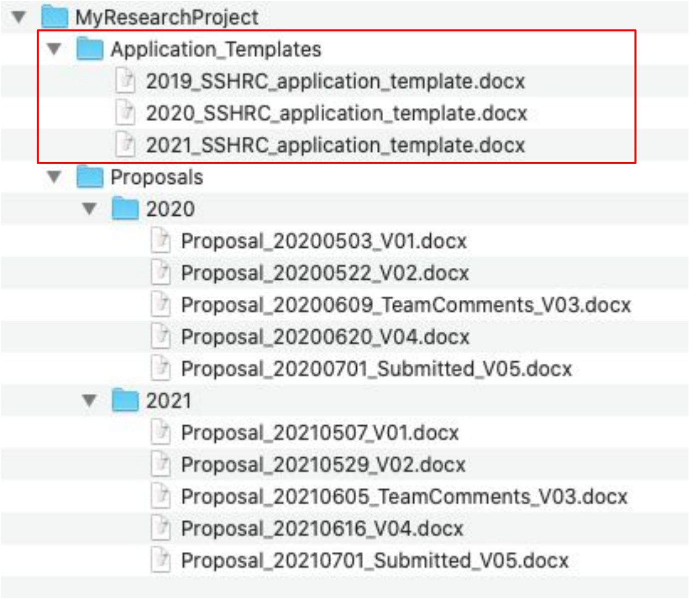
```
▼ MyResearchProject
    ▼ Application_Templates
          2019_SSHRC_application_template.docx
          2020_SSHRC_application_template.docx
          2021_SSHRC_application_template.docx
    ▼ Proposals
        ▼ 2020
              Proposal_20200503_V01.docx
              Proposal_20200522_V02.docx
              Proposal_20200609_TeamComments_V03.docx
              Proposal_20200620_V04.docx
              Proposal_20200701_Submitted_V05.docx
        ▼ 2021
              Proposal_20210507_V01.docx
              Proposal_20210529_V02.docx
              Proposal_20210605_TeamComments_V03.docx
              Proposal_20210616_V04.docx
              Proposal_20210701_Submitted_V05.docx
```

Directory Structures

- Use folder hierarchy from general to specific
    - Don’t go too deep - use file names instead



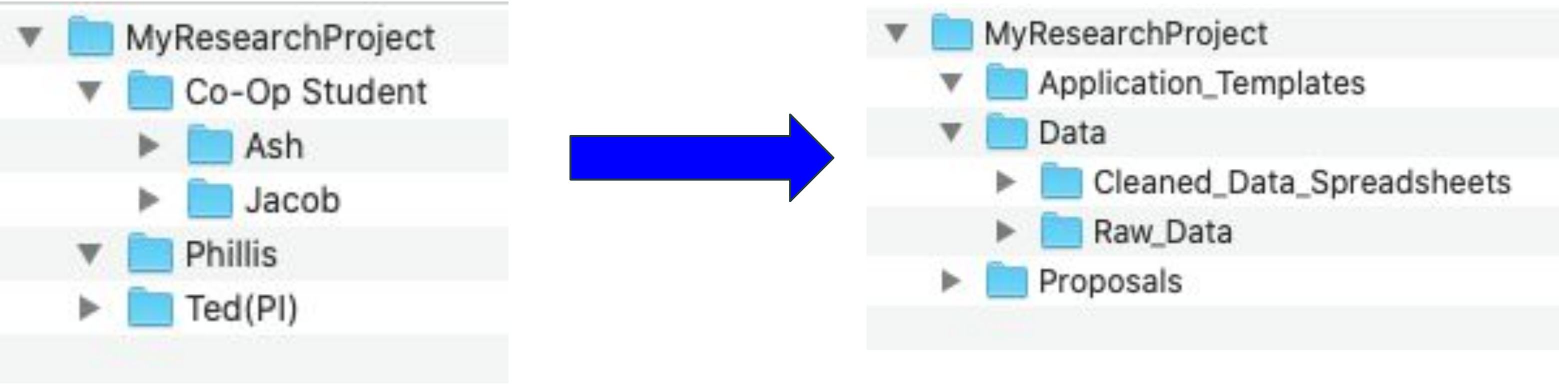
```
▼ MyResearchProject
    ▼ Applications
          new version.docx
          original.docx
          template 2020.docx
    ▼ Proposals
        ▼ 2021
            ▼ drafts
                ▼ comments from team
                      revised.docx
                      with comments.docx
                  draft1.docx
                  draft2.docx
              submitted.docx
        ▼ First year
              proposal_Karen_DRAFT.docx
              proposal_Karen_DRAFT1.docx
              proposal_Karen_FINAL with comments.docx
              proposal_Karen_FINAL.docx
              proposal_Karen_FINALFINAL.docx
```



```
▼ MyResearchProject
    ▼ Application_Templates
          2019_SSHRC_application_template.docx
          2020_SSHRC_application_template.docx
          2021_SSHRC_application_template.docx
    ▼ Proposals
        ▼ 2020
              Proposal_20200503_V01.docx
              Proposal_20200522_V02.docx
              Proposal_20200609_TeamComments_V03.docx
              Proposal_20200620_V04.docx
              Proposal_20200701_Submitted_V05.docx
        ▼ 2021
              Proposal_20210507_V01.docx
              Proposal_20210529_V02.docx
              Proposal_20210605_TeamComments_V03.docx
              Proposal_20210616_V04.docx
              Proposal_20210701_Submitted_V05.docx
```

Folder Organization

- Base folders on factors that will not change over the course of the project
    - People may leave, departments may change, etc.
- List all folders needed and try to group them logically

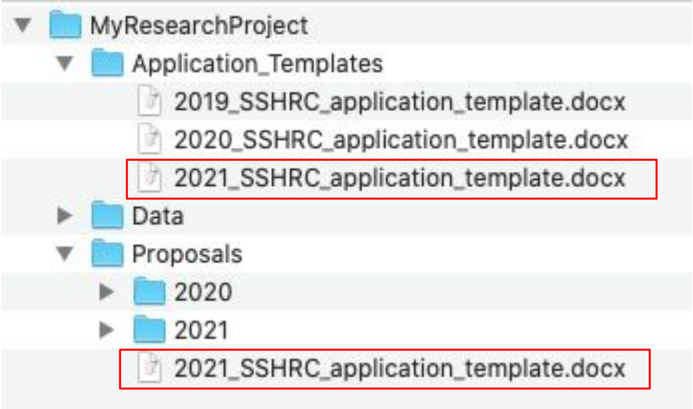


- MyResearchProject
    - Co-Op Student
        - Ash
        - Jacob
    - Phillis
    - Ted(PI)

- MyResearchProject
    - Application_Templates
    - Data
        - Cleaned_Data_Spreadsheets
        - Raw_Data
    - Proposals

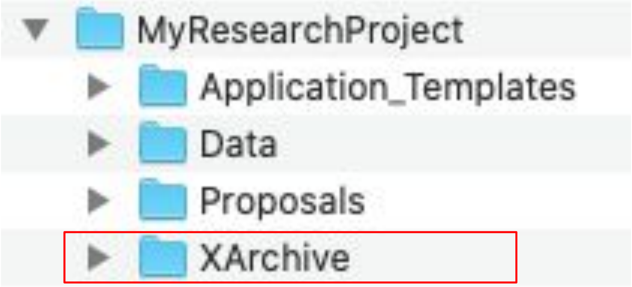
Things to avoid

- Having multiple copies of files in different folders



MyResearchProject
Application_Templates
2019_SSHRC_application_template.docx
2020_SSHRC_application_template.docx
2021_SSHRC_application_template.docx
Data
Proposals
2020
2021
2021_SSHRC_application_template.docx

- Deleting files
  - Intead, keep a separate folder for archiving



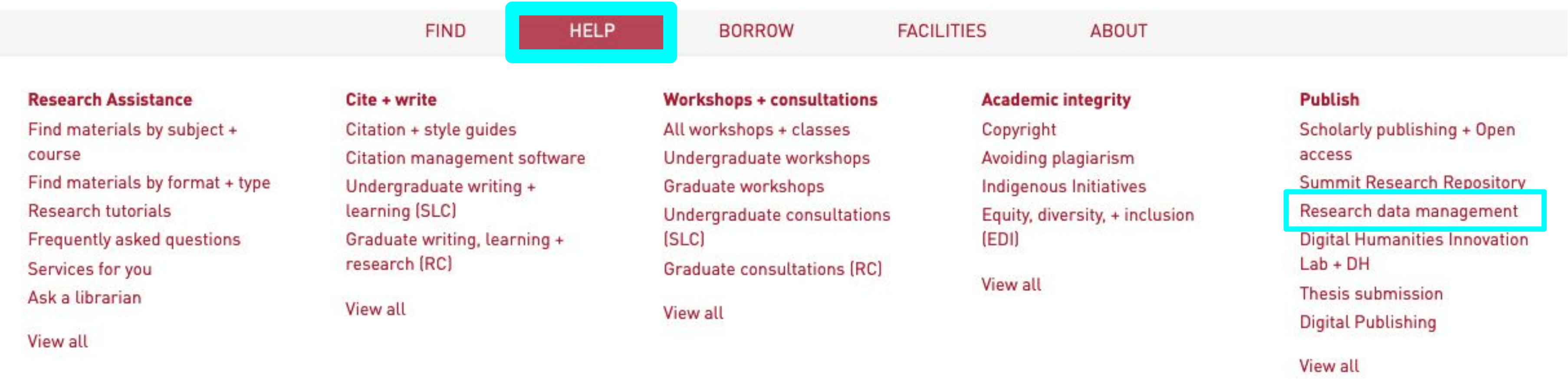
MyResearchProject
Application_Templates
Data
Proposals
XArchive

  - Symbols or letters can be added for sorting

Basics of File Management

- Plan structure ahead of time
  - Periodically review and cleanup the structure
- Apply the structure consistently
  - All members of the team should apply structure to all locations where data is stored
- Document the structure
- Use descriptive file names

For more information



Research Assistance
- Find materials by subject + course
- Find materials by format + type
- Research tutorials
- Frequently asked questions
- Services for you
- Ask a librarian
- View all

Cite + write
- Citation + style guides
- Citation management software
- Undergraduate writing + learning (SLC)
- Graduate writing, learning + research (RC)
- View all

Workshops + consultations
- All workshops + classes
- Undergraduate workshops
- Graduate workshops
- Undergraduate consultations (SLC)
- Graduate consultations (RC)
- View all

Academic integrity
- Copyright
- Avoiding plagiarism
- Indigenous Initiatives
- Equity, diversity, + inclusion (EDI)
- View all

Publish
- Scholarly publishing + Open access
- Summit Research Repository
- Research data management
- Digital Humanities Innovation Lab + DH
- Thesis submission
- Digital Publishing
- View all

Contact us at: data-services@sfu.ca