Unravelling Data: R for Social Science

Marion Walton, Centre for Film & Media Studies, University of Cape Town

Jan 2025

From dataset to data settings

Hello!

Marion Walton (she/her)

```
selection <- "English"
language <- c("English","Afrikaans","Kaaps", "isiZulu", "isiXhosa")
greeting <- c("Hello","Hallo","Aweh","Sawubona","Molweni")
pairs.df <- data.frame(language,greeting)
result <- pairs.df$greeting[pairs.df$language == selection]
print(result)
```

```
## [1] "Hello"
```

About Me

- Associate Professor, Centre for Film & Media Studies, University of Cape Town.
- CODATA-RDA 2024 Alumna 😊
- Carpentries instructor - https://carpentries.org/
- Hons/MA courses in Digital Methods for Social Media Research
- Lesson - Data Carpentry for Media Research (Pre-alpha)
- marion.walton@uct.ac.za
- https://aoir.social/@marionwalton

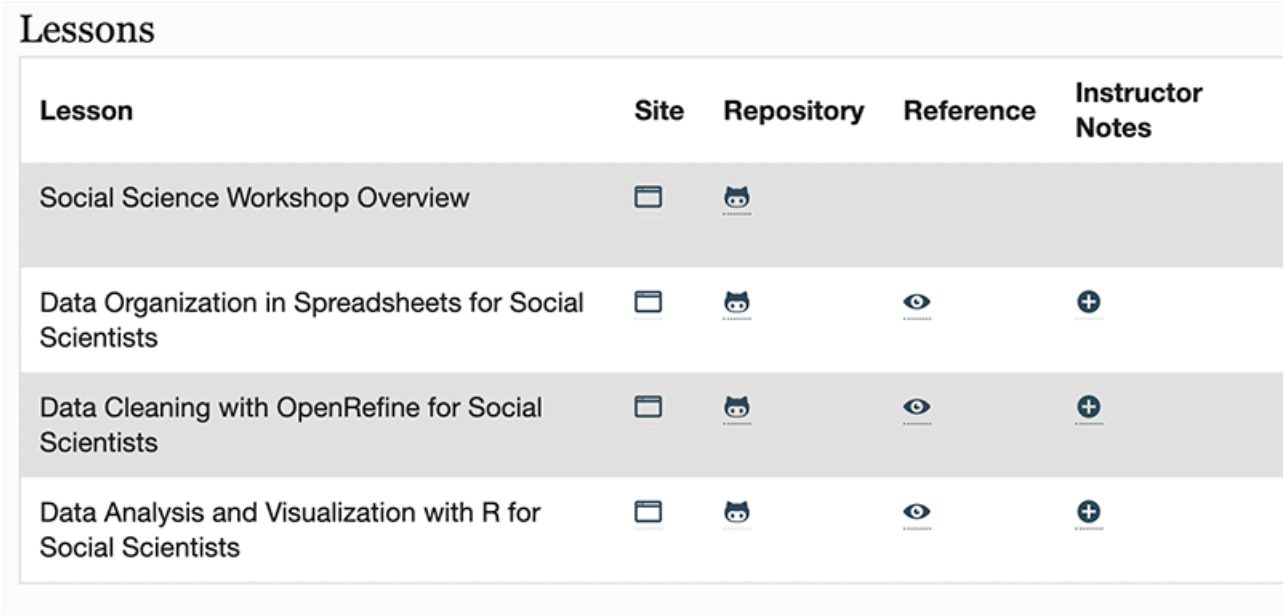
What is Data carpentry (DC)?

“Data Carpentry is a lesson program within The Carpentries building communities teaching universal data skills.”

https://carpentries.org/

“Data carpentry workshops for a **biological/ecological curriculum** were later adapted for a range of different scientific disciplines and datasets, including one designed for **social scientists** (Teal et al., 2015 emphasis added).

Data carpentry for social science curriculum



Lessons

Lesson	Site	Repository	Reference	Instructor Notes
Social Science Workshop Overview				
Data Organization in Spreadsheets for Social Scientists				
Data Cleaning with OpenRefine for Social Scientists				
Data Analysis and Visualization with R for Social Scientists				

. What are the “foundations” of our work with data in social science and humanities?

Foundations

Data ethics and data justice are just as “foundational” as spreadsheets and dataframes, particularly in the Global South and decolonial contexts.

Afro-feminist perspectives - **Relational approach** to data

“knowing is an activity that happens in the relationship between the knower and the known” (De Jaegher in Birhane, 2021)

Data settings and data sets

Relational ethics - All datasets constructed in complex, rich **social settings**

“Using data ethically is a local problem, which requires attention to the differences among data settings and how they might change over time, necessitating continued maintenance and adjustment.” (Loukissas, 2019:194)

Explore construction -> Read, Inquire, Unfold, Represent

Loukissas, Y. A. (2019). All data are local: Thinking critically in a data-driven society. The MIT Press.

About the dataset

SAFI_clean.csv

“The data used in these lessons are taken from interviews of farmers in two countries in eastern sub-Saharan Africa (Mozambique and Tanzania). These interviews were conducted between November 2016 and June 2017 and probed household features (e.g. construction materials used, number of household members), agricultural practices (e.g. water usage), and assets (e.g. number and types of livestock).”

The Carpentries. 2023. Data Organization in Spreadsheets for Social Scientists.

Meet the farmers

Farmer led irrigation development in Africa, example from Mozambique

Reflect on your own responses

What did you feel about the video?

- Do you know anyone who does the kinds of farming work featured in the video?
- Have you ever grown your own food?
- How did you feel while watching the video?
- Do you know anyone who does this kind of research?
- Have you been interviewed for a research project before?

Explore motivations of research

Key social questions raised by the data:

For example:

- [Farmer-led irrigation] may increase crop yields, but at what cost? For example, does it benefit everyone in a given community, or just a few individuals? How do women and non-locals fare? Is it environmentally sustainable?(Woodhouse, 2017).
- World Bank highlights the need for small-scale farmers to adapt to climate change (rather than e.g. requiring an end to fossil fuel production).

Woodhouse, P. (2017, January 24). Invisible irrigators: How small-scale Tanzanian farmers are making a difference. The Conversation.

Opportunities

Make your voice heard in open data:

- Ask questions!
- Contribute to lesson development
- Adapt language and exercises
- New lessons and disciplines?

Conclusions

Foundational skill for Social Science - Build and reflect on the relationship between the knower and the known.

“treat data as a point of contact, a landing, an opportunity to get closer, to learn to care about a subject, or the people and places beyond data. Do not mistake the availability of data as permission to remain at a distance.” (Loukissas, 2019:196)

Loukissas, Y. A. (2019). All data are local: Thinking critically in a data-driven society. The MIT Press.

Thank you!

marion.walton@uct.ac.za

https://aoir.social/@Marionwalton

Selected references

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. Patterns, 2(2), 100205.

Loukissas, Y. A. (2019). All data are local: Thinking critically in a data-driven society. The MIT Press.

Woodhouse, P. (2017, January 24). Invisible irrigators: How small-scale Tanzanian farmers are making a difference. The Conversation.

Part II

Where are the farmers?

What questions did you have about the participants after seeing the data?

A	B	C	D	E	F	G	H
key_ID	village	interview_date	no_membrs	years_liv	respondent_wall_type	rooms	memb_assoc
1	God	2016-11-17T00:00:00Z	3	4	muddaub	1	NULL
2	God	2016-11-17T00:00:00Z	7	9	muddaub	1	yes
3	God	2016-11-17T00:00:00Z	10	15	burntbricks	1	NULL
4	God	2016-11-17T00:00:00Z	7	6	burntbricks	1	NULL
5	God	2016-11-17T00:00:00Z	7	40	burntbricks	1	NULL
6	God	2016-11-17T00:00:00Z	3	3	muddaub	1	NULL
7	God	2016-11-17T00:00:00Z	6	38	muddaub	1	no
8	Chirodzo	2016-11-16T00:00:00Z	12	70	burntbricks	3	yes
9	Chirodzo	2016-11-16T00:00:00Z	8	6	burntbricks	1	no
10	Chirodzo	2016-12-16T00:00:00Z	12	23	burntbricks	5	no
11	God	2016-11-21T00:00:00Z	6	20	sunbricks	1	NULL
12	God	2016-11-21T00:00:00Z	7	20	burntbricks	3	yes
13	God	2016-11-21T00:00:00Z	6	8	burntbricks	1	no
14	God	2016-11-21T00:00:00Z	10	20	burntbricks	3	NULL
15	God	2016-11-21T00:00:00Z	5	30	sunbricks	2	yes
16	God	2016-11-24T00:00:00Z	6	47	muddaub	1	NULL
17	God	2016-11-21T00:00:00Z	8	20	sunbricks	1	NULL

SAFI_clean.csv

“Tidy” data?

Rows and columns of ‘tidy’ **machine-readable** data look deceptively simple, yet there are many complex issues to address when we use data to represent people, or to make decisions that can affect people’s lives and our societies.

Missing context

“the data scientist decides what is worth measuring (making some things visible and others invisible by default) and how. In the process of data cleaning, rich information that provides context about which data are collected and how datasets are structured is stripped away.”

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. Patterns, 2(2), 100205.

Why does the data look this way?

- Review the lesson datasets and find out more about the SAFI project
- Metadata for table used in R for Social Sciences lesson
- Who/where are the people? (Project video Farmer led irrigation development in Africa, example from Mozambique)
- Why was this data collected? (Read press coverage and papers)
- Who are the researchers? (Email them!)