Organize Your Research Data
Part 1

SFU Library Data Services Team

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

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)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable Name</th>
<th>Format</th>
<th>Measurement Unit</th>
<th>Allowed Values</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date_Collected</td>
<td>Date</td>
<td>yyyy-mm-dd</td>
<td></td>
<td>2015-03-12 to 2016-05-06</td>
<td>When the data was collected</td>
</tr>
<tr>
<td>Species</td>
<td>Species</td>
<td>Text</td>
<td>Cat, Dog, Raccoon</td>
<td></td>
<td>Species that was observed</td>
</tr>
<tr>
<td>Sex</td>
<td>Sex</td>
<td>Numeric</td>
<td>1 = Female, 2 = Male</td>
<td></td>
<td>Sex of animal</td>
</tr>
<tr>
<td>Hgt.</td>
<td>Height</td>
<td>Numeric</td>
<td>Centimeters</td>
<td>0-999</td>
<td>Height of animal in centimeters</td>
</tr>
</tbody>
</table>
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
City of Vancouver Intangible Transit Costs

Description: Intangible cost of transportation modes (walking, cycling, transit, and driving) within the City of Vancouver. The data includes Cycling Quality Data (cycling.zip), Road Quality Data (road.zip), and 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 file is deposited in the 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
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 Deterministic 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. `precip_y2010m01`.

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

   B. Filename:
   `csw_yYYYYMMM`: short wave radiation, corrected flux by paquin.jeanphili

   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).

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

[https://doi.org/10.7939/DVN/10963; One_Time_Expenditures_DataDictionary.pdf](https://doi.org/10.7939/DVN/10963; One_Time_Expenditures_DataDictionary.pdf)
Metadata is essential
For more information

Contact us at: data-services@sfu.ca
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

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)

- Consider version control system (e.g., [Open Science Framework](https://osf.io), Git, Wiki, etc)

---

<table>
<thead>
<tr>
<th>Smith_interview_July2010_1</th>
<th>Smith_interview_201006_V01</th>
</tr>
</thead>
<tbody>
<tr>
<td>lipid analysis rate edited2</td>
<td>LipidAnalysisRate_V02_02</td>
</tr>
<tr>
<td>Nov2801_ILB_AB_CS3_6</td>
<td>Nov20011128_ILB_AB_CS3_V06</td>
</tr>
</tbody>
</table>
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
Directory Structures

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

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

- Having multiple copies of files in different folders

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

Contact us at: data-services@sfu.ca