Research Data Documentation:

Codebooks, Data Dictionaries & ReadMe Files

Summer 2019

Sawyer Newman
Data Librarian for the Health Sciences

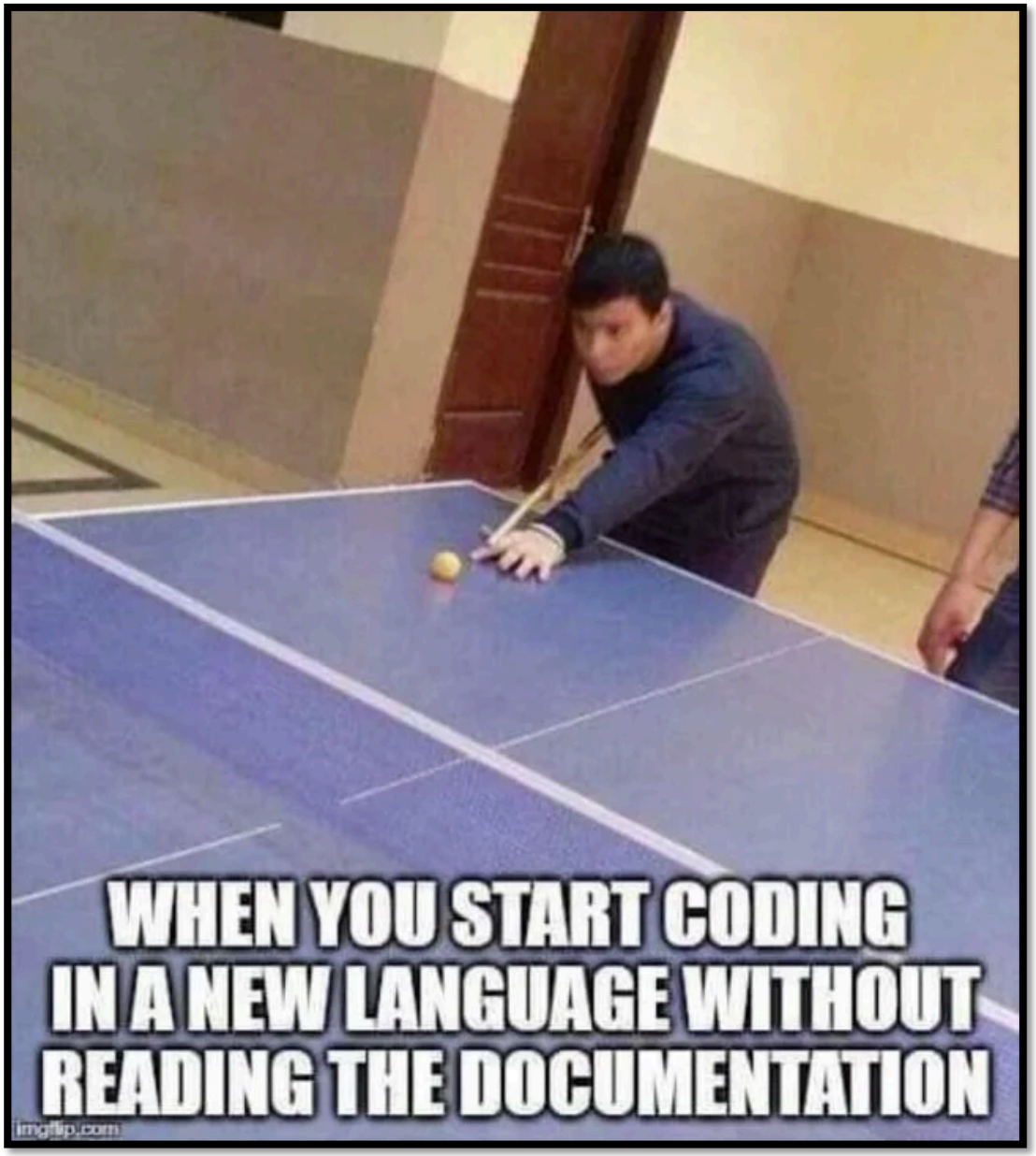
Yale *Harvey Cushing / John Hay Whitney Medical Library*

Documentation - what is it?

Information you note about your data and research

Documentation may make note of:

- Data variable meanings
- File organization structures
- Methodologies employed
- Context of your data and research
- Project contributors

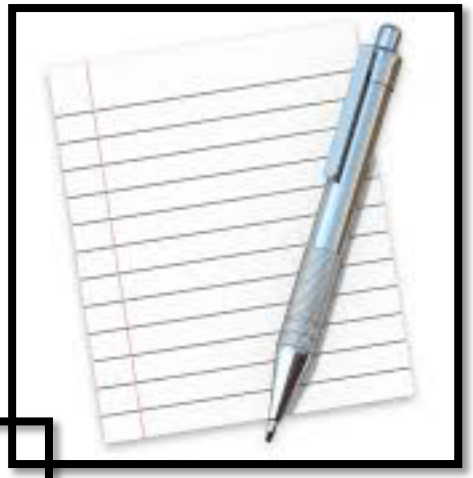

WHEN YOU START CODING
IN A NEW LANGUAGE WITHOUT
READING THE DOCUMENTATION
imgflip.com

Documentation in plain text

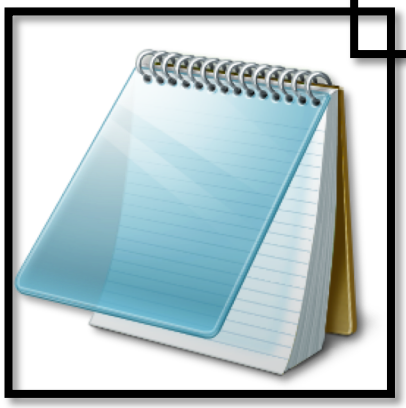
Plain text file formats are universally accessible

Formatting in plain text

- Markdown
  - LaTex
- XML | JSON



Mac TextEdit



Microsoft Notepad

Types of documentation

1. ReadMe files
2. Data dictionaries
3. Codebooks

ReadMe files

Content described	Files and folders in working directories
Where they are found	Working directories
Use	• Navigation • Instructions for using files
Format	Plain text file

Information to include in a ReadMe file

Project level ReadMe

- Organization of directory
  - Any organization rules
- Context of your project/directory
- Software & OS used
- People involved and contact info

File/subfolder level ReadMe

- Purpose of file/folder
- Interdependencies of files in a folder
- How and why files were created

ReadMe file example

Link to example

Data Dictionaries

Content described	Data files
Where they are found	Working directories (near data files)
Use	• Interpretation of data fields • Database management
Format	Spreadsheet

Data dictionary example

DATA

emlployee_id	first_name	last_name	nin	dept_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Cornelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1

DATA DICTIONARY (METADATA)

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
dept_id	int	Employee deparmtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date.

Data dictionary fields

Field	Data type	Field format	Field length	Description	Null values
last_name	char		60	Patient's last name	
gender	char	n, m, f, u	1	Patient's gender n = nonbinary m = male f = female u = unidentified	Null value indicates information was not collected
mrn	int	XXX-XXX-XXX	11	Medical record number (MRN) to serve as a patient's unique identifier	Not null
date	date	YYYY-MM-DD	10	Date of visit	

Codebooks

Content described	Research projects, data collection instruments (surveys), and collected data
Where they are found	Data repositories
Use	Interpretation of data fields and variables and data collection proceedures
Format	• PDFs (when downloaded) • Formatted plain text documents (when uploaded)

Codebook Example

```
H00034.00    [H40-SF12-2]                                    Survey Year: 2002

             SF12 - ASSESSMENT OF R'S GENERAL HEALTH

In general, would you say your health is ....

NOTE: SF-12(r) Health Survey (Medical Outcomes Trust)
(c) Medical Outcomes Trust and John E. Ware, Jr., All Rights Reserved
SF-12(tm) (QualityMetric, Inc.)

    1232        1 Excellent
    2111        2 Very Good
    1531        3 Good
     563        4 Fair
     145        5 Poor
  -------
    5582

Refusal(-1)              6
Don't Know(-2)           0
TOTAL ==========>    5588   VALID SKIP(-4)    7098     NON-INTERVIEW(-5)       0

Lead In: H00033.00[Default]
Default Next Question: H00035.00
```

Taken from: ICPSR - What is a Codebook

Information to include within a Codebook

- Variable names
- Variable labels
- Question text
- Values
- Value labels
- Summary statistics
- Missing data
- Universal skip patterns
- Notes

Variable names

15. Siblings

Variable name: **sibname1 - sibname7**

Siblings name. Information about siblings was submitted to the Pension Board when a recruit needed to prove his age in order to receive an age-dependent pension. Sibling names were collected from family Bibles and other sources. If the Pension Board conducted a census search, the generated document also contained siblings names and ages. Sibling names were also extracted from affidavits and depositions. This variable was cleaned according to the rules for names (see General Information, V.A.2). Comments included the relationship of the sibling to the recruit, especially in the cases when it was a step- or half-sibling, as well as dates and places. SIS and BRO were expanded to SISTER and BROTHER, and 1/2 was changed to HALF.

ILTOT31 – Illegal Activities – Wave 3

The total score was calculated by taking the mean of the z-scores of the following items: ril2ar, ril4ar, ril6ar, ril7ar, ril8ar, ril11ar, ril13ar, ril14ar, ril15ar, ril17ar, ril22ar. Eight of the 11 items need valid responses for a score to be calculated. To address the skewed distribution of the scale, a transformed score was computed by adding 1 to the mean and taking the natural log of that value.

Summary statistics

w1e47. Does R like or dislike Joe Biden

w1e47	Frequency	Cumulative Frequency
-7. No answer	11	11
-6. Not asked, unit non-response	2553	2564
-5. Not asked, terminated	63	2627
-4. Error, see documentation	1	2628
1. Like	240	2868
2. Dislike	209	3077
3. Neither like nor dislike	1163	4240

The following variable, MTHOTHW2, was recoded to SAS missing if a respondent indicated past month methamphetamine use (CPNMTHMN=1) and either of these situations occurred: (1) the respondent indicated that during the past 30 days they got methamphetamine in "some other way" (MTHOTHWY=1) and their write-in response was classified as an invalid or unknown source (MTHOTHS2=97), or (2) a response to getting methamphetamine "some other way" was unknown (MTHOTHWY=93,94,97,98).

MTHOTHW2	1	GOT MTH IN SOME OTHER WAY - PST MON		
		. = Unknown Source (See comment above).....	20	0.04
		1 = Yes (CPNMTHMN=1 and MTHOTHWY=1 and MTHOTHS2=10)...	6	0.01
		2 = No (CPNMTHMN=1 and MTHOTHWY=6).....	96	0.17
		3 = No Past Month Use (CPNMTHMN=0).....	55650	99.78

Codebook Creation Tools

- R, Python, SPSS, Stata & SAS
- Nesstar Publisher
- Colectica

Codebook Example

The National Survey of Fertility Barriers, 2010 [United States] (ICPSR 36902)

link

Research Data Services @ CWML

Data Management Plans

Data Tools & Software

Data Policy Guidance

Find Datasets

Data Storage

Best Practices & Definitions

Data Support Groups at Yale

Consultations & Drop-Ins

medicaldata@yale.edu

Resources

- DDI - Create a Codebook
- Cornell Guide to writing a ReadMe file for research data
- AHIMA Data Dictionary
- LaTex