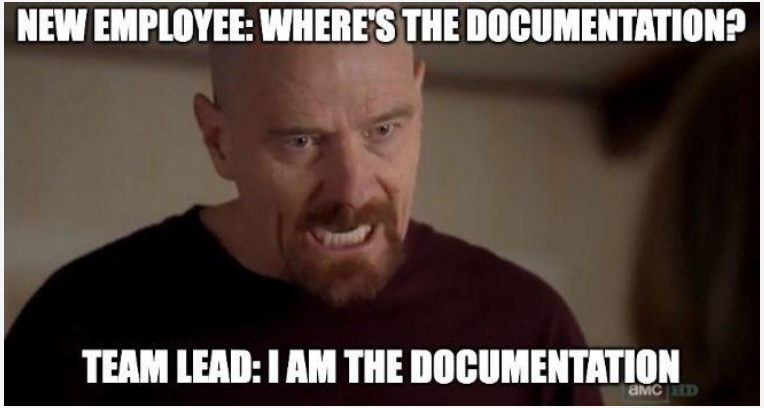
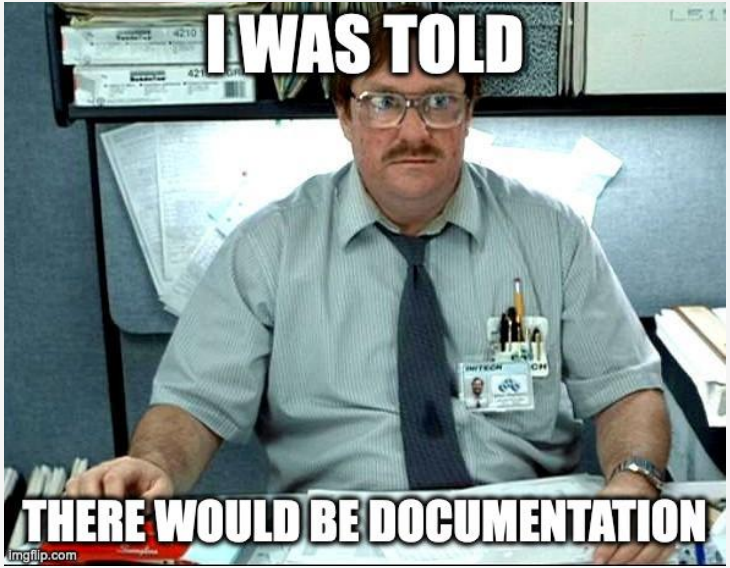

NEW EMPLOYEE: WHERE'S THE DOCUMENTATION?
TEAM LEAD: I AM THE DOCUMENTATION
amc HD

source

Documentation

Session Overview

- Identify the importance of documentation as it relates to RDM and the FAIR Principles
- Key concepts to cover in a README document
- Data dictionaries as an alternative/additional form of documentation
- Discuss general best practices of data licensing, and how it pertains to the project's data



source

Why Document Your Files?

Documentation

- A way to ensure that others (and your future selves) can navigate and correctly interpret that files and data of a project
- Crucial in achieving FAIR data

Findable
Accessible
Interoperable
Reusable

source

Documentation

Questions to ask yourself:

- Can you and your collaborators easily find and interpret files?
- Could people outside of your group be able to find and interpret your files?

README Files

- A file that sits in a project’s root directory (sometimes there can be multiple README files for a project), and provides information about the files and their content

- During a project, keeping an updated README file will help you and your team having a source of truth regarding your project’s files

- After a project’s completion, a README file can be used by those who might be accessing your data, as a sort of instruction manual on how to navigate and use the data

README Files

Things to include in a README file:

- Contact information for the researcher(s)
- Data collection methods (protocols, sampling, instruments, coverages, etc.)
- File structures
- Naming conventions of files, if applicable
- Description of data cleaning, analysis, manipulations, or modifications
- Descriptions of variables and explanations of codes and classifications
- Data confidentiality and permissions, if applicable
- Data use license

README Files

More considerations:

- Create README files for logical clusters of related files/data
- Write your README as a plain text document (.txt or .md)
- Prepend the filename with an _ so that it shows up at the top of the file list
- If using multiple README files, place them in sensical locations and format identically
- Be sure to update!

Exercise

Take a look at the following datasets:

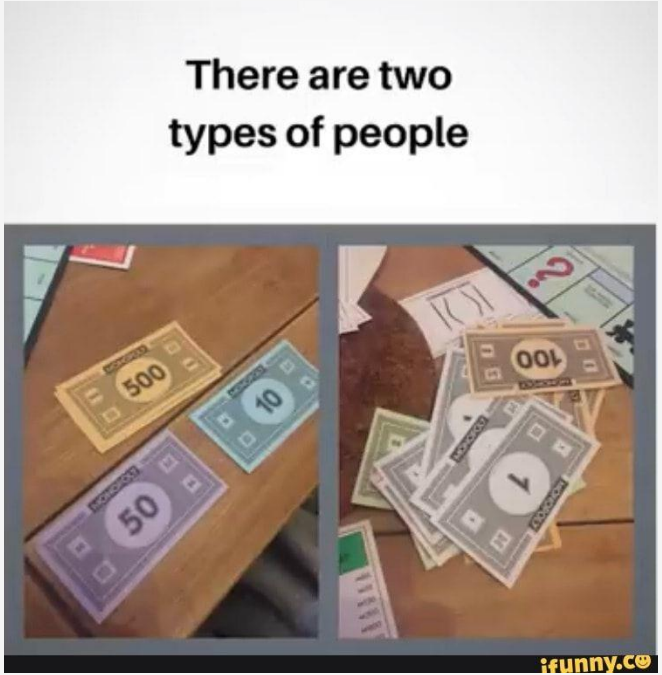
- Kampen, Andrea; Pearson, Maggie; Smit, Michael, 2018, "Replication Data for: Digital Tools and Techniques in Scholarship and Pedagogy in the Social Sciences and Humanities", https://doi.org/10.23685/1H9TOV

- Livingstone, D.W., 2021, "7 Replication Data for: 2017 CWKE Registered Nursing Dataset", https://doi.org/10.5683/SP2/I98O1W

- Perron, Maxime, 2023, "Interindividual variability in the benefits of personal sound amplification products on speech perception in noise: a randomized cross-over clinical trial", https://doi.org/10.5683/SP3/HTMDLI

Exercise

- What kind of documentation do you see?
- Can you tell what each of the files is?
- When looking at a data file, can you understand what you’re looking at?
- Is there anything that sticks out to you as interesting? Good? Bad?



source

Caveat: 2 Types of READMEs

Caveat

Things to include in a **DEPOSIT** README file:

- Contact information for the researcher(s)
- Data collection methods (protocols, sampling, instruments, coverages, etc.)
- File structures
- Naming conventions of files, if applicable
- Description of data cleaning, analysis, manipulations, or modifications
- Descriptions of variables and explanations of codes and classifications
- Data confidentiality and permissions, if applicable
- Data use license

Caveat

Things to include in an **ACTIVE** README file:

- Contact information for the researcher(s)
- Data collection methods (protocols, sampling, instruments, coverages, etc.)
- **File structures**
- **Naming conventions of files, if applicable**
- Description of data cleaning, analysis, manipulations, or modifications
- **Descriptions of variables and explanations of codes and classifications**
- **Data confidentiality and permissions, if applicable**
- Data use license

Questions?

Data Dictionaries / Codebooks

- A file that describes each element of tabular datasets
- Details of variagble names, labels, units, and constraints such as acceptable range of values
- Can enable software programs (R, Python, etc.) to read and process a data file, enhancing machine-readability, interoperability, and data reuse
- Provides human-readable details to support interpretation and analysis

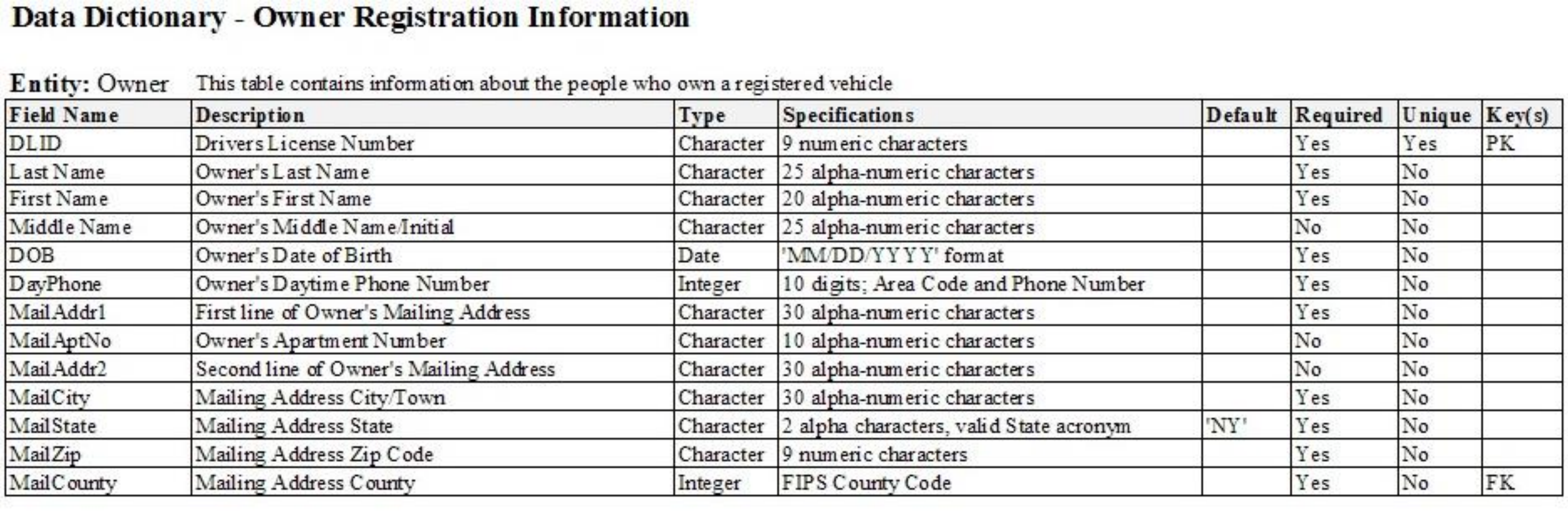
source

Caveat: 2 Types of Data Dictionaries

Our Data Dictionary (Stats Can)

https://osf.io/p7cv8

Machine Readable



Data Dictionary - Owner Registration Information

Entity: Owner This table contains information about the people who own a registered vehicle

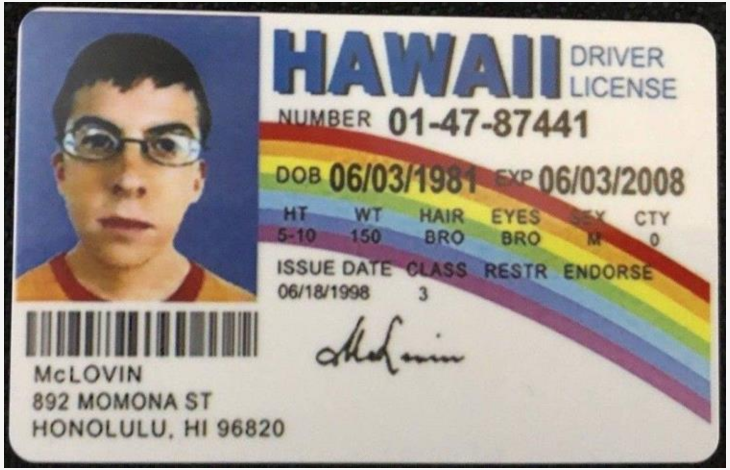
Field Name	Description	Type	Specifications	Default	Required	Unique	Key(s)
DLID	Drivers License Number	Character	9 numeric characters		Yes	Yes	PK
Last Name	Owner's Last Name	Character	25 alpha-numeric characters		Yes	No	
First Name	Owner's First Name	Character	20 alpha-numeric characters		Yes	No	
Middle Name	Owner's Middle Name/Initial	Character	25 alpha-numeric characters		No	No	
DOB	Owner's Date of Birth	Date	'MM/DD/YYYY' format		Yes	No	
DayPhone	Owner's Daytime Phone Number	Integer	10 digits; Area Code and Phone Number		Yes	No	
MailAddr1	First line of Owner's Mailing Address	Character	30 alpha-numeric characters		Yes	No	
MailAptNo	Owner's Apartment Number	Character	10 alpha-numeric characters		No	No	
MailAddr2	Second line of Owner's Mailing Address	Character	30 alpha-numeric characters		No	No	
MailCity	Mailing Address City/Town	Character	30 alpha-numeric characters		Yes	No	
MailState	Mailing Address State	Character	2 alpha characters, valid State acronym	'NY'	Yes	No	
MailZip	Mailing Address Zip Code	Character	9 numeric characters		Yes	No	
MailCounty	Mailing Address County	Integer	FIPS County Code		Yes	No	FK

source

Sneak Peak - README Template

- In the next session we’ll begin filling out a README for our project. We’ll take a quick look at the temlate that we’re using, which can be found in the ***docs*** folder in OSF

- https://data.research.cornell.edu/data-management/sharing/readme/



HAWAII DRIVER LICENSE

NUMBER 01-47-87441

DOB 06/03/1981 EXP 06/03/2008

HT	WT	HAIR	EYES	SEX	CTY
5-10	150	BRO	BRO	M	0

ISSUE DATE	CLASS	RESTR	ENDORSE
06/18/1998	3		

McLovin

McLOVIN
892 MOMONA ST
HONOLULU, HI 96820

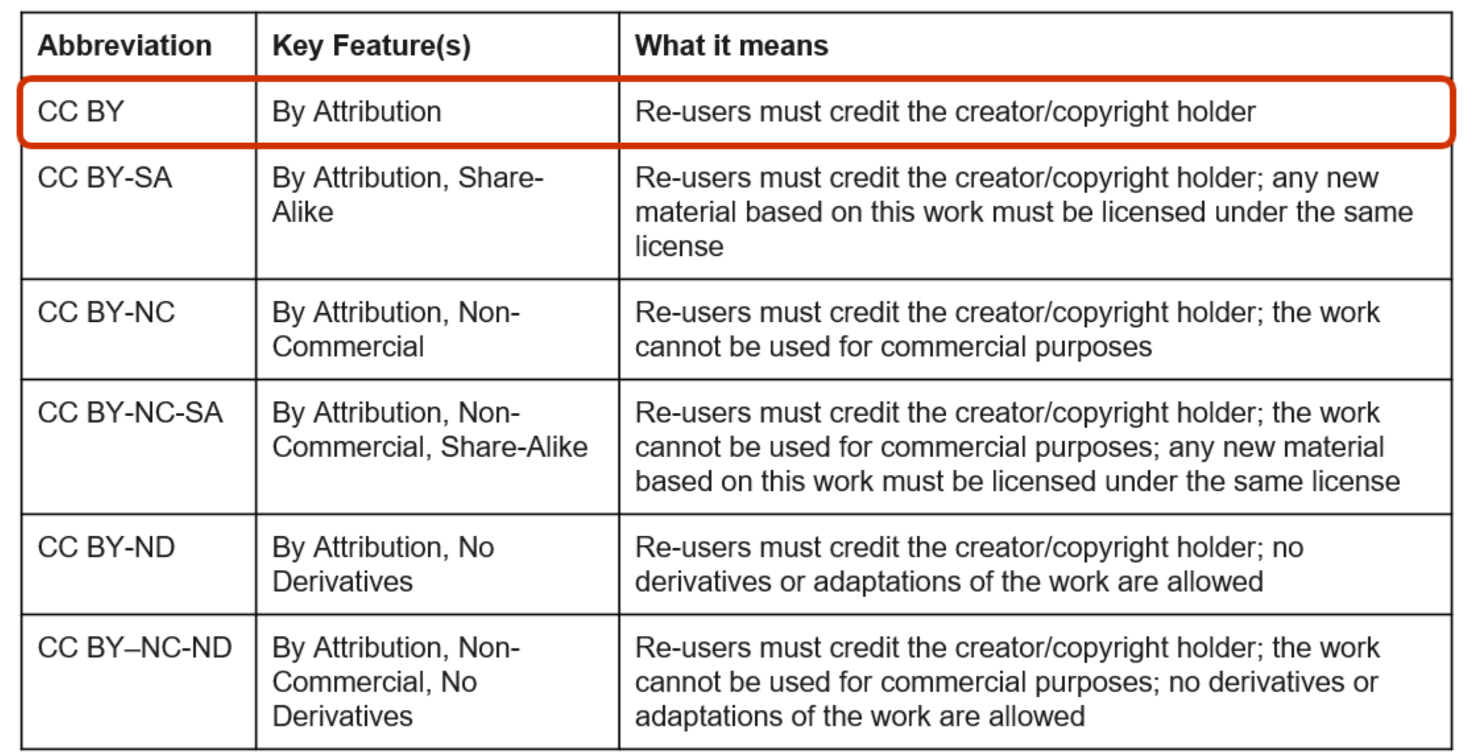
Licenses

source

Data Licenses

- A data license is a legal arrangement between the creator of the data and the end-user, specifying what can be done with the data.
- As a producer of data, this allows your work to be shared/used in the ways that you are comfortable with.
- As a consumer of data, this provides boundaries of what you are able to do with data you encounter.

Creative Commons Licenses

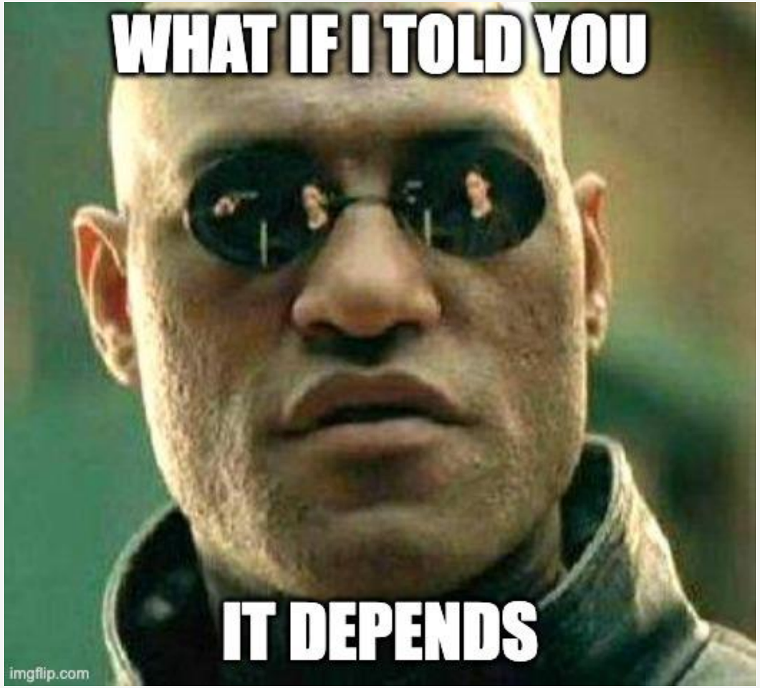


Abbreviation	Key Feature(s)	What it means
CC BY	By Attribution	Re-users must credit the creator/copyright holder
CC BY-SA	By Attribution, Share-Alike	Re-users must credit the creator/copyright holder; any new material based on this work must be licensed under the same license
CC BY-NC	By Attribution, Non-Commercial	Re-users must credit the creator/copyright holder; the work cannot be used for commercial purposes
CC BY-NC-SA	By Attribution, Non-Commercial, Share-Alike	Re-users must credit the creator/copyright holder; the work cannot be used for commercial purposes; any new material based on this work must be licensed under the same license
CC BY-ND	By Attribution, No Derivatives	Re-users must credit the creator/copyright holder; no derivatives or adaptations of the work are allowed
CC BY–NC-ND	By Attribution, Non-Commercial, No Derivatives	Re-users must credit the creator/copyright holder; the work cannot be used for commercial purposes; no derivatives or adaptations of the work are allowed

Licenses – Our Dataset

- At the end of the program, we'll be looking at data repositories and depositing data, and we'll be applying a license to our data set (stay tuned!).
- However, because we'll be working with an existing dataset, we need to make sure that we understand its license and what that entails.
- Let's take a look!

https://doi.org/10.5683/SP3/RDS0CK



Ethical, legal, and commercial considerations

source

Ethical, legal, and commercial considerations

- Research involving any of the following requires additional considerations:
  - Human participants / personal information
  - Animals
  - Collaborators at other institutions
  - Industry partners
  - Community organizations
  - Indigenous communities
  - Communities that have traditionally been marginalized or tokenized
  - Others?

Ethics and Consent Forms

- Research involving human participants requires an ethics application and informed consent.

- Consent/information letters to participants must describe how data are handled during active phases of research as well as post-project.

- It can be quite difficult, or even impossible, to revise participant consent, so getting things right at the start of a project is very helpful!

- UVic Ethics applications contain sections on Data Management and Informed Consent to help guide this process.

- Opportunities for collaboration between ORS and the Library!

Partnerships and Contracts

- Research involving industry partners, community organizations, or even researchers from different institutions, may require data sharing agreements and other contracts.

- These can dictate where the data must reside, how it can be transferred, and what can be done with the data both during active research phases and after a project's completion.

Research Involving the First Nations, Inuit, and Métis Peoples of Canada

- The Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2022) provides a detailed framework for approaching this research and data/knowledge in respectful ways that are beneficial to the involved communities.

- A large part of conducting this type of research involves relationship building and a slower timeline than other research projects.

- No clear-cut paths or solutions, and each project generally requires a unique approach that reflects the values and desires of the communities.

Research involving communities that have traditionally been marginalized or tokenized

- Should be approached in similar ways to research involving Indigenous communities, but there may be less formal guidance to support.

- Integrated Knowledge Translation Guiding Principles

Questions?

source