rbtl - Research Beyond the Lab

Research Data Management

Lars Schöbitz

2022-03-23

Today

1. Homework Assignment 4 - Solutions
2. Week 5 - Learning Objectives
3. Research Data Management
4. Data Organisation in Spreadsheets
5. Homework Assignment 5

Homework Assignment 4

Think, Pair, Share

1. What is the difference between markdown and R Markdown?
2. How does a file written in markdown differ from a file in a proprietary file fornat (e.g. .docx)?

- **Think** for 2 minutes
- **Pair** with your neighbour for 4 minutes
- **Share** your answer with the class

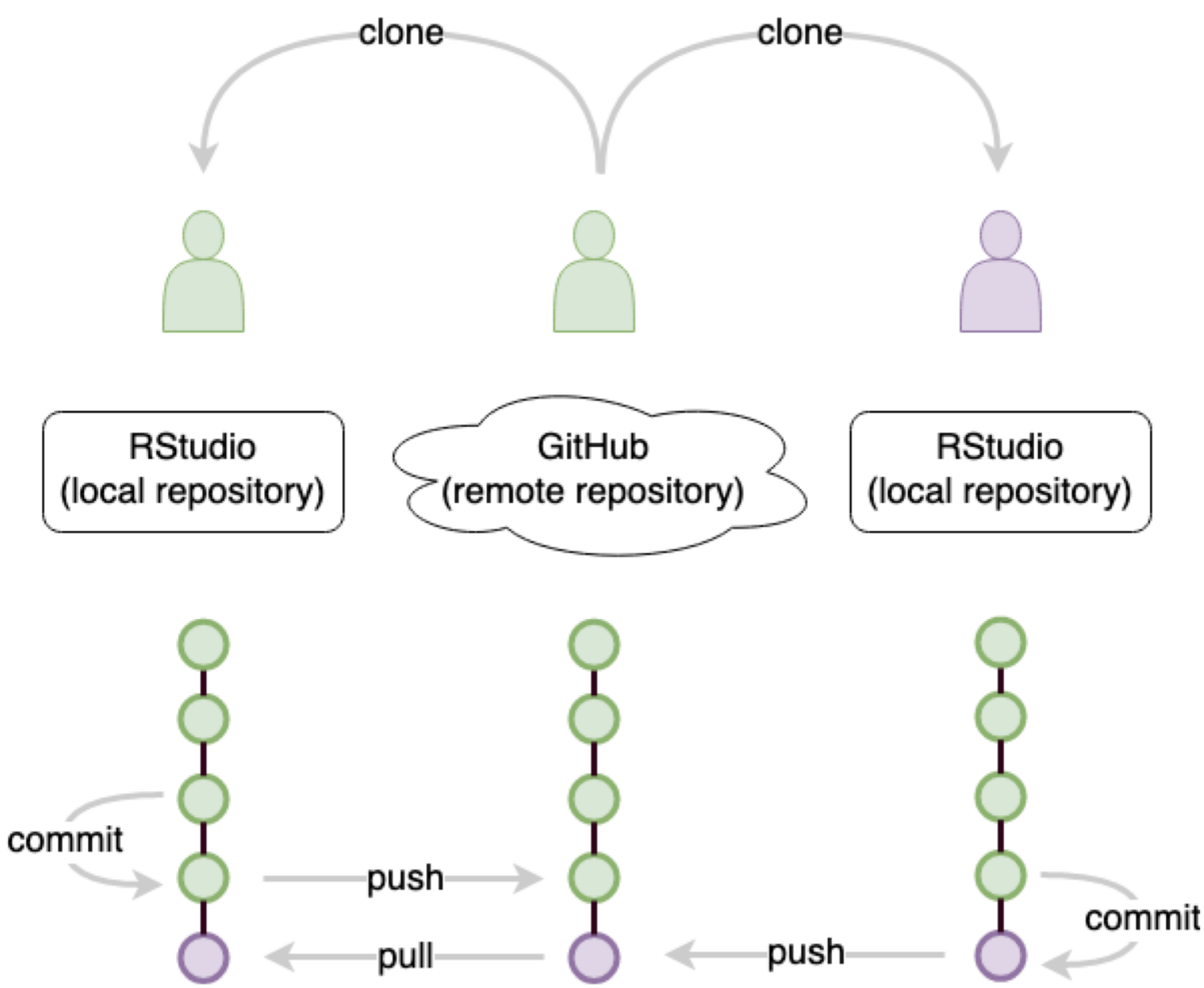
02:00

Homework Assignment 4 - Solutions

main · 1 branch · 0 tags

larnsce Added solutions file for ae-04-rmarkdown — 76c1239 1 minute ago — 2 commits

.gitignore	Initial commit	5 days ago
LICENSE	Initial commit	5 days ago
README.md	Initial commit	5 days ago
ae-04-rmarkdown-solutions.Rmd	Added solutions file for ae-04-rmarkdown	1 minute ago
ae-04-rmarkdown.Rmd	Initial commit	5 days ago
ae-04-rmarkdown.Rproj	Initial commit	5 days ago


clone
clone
RStudio
(local repository)
GitHub
(remote repository)
RStudio
(local repository)
commit
push
pull
push
commit

Week 5 - Learning Objectives

Week 5 - Learning Objectives

1. Learners can use a tool for survey design for 12 to 15 questions ~~using at least two elements of survey logic~~

2. Learners can apply 12 principles for data organisation in spreadsheets in the layout of a collected dataset

3. Learners understand the importance of documentation and metadata for (research) data management

Research Data Management

Research Data Management

> Acronym: RDM. Refers to the organisation, storage and preservation of data created during a research project. It covers initial planning, day-to-day processes and long-term archiving and sharing. Shortened to RDM. - Reference

Research Data Management

- ensures efficiency in research workflows
- enables greater reach and impact (FAIR principles)
- prevents loss and data corruption
- documentation and metadata ensures access and enables reuse for others

The FAIR principles

The Turing Way Community, & Scriberia. (2021). Illustrations from the Turing Way book dashes. Zenodo. https://doi.org/10.5281/zenodo.5706310

Research Data Lifecycle

Research Data Lifecycle

Reference: https://www.youtube.com/watch?v=-wjFMMQD3UA

Data Management Plan

Data Management Plan

Provides information on five main topics:

1. Roles and Responsibilities
2. Type and size of data collected and documentation/metadata generated
3. Type of data storage used and back up procedures that are in place
4. Preservation of the research outputs after the project
5. Reuse of your research outputs by others

Reference: https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-dmp.html

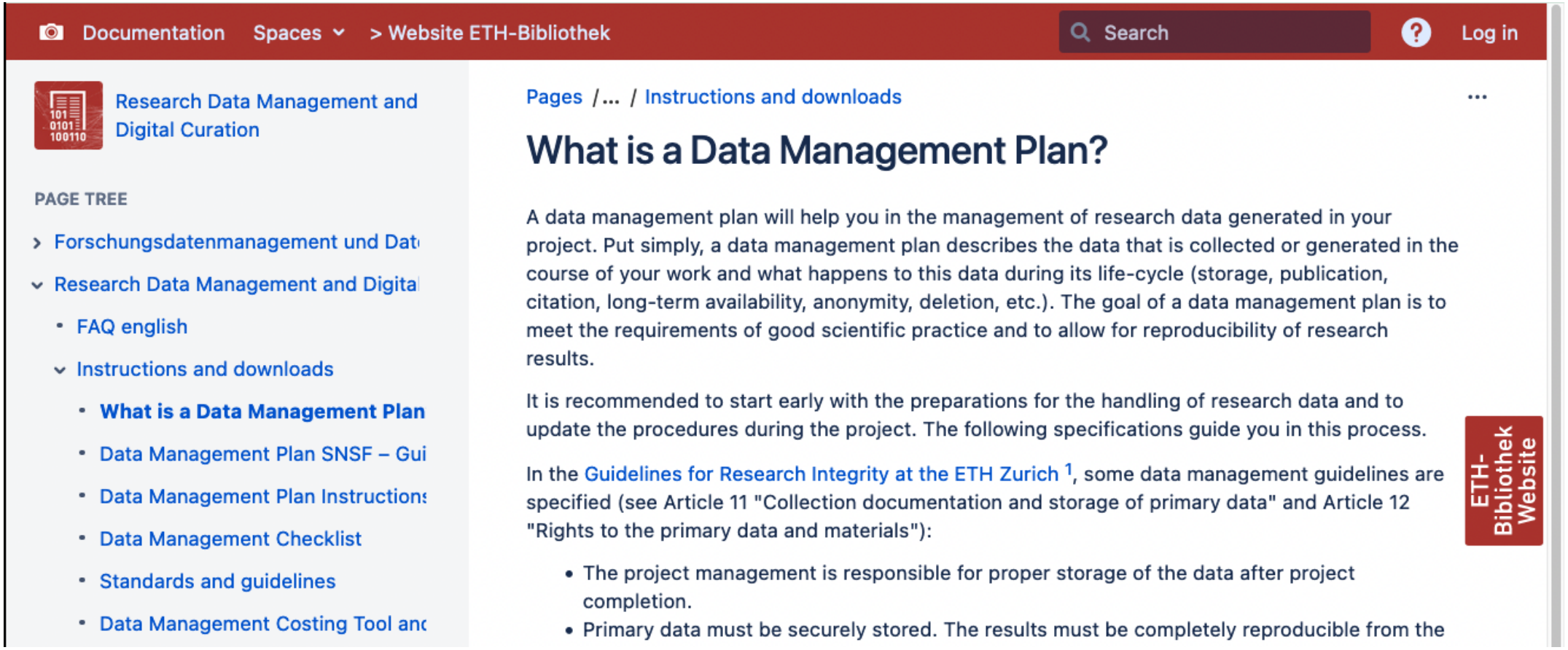
Data Management Plan

Provides information on five main topics:

1. Roles and Responsibilities
2. Type and size of data collected and **documentation/metadata** generated
3. Type of data storage used and back up procedures that are in place
4. Preservation of the research outputs after the project
5. Reuse of your research outputs by others

Reference: https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-dmp.html

Data Management Plan



Screenshot, 2022-03-22: ETH Library - Research Data Management and Digital Curation - What is a Data Management Plan?

Documentation and Metadata

You got data. Is it enough?

Erika Berenguer
@Erika_Berenguer

Replying to @Erika_Berenguer

@tomjwebb I see tons of spreadsheets that i don't understand anything (or the stduent), making it really hard to share.

4:32 PM · Jan 16, 2015

1 Reply Copy link to Tweet

Explore what's happening on Twitter

Sven Kochmann
@indianalytics

Replying to @tomjwebb

@tomjwebb @ScientificData "Document. Everything." Data without documentation has no value.

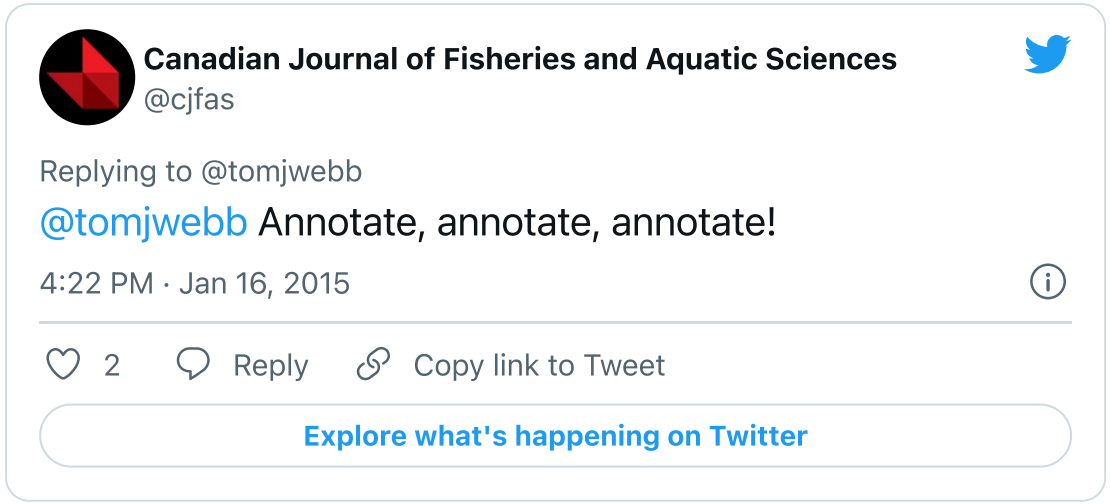
5:08 PM · Jan 16, 2015

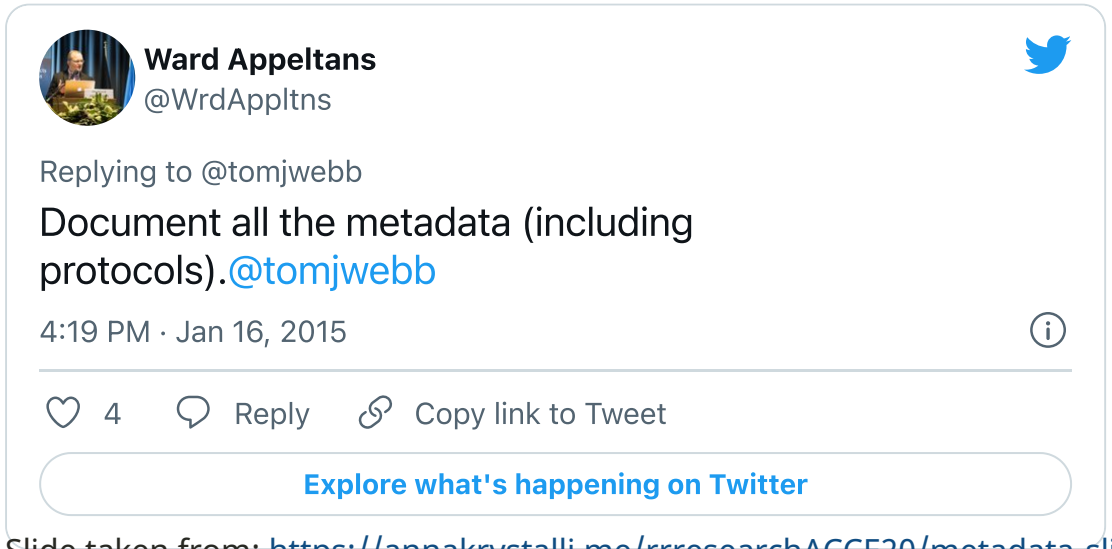
1 Reply Copy link to Tweet

Explore what's happening on Twitter

Slide taken from: https://annakrystalli.me/rrresearchACCE20/metadata-slides.html#metadata-slides

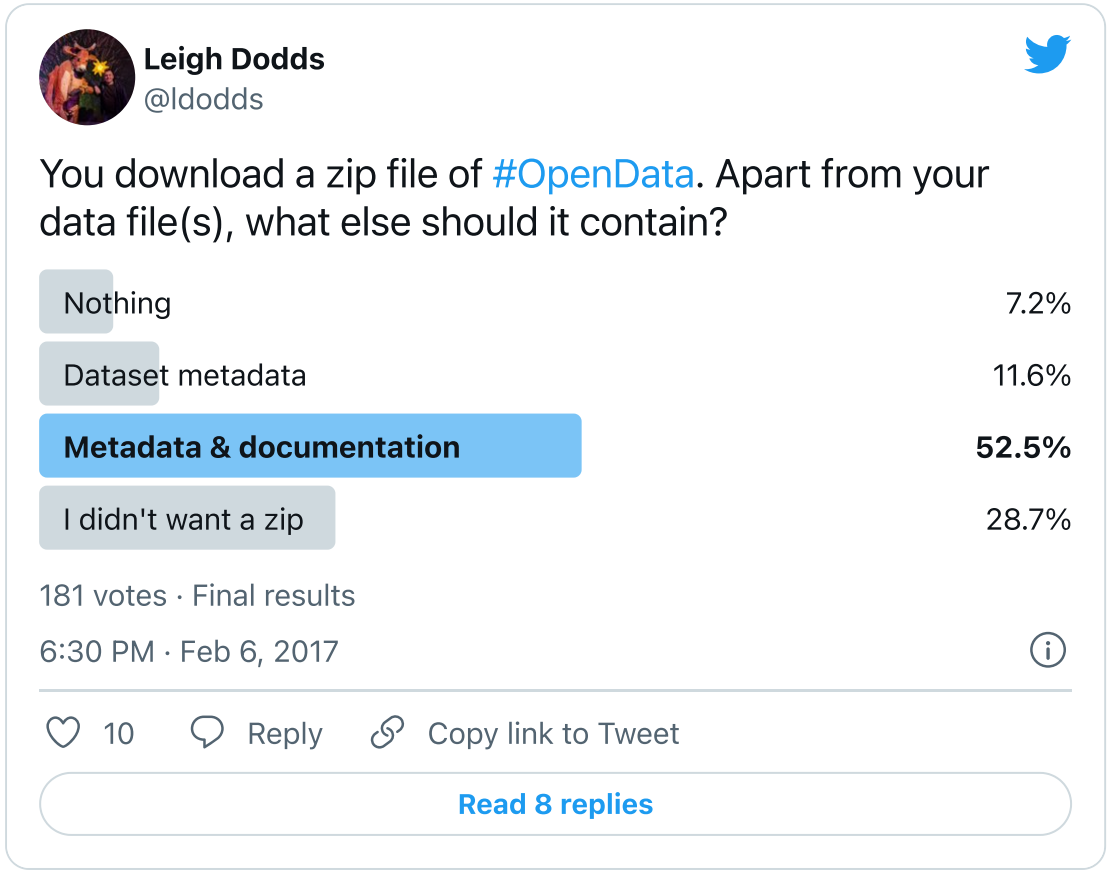
You got data. Is it enough?





Slide taken from: https://annakrystalli.me/rrresearchACCE20/metadata-slides.html#metadata-slides

Think, Pair, Share



Slide taken from: https://annakrystalli.me/rrresearchACCE20/metadata-slides.html#metadata-slides

Think, Pair, Share

Thought experiment: Imagine a dream open data set

1. How would you locate it?
2. What details would you need to know to determine relevance?
3. What information would you need to know to use it?

- **Think** for 2 minutes
- **Pair** with your neighbour for 4 minutes
- **Share** your answer with the class

02:00

Slide adapted from: https://annakrystalli.me/rrresearchACCE20/metadata-slides.html#metadata-slides

ETH library

Data Management Plan Instructions for ETH Zurich Researchers

Section 1: Data collection and documentation

- 1.3 What documentation and metadata will you provide with the data?
  - What information about your data (i.e., metadata) is required to make reuse of your data in the future?
  - Are you using certain community standards for the annotation of metadata?
  - How will data documentation be carried out?

Q1: What information about your data (i.e., metadata) is required to make reuse of your data in the future?

- Codebook or data dictionaries that define and explain all variables in your data
- A human and machine-readable file (e.g. README.md) that contains general information on:
  - Title of the dataset
  - Description of the dataset
  - Author information (Name, Institution, Address, Email)
  - Date of data collection (begin, end)
  - Geographic location of data collection
  - etc.
- A license to clarify reuse possibilties

Q2: Are you using certain community standards for the annotation of metadata?

- General metadata standards (e.g. DublinCore, schema.org, R3data.org)
- Domain specific standards (**we won't get into this**)

Q3: How will data documentation be carried out?

- Repository hosted on GitHub with license that clarifies use: CC-BY-4.0
- Respository has `/data` directory which contains:
  - **Codebook** (stored as `attributes.csv`) that describes all variables following schema.org standards
  - **README** hat contains information for each dataset, following variation of Cornell University README template

Let's take a break

10:00

Photo by: Blake Wisz

Data Collection Tools

Data Collection Tools

- Questionnaires for survey based data collection
- Spreadsheets for experimental data collection

Surveys

Surveys

Commonly used in our sector

- KOBO Toolbox
- mWater
- OpenDataKit

ETH tool - Select Survey

- Select Survey Portal https://selectsurvey.ethz.ch/
- Select Survey User Manual
- Select Survey IT Knowledge Base

Live Demonstration - Select Survey

Task - Fill out questionnaire

- Open Slack and find the link to the questionnaire we have just created
- Fill out the questionnaire

Live Demonstration - Writing a codebook

Let's take a break

10:00

Photo by: Blake Wisz

Data Organisation in Spreadsheets

Data Organisation in Spreadsheets

Screenshot taken on 2022-03-23

Data Organisation in Spreadsheets

Read the paper (it's part of your homework), but you can also:

- Go through the annoated slides: https://kbroman.org/Talk_DataOrg/dataorg_notes.pdf
- Watch Karl Broman give the talk (02:36 to 45:00): https://youtu.be/t74E0a90gkA?t=156
- Read the content on a website: https://kbroman.org/dataorg/

But, especially apply it to your data

Why?

Because it will make your life easier!

Jorunal: The American Statistician, Screenshot taken on 2022-03-23

License? CC0 (!)

README.md

Data organization in spreadsheets

Slides for a talk for the OSGA Webinar Series, on 24 Sept 2021, based on my paper of the same title with Kara Woo. Also see the related website.

PDF of slides: https://kbroman.org/Talk_DataOrg/dataorg.pdf

PDF of slides with notes: https://kbroman.org/Talk_DataOrg/dataorg_notes.pdf

Video of presentation: https://youtu.be/t74E0a90gkA

License

To the extent possible under law, Karl Broman has waived all copyright and related or neighboring rights to "Data organization in spreadsheets". This work is published from the United States.

Screenshot taken on 2022-03-23

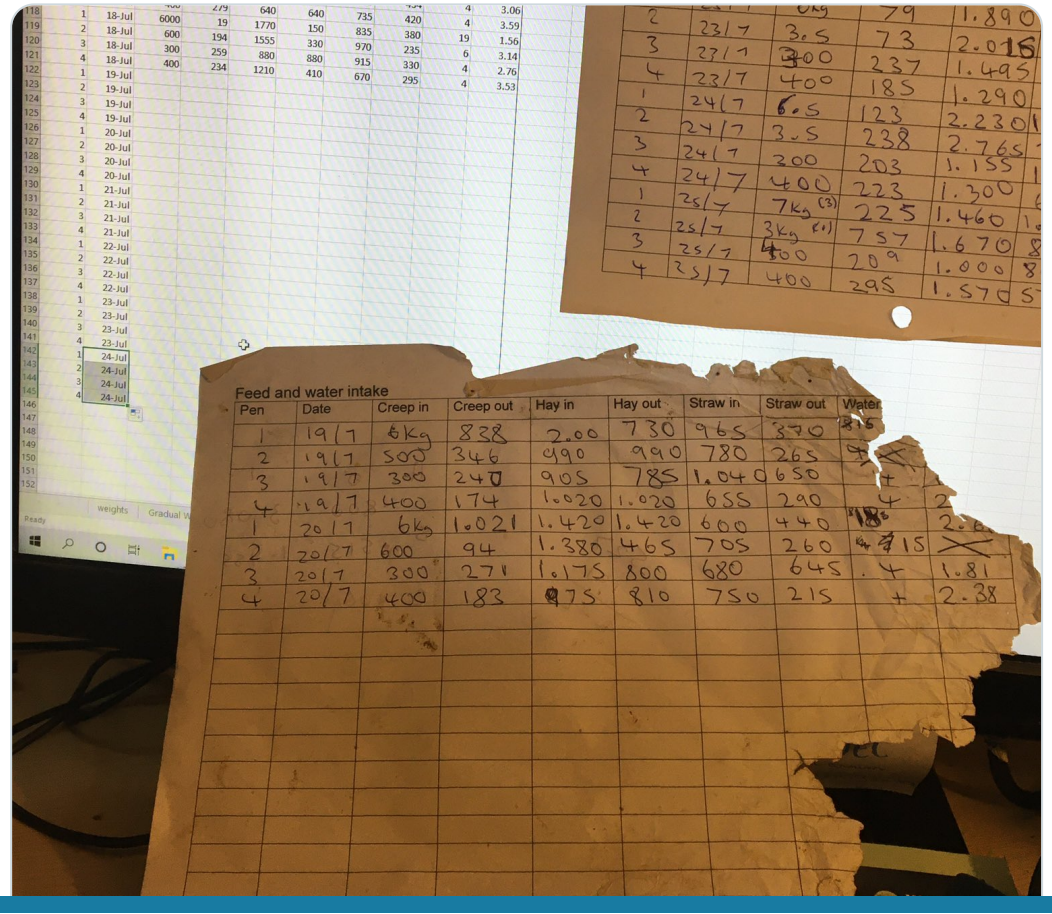
Waste Characterisation - Data example

bin_id	collection_date	waste_category	weight
234	2022-03-22	plastics	2.5
234	2022-03-22	metal	1.7
234	2022-03-22	pet	0.8
234	2022-03-22	other	15.0
682	2022-03-22	plastics	6.5
682	2022-03-22	metal	1.1
682	2022-03-22	pet	7.8
682	2022-03-22	other	9.1

Waste Characterisation - How to collect your data?

Holly Vickery
@SkylarkHolly

Working with goats 🙄🤦‍♀️



Assignment 5

Assignment 5

Submit via GitHub by 2022-03-29 23:59

1. ~~Groups~~ Students will program their revised questionnaire onto a digital platform so it is ready to use for the following class

2. ~~Groups~~ Students will transfer their variables into a spreadsheet database

3. ~~Groups~~ Students will write a codebook for their spreadsheet database

4. ~~Groups~~ Students will write a README for documentation of metadata

All details on website: https://rbtl-fs22.github.io/website/am-05-data-management.html

Final excursion - File naming conventions

File naming conventions

Why?

- to stay organised
- to not end up with: `'text is FINAL_v2-Ls.docx'`

File naming conventions - Directories and files

Rule 1: Avoid capital letters

`project/rbtl/website/data/`

(if you want them, be consistent)

File naming conventions - Directories and files

Rule 2: Avoid empty spaces

`project/rbtl/website/data/tab-01-rbtl-assignments.csv`

File naming conventions - Directories and files

Rule 3: Use dash '-' to connect text strings in file names

`tab-01-rbtl-assignments.csv`

File naming conventions - Directories and files

Rule 4: Avoid special characters (e.g %, £)

File naming conventions - Directories and files

Rule 5: Use ISO 8601 date format: YYYY-MM-DD

`2022-03-21-waste-characterization-erz.csv`

File naming conventions - Directories and files

Rule 6: Use names that are descriptive for the content

`2022-03-21-waste-characterization-erz.csv`

Thanks! 🌻

Slides created via the R packages:

xaringan
gadenbuie/xaringanthemer

The chakra comes from remark.js, **knitr**, and R Markdown.

Access slides as PDF on GitHub