Research Data Documentation:

Codebooks, Data Dictionaries & ReadMe Files

Yale UNIVERSITY LIBRARY
Harvey Cushing/John Hay Whitney Medical Library

Documentation – what is it?

Documentation is 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

Why should you spend time on documentation?

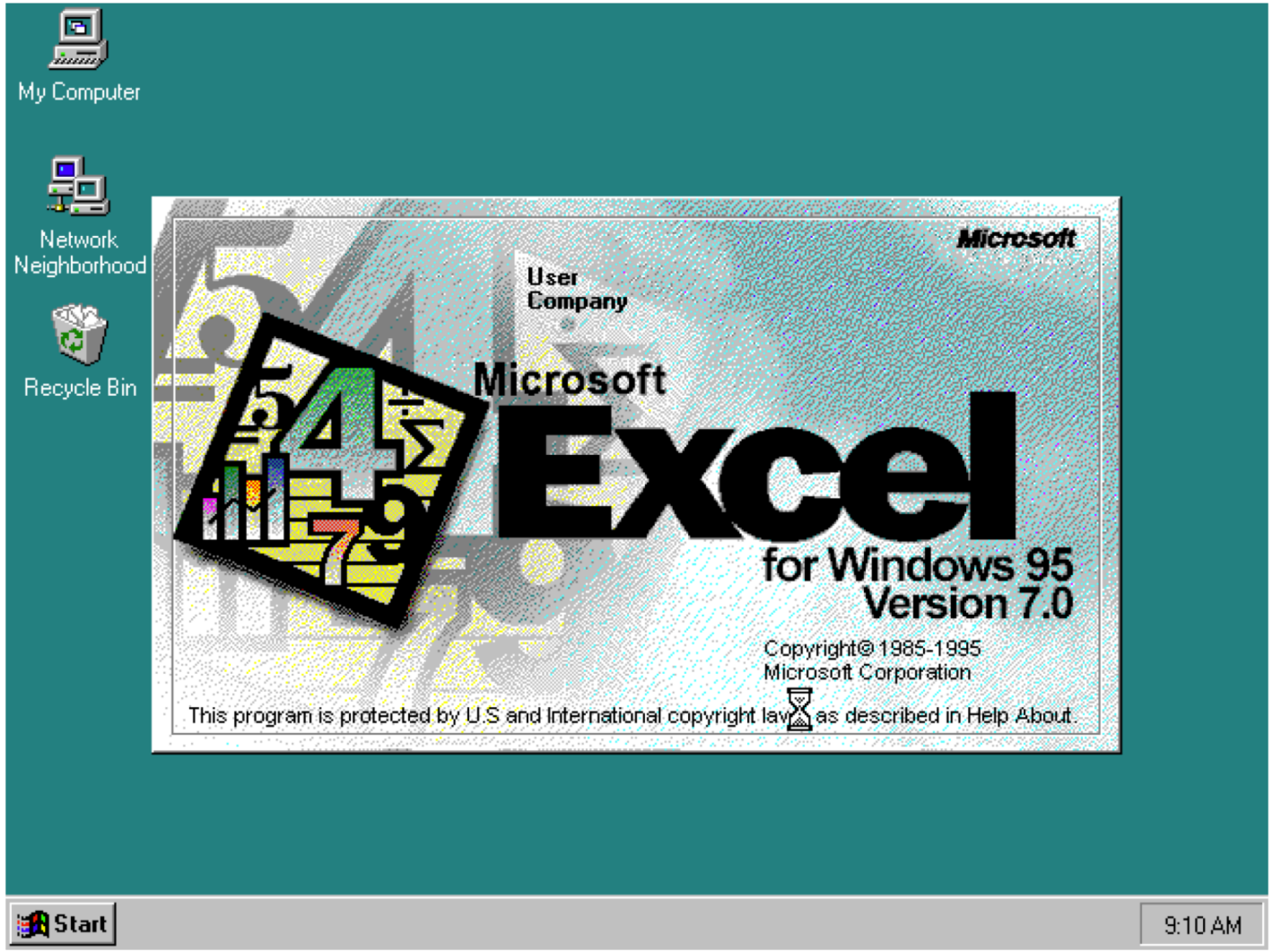
- Help others understand what your spreadsheet data means
- Keep notes about how your files are organized – prevent yourself from loosing files
- Track how files are related
- Help your future self understand your research projects
- In the long run, documentation saves time

Documentation in Plain Text

Why use plain text files?

- Plain text file formats are universally accessible across machines and across time

Could you open a file saved by this machine?



How do you create plain text files?

Text editors

- Mac: TextEdit
- Windows: Microsoft Notepad
- Sublime, Brackets, etc.

Exportation from proprietary software

- Word: save file as .txt
- Excel: save file as .csv

Types of Documentation

ReadMe Files, Data Dictionaries, & Codebooks

Yale UNIVERSITY LIBRARY
Harvey Cushing/John Hay Whitney Medical Library

Types of research documentation

ReadMe Files

Describe local files and folders

Data Dictionaries

Describe data in spreadsheets or databases

Codebooks

Describe survey design, variable meanings, and coded data

ReadMe Files

What information do the files in this folder contain?

Yale UNIVERSITY LIBRARY
Harvey Cushing/John Hay Whitney Medical Library

ReadMe files

Content described	Files and folders in working directories
Where to create them	Working directories (i.e. active working folders)
Use	• Navigation • Instructions for using files
Format	Plain text files

Information to include in a ReadMe file

Project Level ReadMe

- Organization of a directory
- 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

https://github.com/sauuyer/Med-Lib-Data-Classes/blob/master/data-documentation/README_template_example.txt

Data Dictionaries

What does that column header mean?

Yale UNIVERSITY LIBRARY
Harvey Cushing/John Hay Whitney Medical Library

Data Dictionaries

Content described	Spreadsheets
Where to create them	Working directories (near the spreadsheets they describe)
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	Length	Description	Nulls
date	date	YYYY-MM-DD	10	Date of exam	Not null
gender	char	m, f, n	1	Patient's gender m = male f = female n = nonbinary	Null value indicates information not collected
mrn	int	X-XXXX	6	Medical record number (MRN) to serve a a patient's uid	Not null

Codebooks

What does this survey response mean?

Codebooks

Content described	Research projects, data collection instruments and collected data
Where to create them	Data repositories
Use	• Interpretation of data fields and variables • Interpretation of data collection proceedures
Format	• PDFs (when downloaded)

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
```

Information to include within a Codebook

- Variable names
- Question text
- Summary statistics
- Missing data
- 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

Wrap up

Where should you look for additional information and tools on documentation?

Tools for documentation creation

DDI's Create a Codebook

- https://ddialliance.org/training/getting-started-new-content/create-a-codebook

Best Practices for Research Data Management

- https://library.medicine.yale.edu/research-data/best-practices

Cornell Guide to Writing a ReadMe file for research data

- https://data.research.cornell.edu/content/readme

Data Documentation Initiative (DDI)

- DDI is an international standard for describing research data

ddialliance.org

Yale UNIVERSITY LIBRARY

Harvey Cushing/John Hay Whitney Medical Library