

Hello, I am a Data Steward

Research Software Engineering and Data Stewardship
Career Talk

Lars Schöbitz

Global Health Engineering - ETH Zurich

Prof. Elizabeth Tilley

Global Health Engineering - ETH Zurich

June 17, 2025



Meet a data steward

Meet a data steward

I have:

- 10+ years work experience (5 in research, at Eawag)
- empathy, compassion, patience, persistence
- an affinity for IT
- teaching experience
- learned how people learn

I don't have:

- a doctoral degree
- a qualification in computer science
- a qualification in statistics
- a lot of time



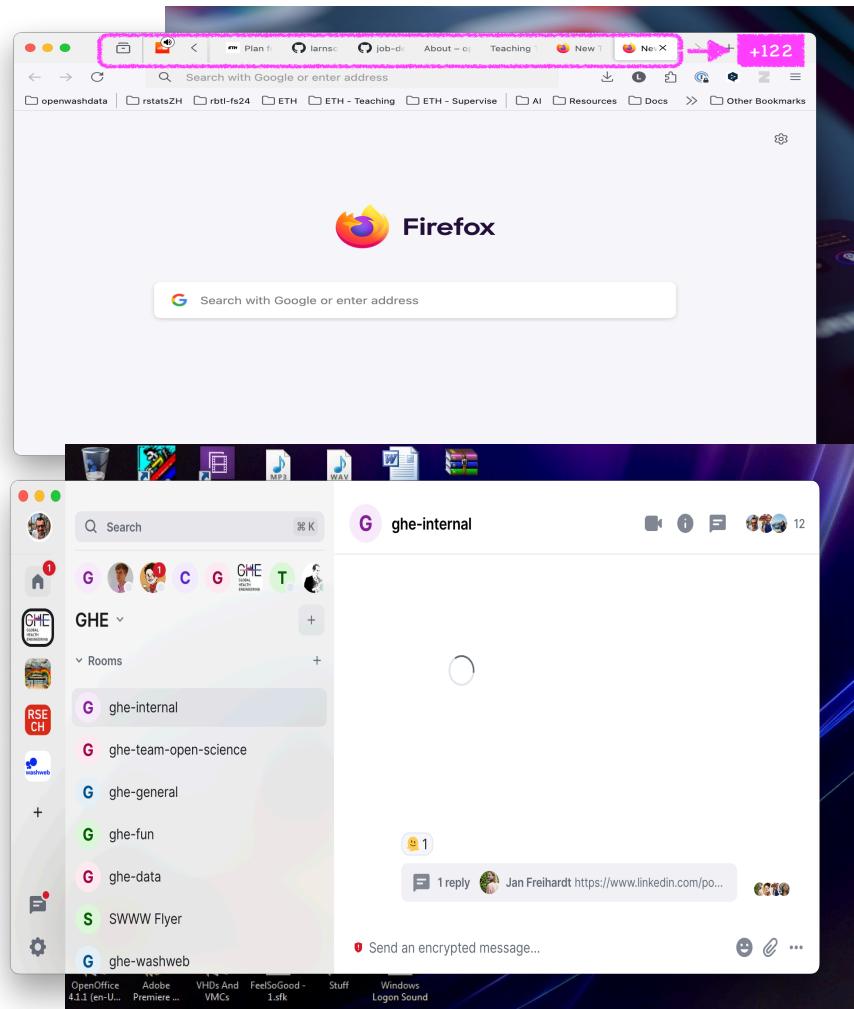
8 learnings from 4
years



#1 Technology is not on
our side

The Modern Academic's Challenges

- Overflowing email inboxes
- Browsers with hundreds of tabs
- Files stored on Desktops
- MS Teams, Slack, Element, NAS, Google Drive, ...
- Credentials, Passwords, OTPs, 2FAs, PATs, ...





#2 ETH wants
reproducibility

ETH RDM Guidelines

4 The structure and the processing steps of all *Research Data* must be digitally documented in order to ensure adherence to the *FAIR principles*. Where documentation includes a lab journal, Electronic Laboratory Notebooks (ELN) are recommended.

Art. 6 Publication of *Research Data* and *Programming Code*

¹ Publication

a. *Research Data and Programming Code* that are considered as directly relevant for a result publication based on *Community Standards* must be published and deposited in a FAIR repository along with rich, openly available Metadata.

(i) If there are limitations for sharing relevant raw data online because sharing is technically or economically not feasible, FAIR allows for publishing *Metadata* only which contain information on how raw data can be accessed if necessary.

(ii) In the case of long-range data collection projects, the *Research Data* and *Programming Code* that are relevant for a result publication may be defined as a subset and may be aggregated.

FAIR data sharing principles

The screenshot shows a web browser window with a purple header bar. The title bar reads "The FAIR Guiding Principles for scientific data management and stewardship". The address bar shows the URL "nature.com/articles/sdata201618". The main content area features the title "scientific data" in large bold letters, followed by navigation links "View all journals", "Search", and "Log in". Below this is a menu with "Explore content", "About the journal", and "Publish with us". The breadcrumb navigation shows "nature > scientific data > comment > article". On the right side, there is a blue button labeled "Download PDF". At the bottom left, it says "Comment | Open access | Published: 15 March 2016". The main title of the article is "The FAIR Guiding Principles for scientific data management and stewardship". The authors listed are Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim



FAIR data sharing principles

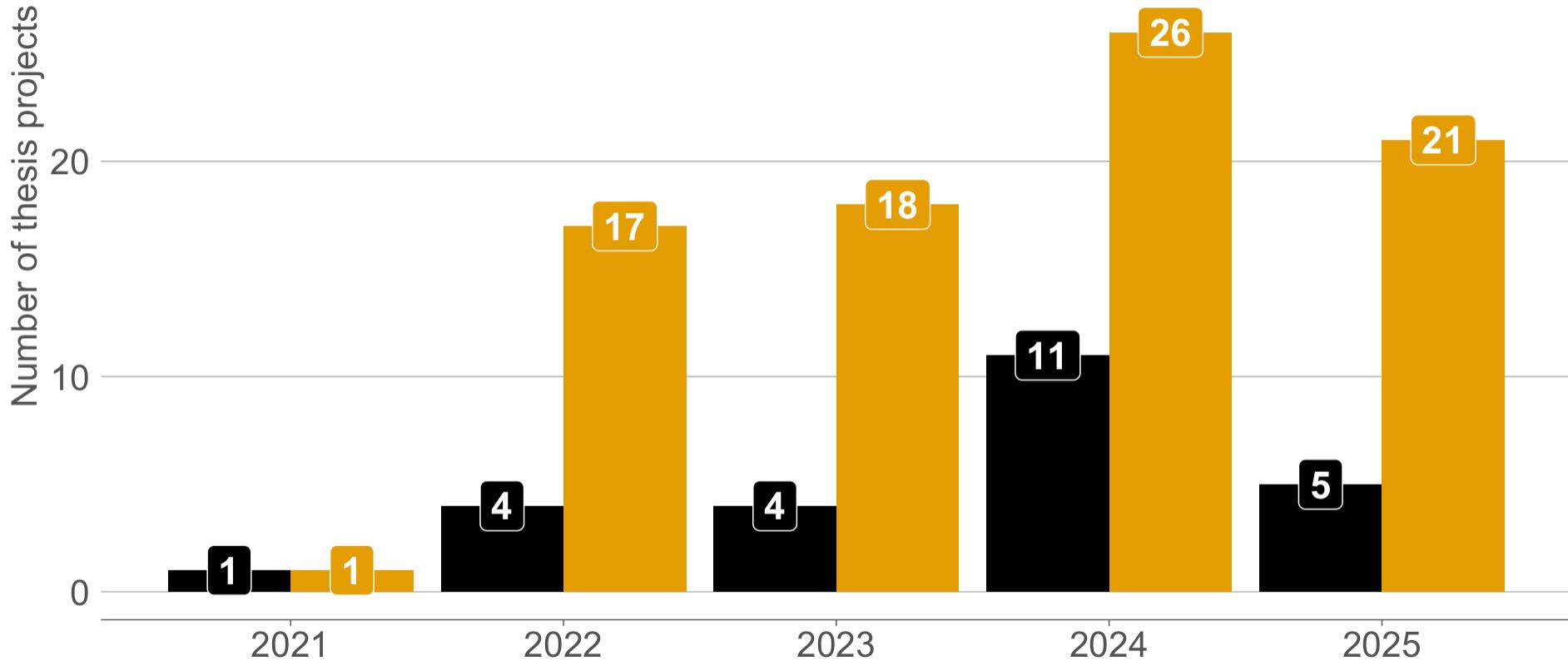
- Technical in nature
- Require data management strategy to establish workflows
- Not a checkbox, but a process

Findable
Accessible
Interoperable
Reusable



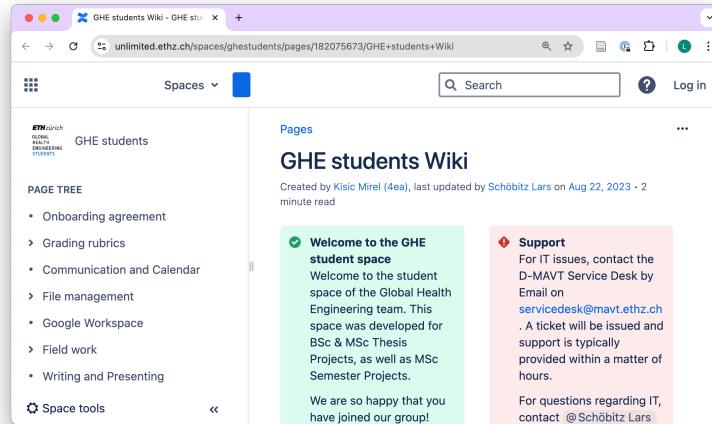
#3 Data management is
project management

Project: ■ BSc thesis ■ MSc thesis



GHE Student Wiki (public)

- Grading criteria
- Communication expectations
- Data storage and data management guidelines
- Presentation standards
- Proposal and thesis writing requirements



Grading rubric & data publication

Four areas of evaluation with 31 sub-areas

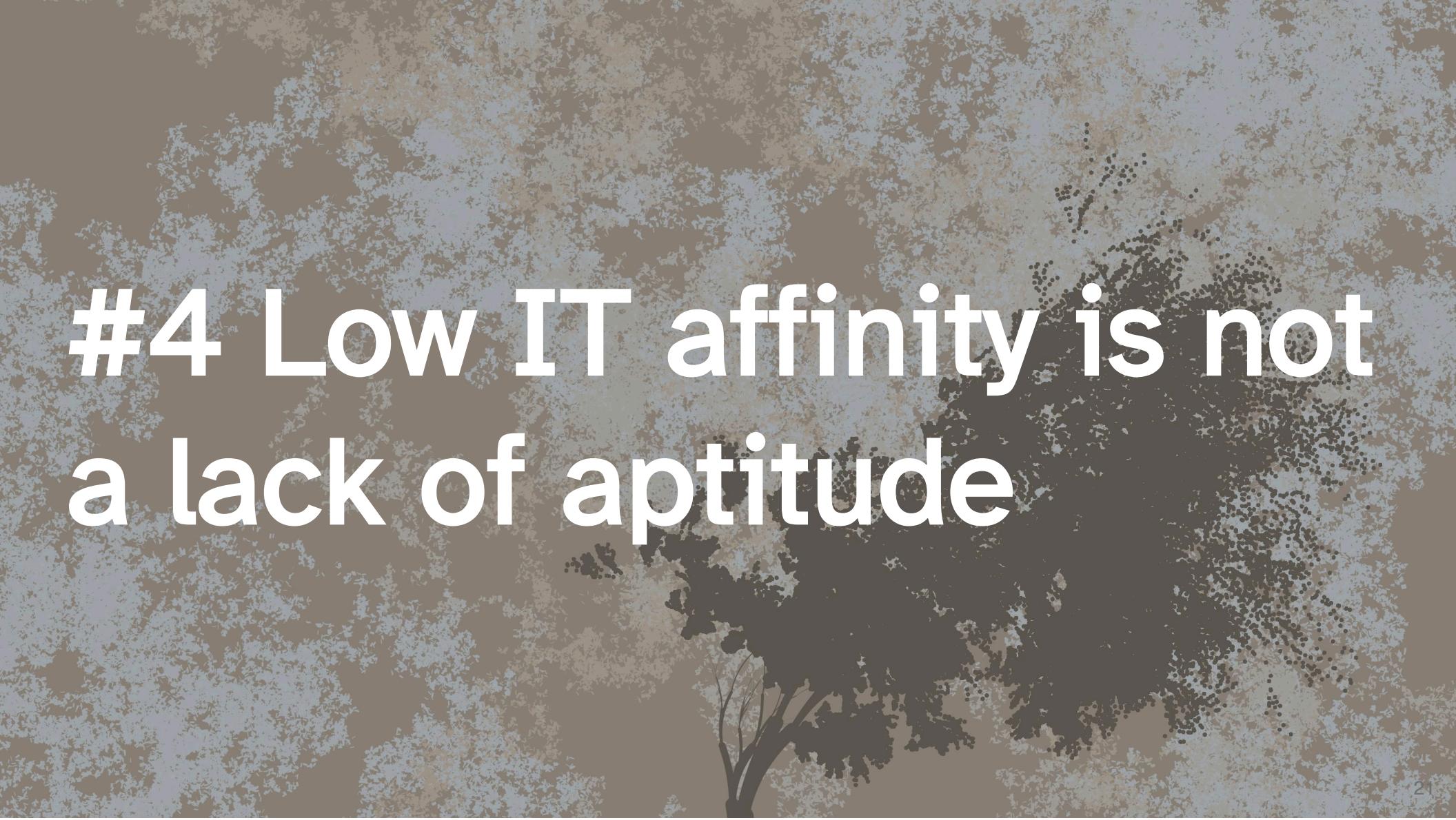
- 40/100: Research competence
- 40/100: Thesis report
- 10/100: Colloquium
- 10/100: Examination

‘Data Management’ under ‘Research Competence’

6: Data is fully documented, organized, easy to reproduce, and publication ready. Everything is stored on Google Drive.

But, data publication requirement

Obtaining a 6 from all sub-areas but not publishing the data in the form of a repository will result in a maximum allowed grade of 5.75.



#4 Low IT affinity is not a lack of aptitude

Safe learning environments

Growth-mindset for better learning outcomes

- **Fixed mindset:** ‘I’m not good’
- **Growth mindset:** ‘I can learn’

Create safe learner environments

- Regular 1:1 research data management meetings
- Bi-monthly half day team events
- Yearly retreat

msc-thesis - ghe-supervision Course Overview – Research +

rbtl-fs25.github.io/website/ ⌂

Research Beyond the Lab: Open Science and Research Methods for a Global Engineer ⓘ

Course Overview Course Calendar

Module 01 >
Module 02 >
Module 03 >
Module 04 >
Module 05 >
Module 06 >
Module 07 >
Module 08 >
Module 09 >

Learning Goals

1. Be able to use a common set of data science tools (R, RStudio IDE, Git, GitHub, tidyverse, Quarto) to illustrate and communicate the results of data analysis projects.
2. Learn to use the Quarto file format and the RStudio IDE visual editing mode to produce scholarly documents with citations, footnotes, cross-references, figures, and tables.
3. Be able to design a questionnaire to collect information that can be analysed to answer a waste-related research question that is relevant for Zurich.
4. Understand the main challenges associated with managing different types of waste, and how they differ between Europe and Africa.

Textbooks and Materials

On this page

Course Information
Learning Goals
Textbooks and Materials
Course Calendar (subject to change)
Weekly Structure (subject to change)
Performance assessment
Policies

#5 Data != Data

Disclaimer: Data at GHE

- small (few MBs)
- tabular
- non-sensitive
- topics
 - waste management
 - sanitation
 - air quality
 - etc.



Three terms for three stages

Three terms for three stages

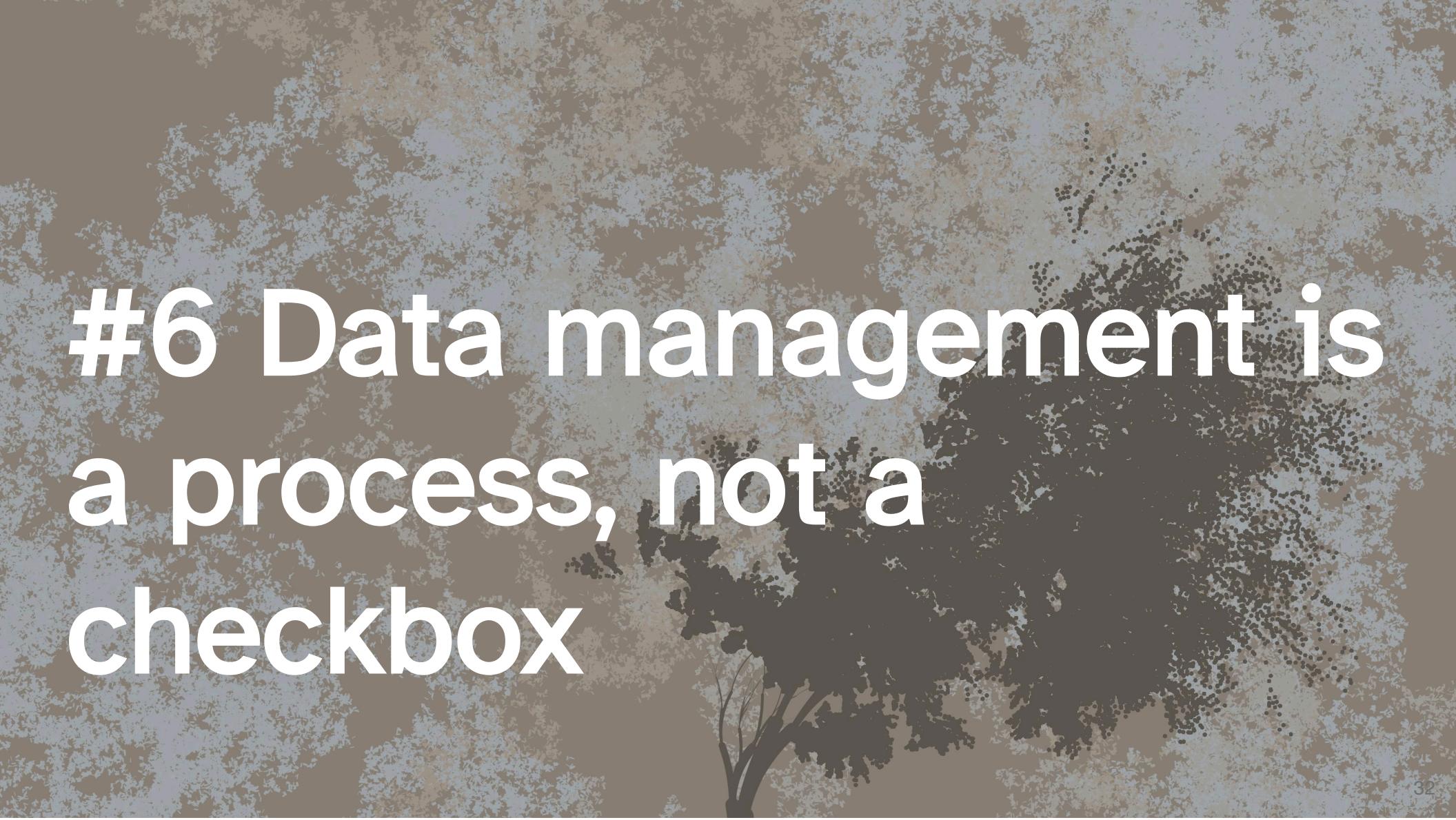
term	explanation	file format
unprocessed raw data	data that is not processed and remains in its original form and file type	often XLSX, also CSV and others

Three terms for three stages

term	explanation	file format
unprocessed raw data	data that is not processed and remains in its original form and file type	often XLSX, also CSV and others
processed analysis-ready data	data that is processed to prepare for an analysis and is exported in its new form as a new file	CSV, R data package

Three terms for three stages

term	explanation	file format
unprocessed raw data	data that is not processed and remains in its original form and file type	often XLSX, also CSV and others
processed analysis-ready data	data that is processed to prepare for an analysis and is exported in its new form as a new file	CSV, R data package
final data underlying a publication	data that is the result of an analysis (e.g descriptive statistics or data visualization) and shown in a publication, but then also exported in its new form as a new file	CSV



#6 Data management is
a process, not a
checkbox

research
questions

experimental
design

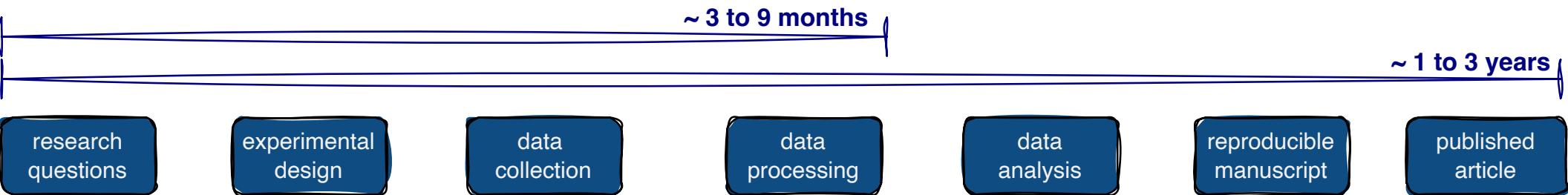
data
collection

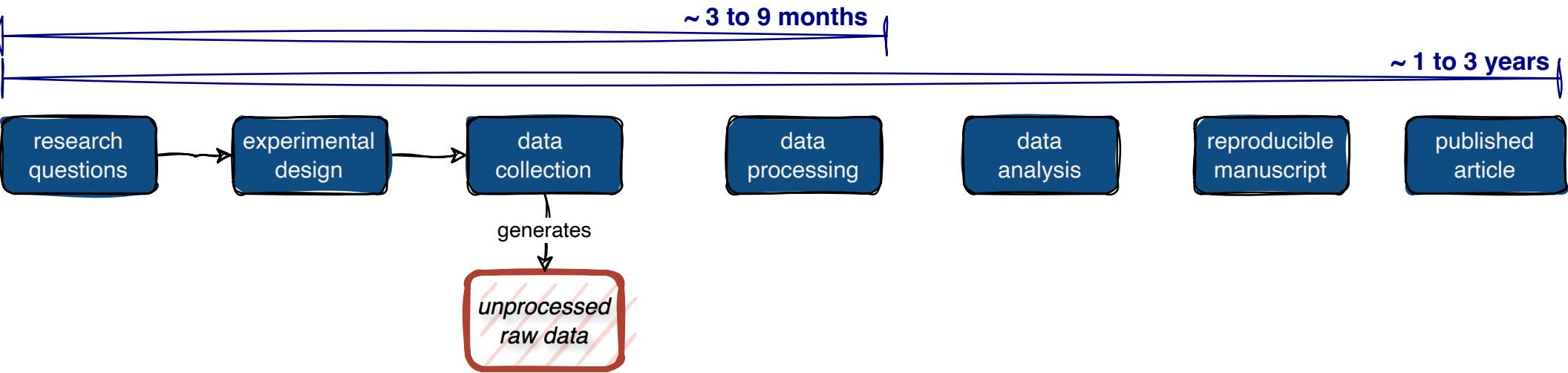
data
processing

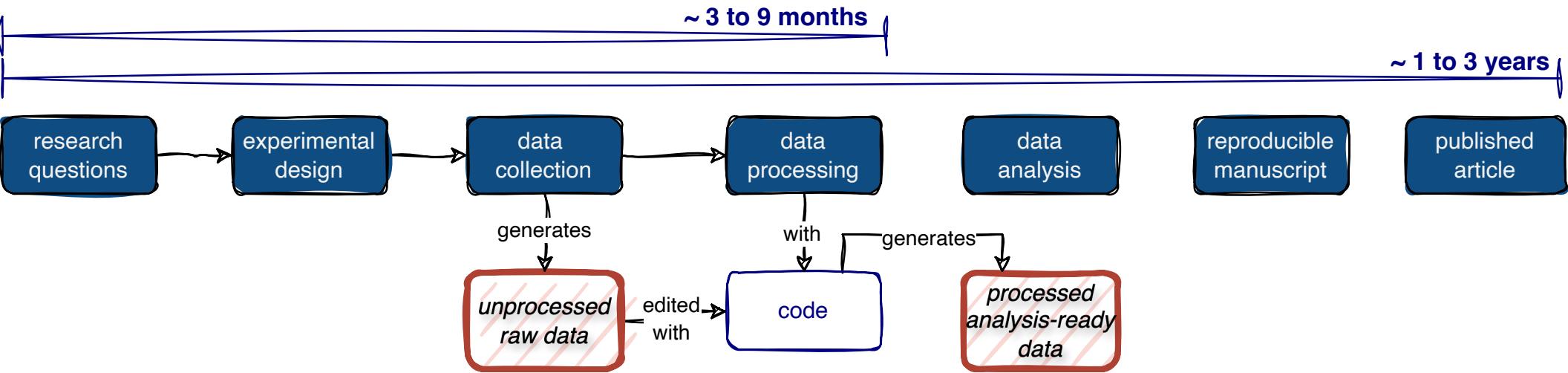
data
analysis

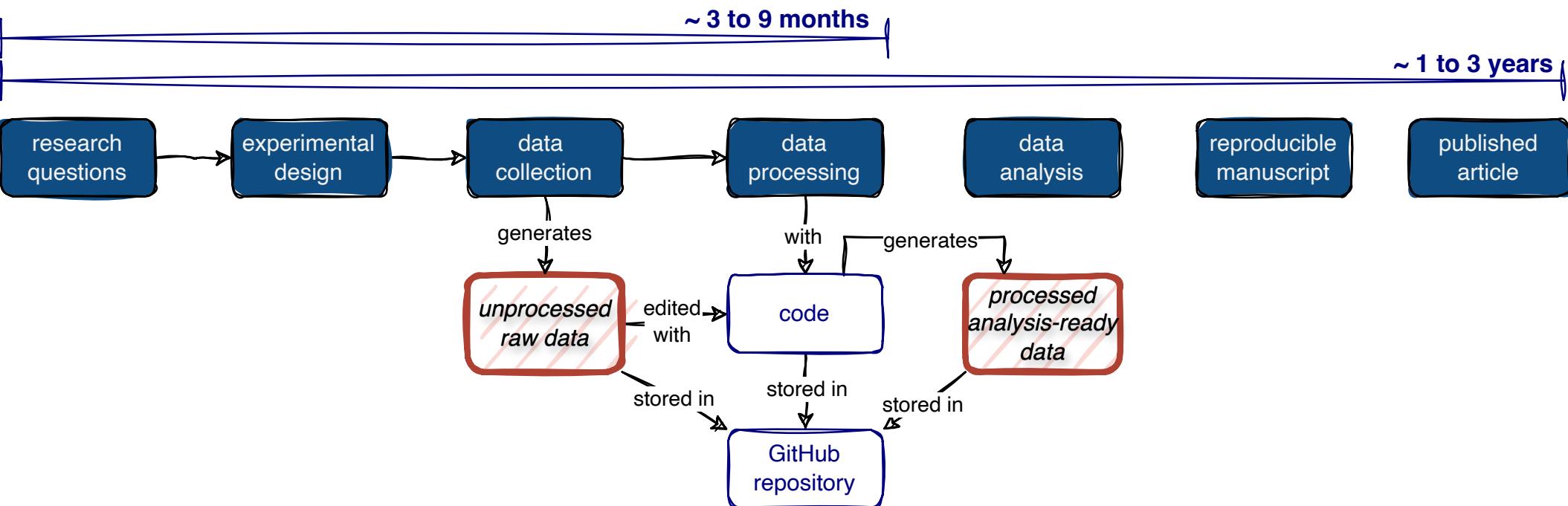
reproducible
manuscript

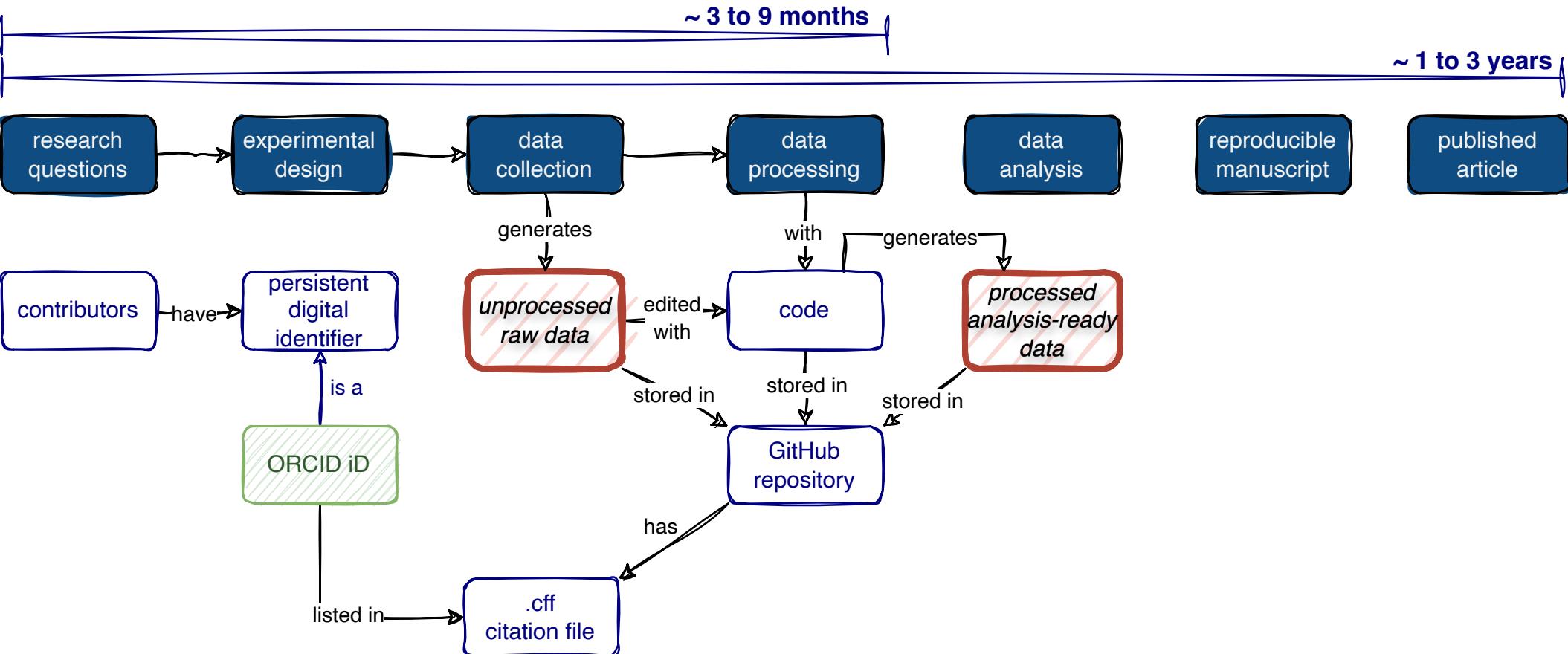
published
article

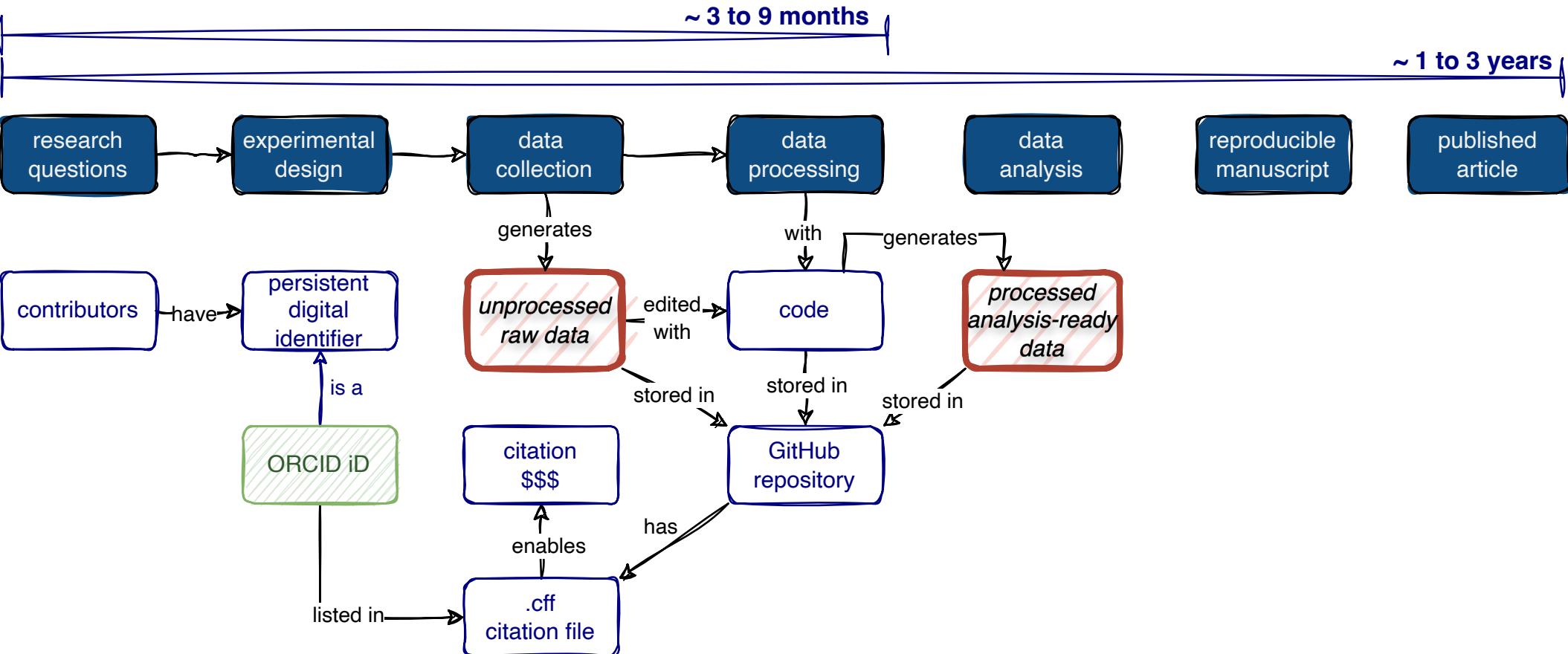


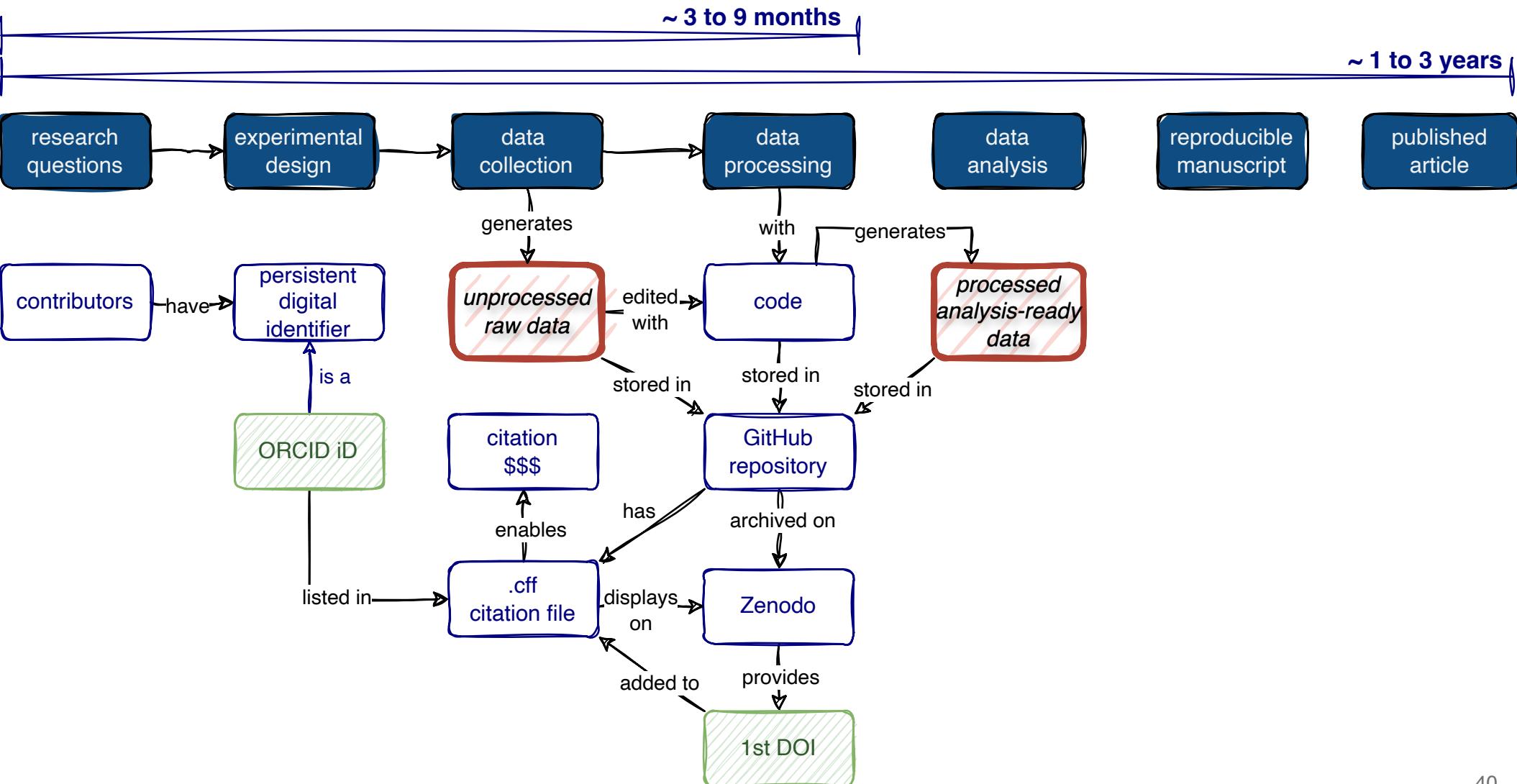


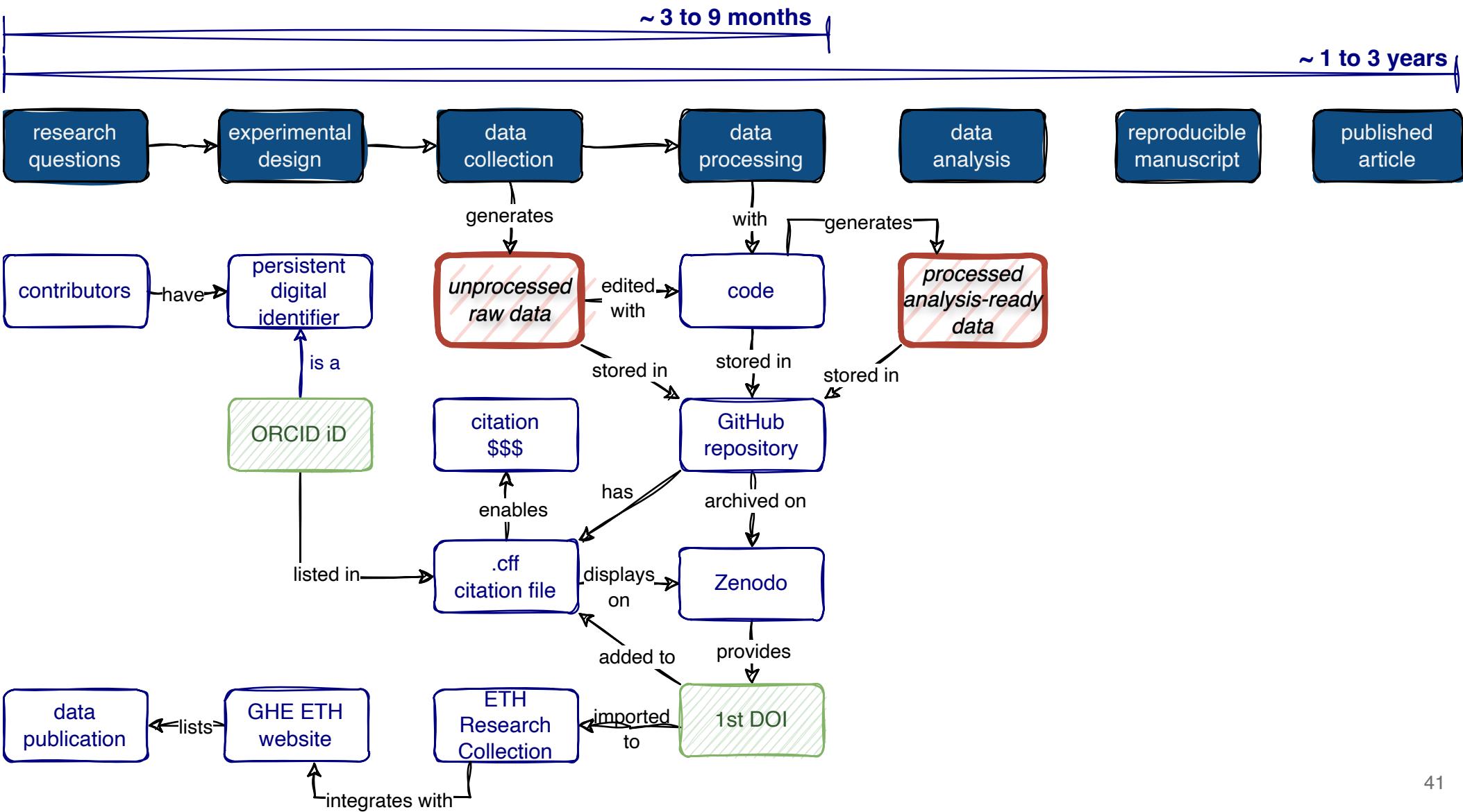


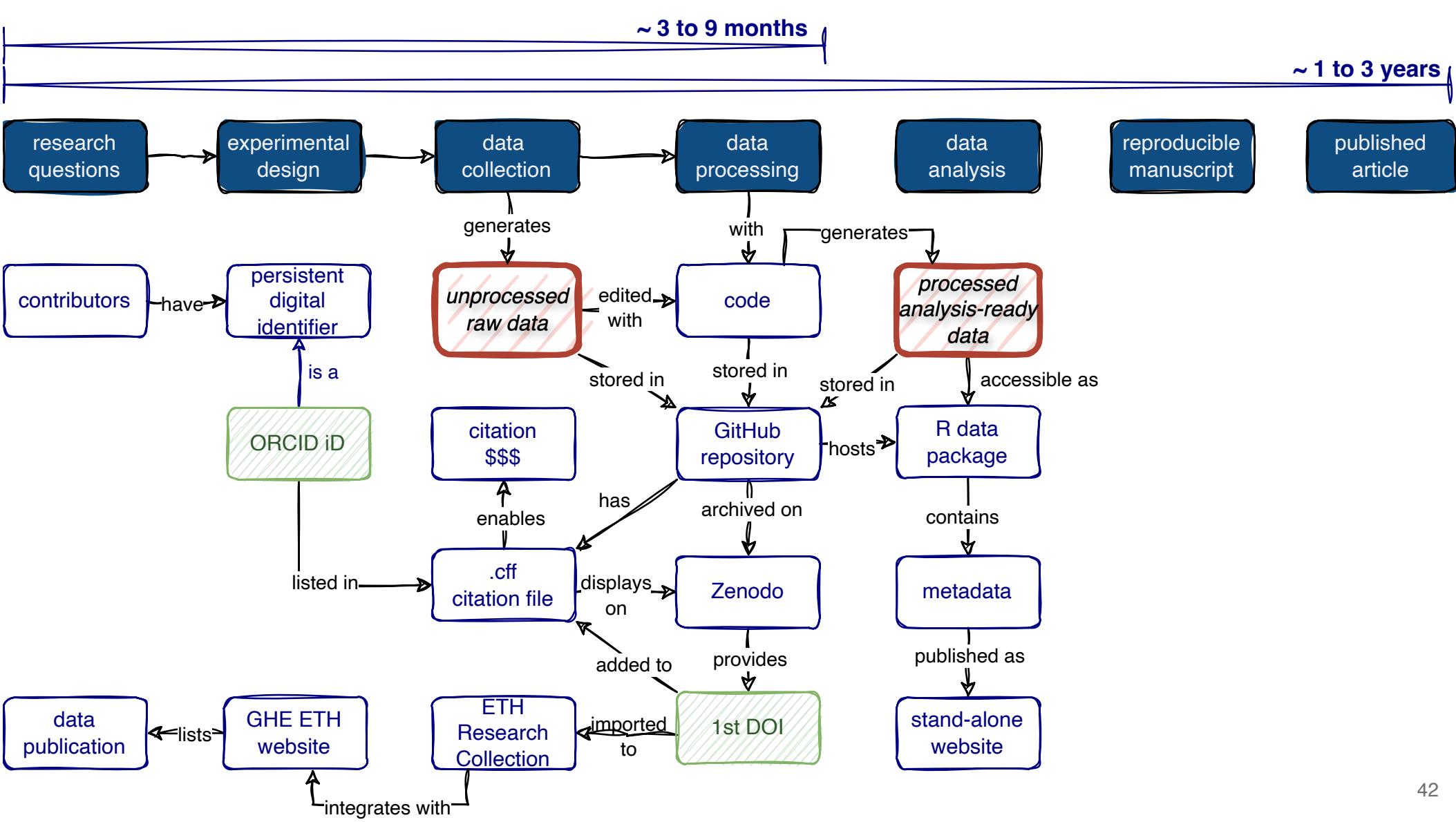


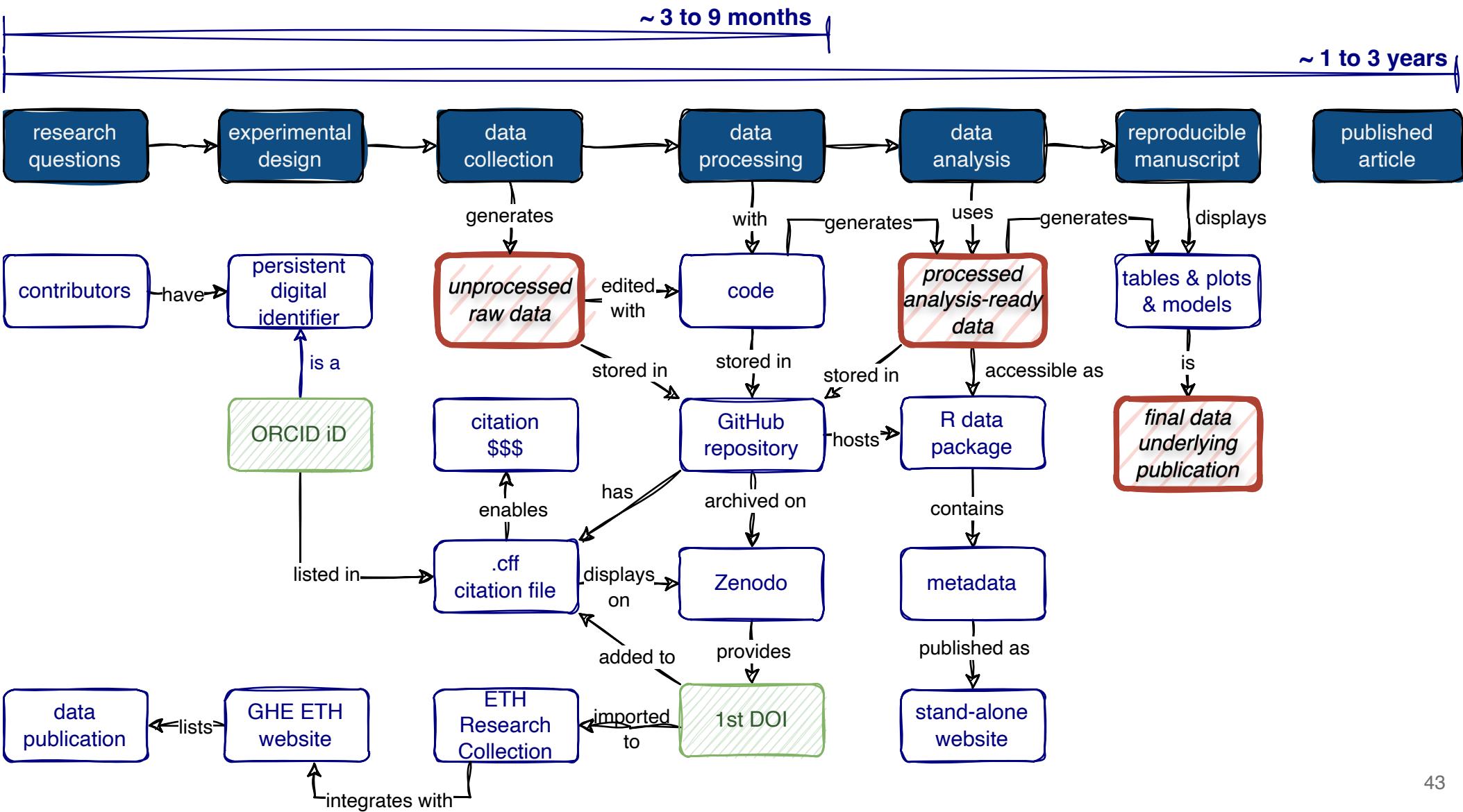


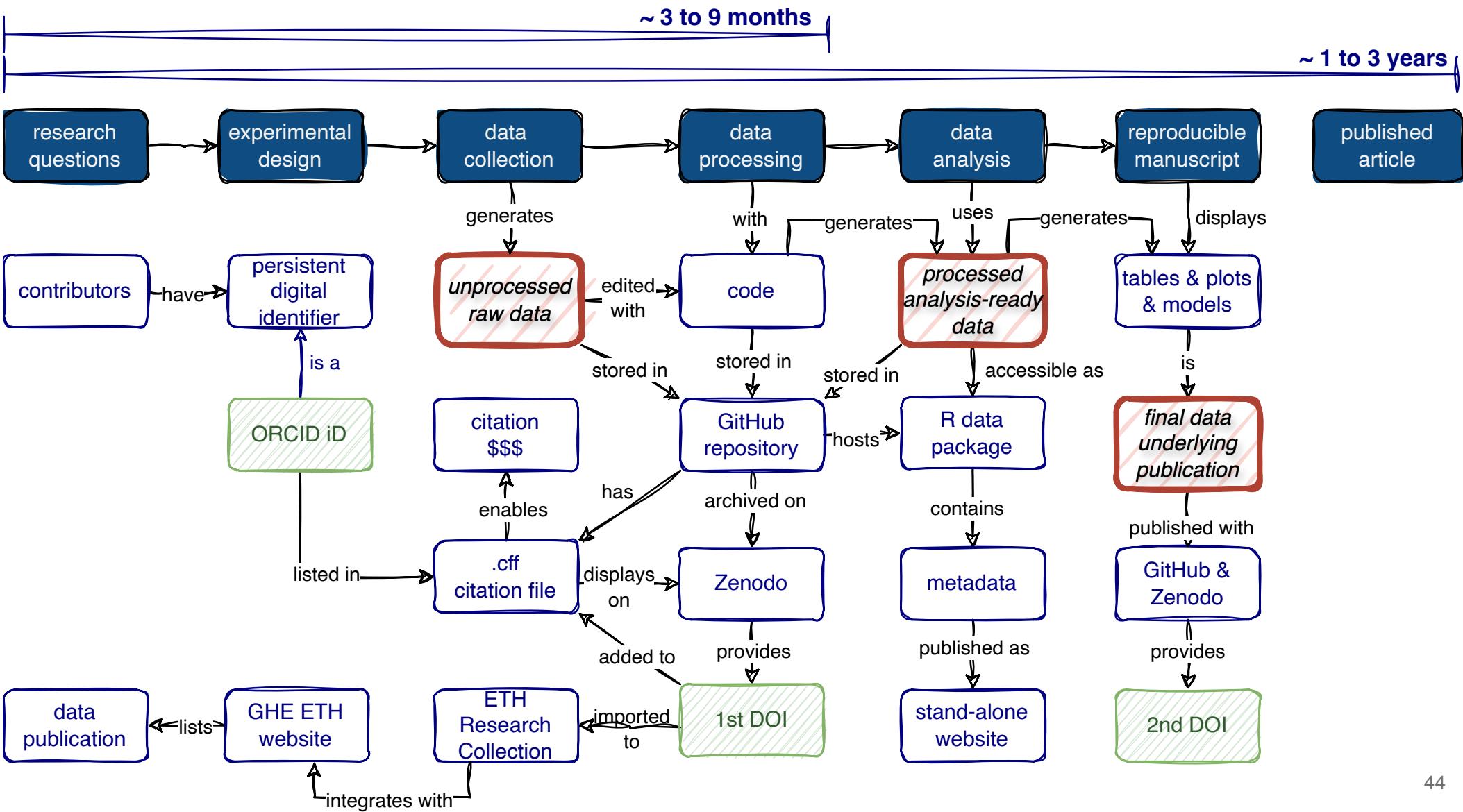






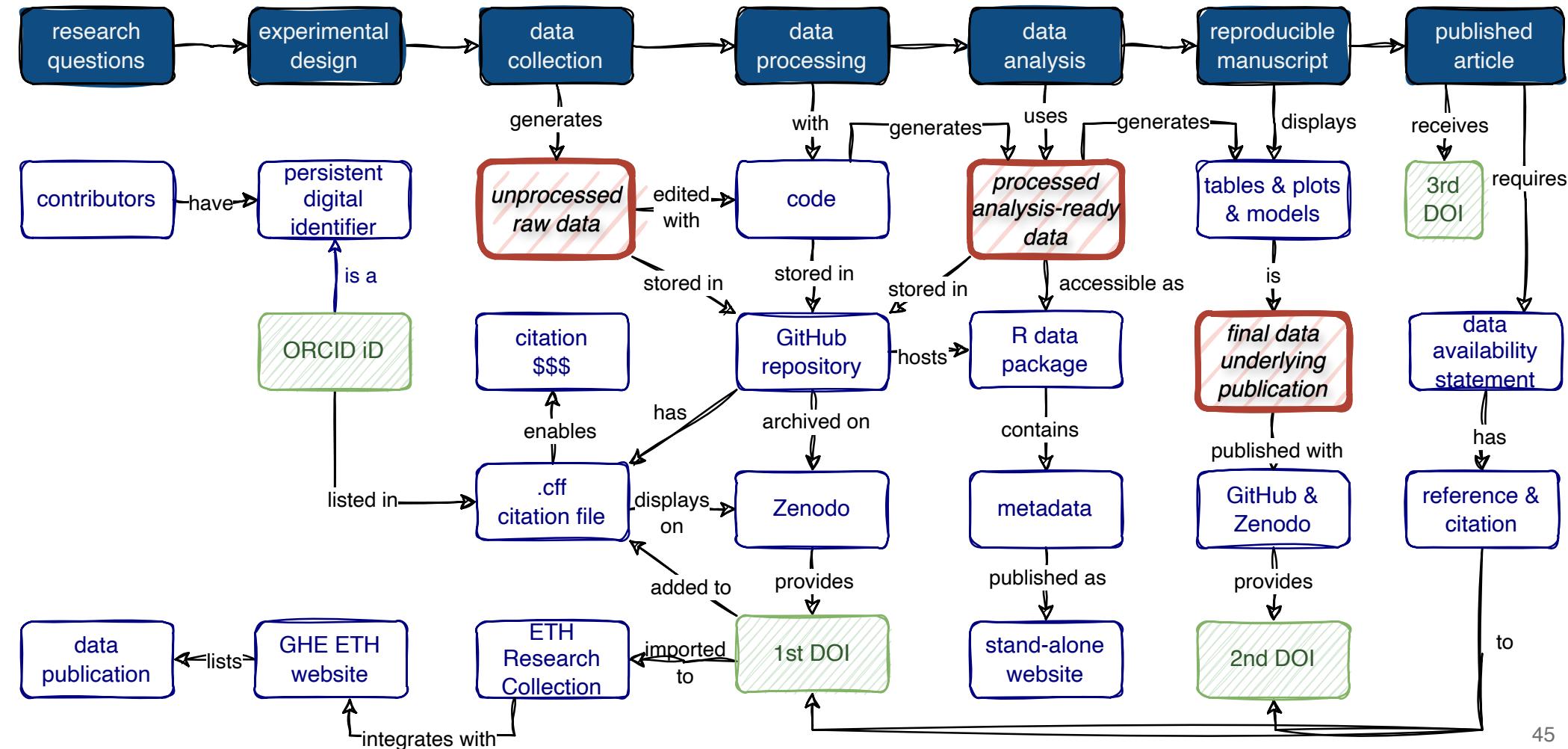






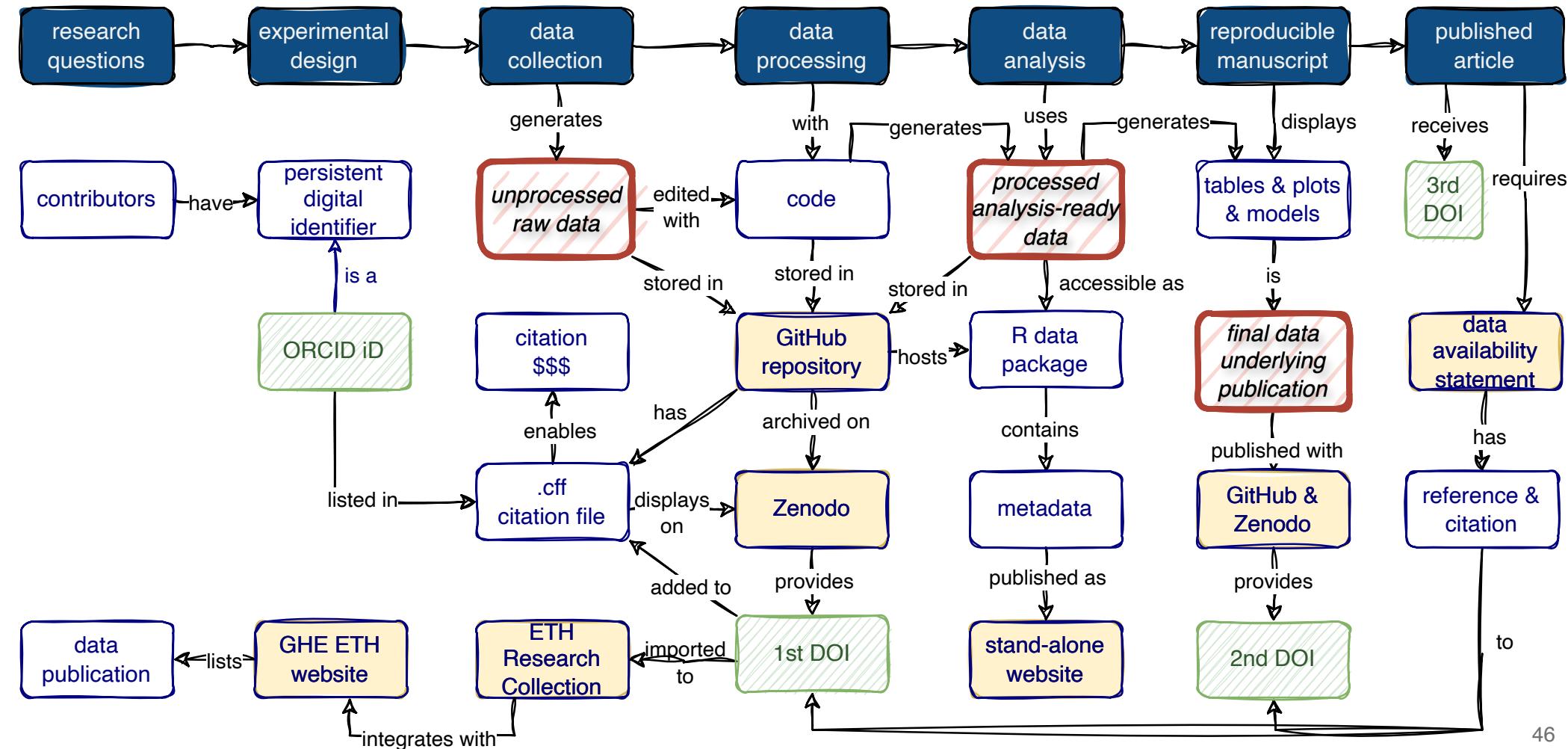
~ 3 to 9 months

~ 1 to 3 years

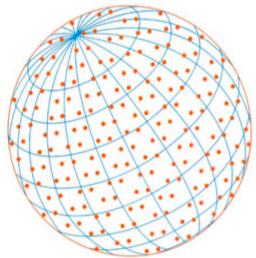


~ 3 to 9 months

~ 1 to 3 years



#7 Findable: Publish for humans and computers



Aerosol and Air Quality
Research

Special Issue:
Carbonaceous Aerosols in the
Atmosphere (VII)

ORIGINAL RESEARCH

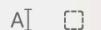
<https://doi.org/10.4209/aaqr.240095>

Absorption Ångström Exponent Values to Identify Light-absorbing Carbonaceous Aerosol Sources in Blantyre, Malawi

Saloni Vijay^{ID 1*}, Lars Schöbitz^{ID 1}, Hope Kelvin Chilunga^{ID 2}, Elizabeth Tilley^{ID 1}

¹ Global Health Engineering, Department of Mechanical and Process Engineering, ETH Zürich,
Zürich 8092, Switzerland

² Department of Computer Science and Information Technology, Malawi University of Business
and Applied Science, Blantyre 265, Malawi



LAC aerosols concentration data acquired from the two MAZUO monitors, we calculated the κ -score between the monitored data from both devices at an 880 nm wavelength channel. An R^2 score of 0.97 both times was obtained implying high data reliability.

2.6 Computational Reproducibility and Data Availability

R Statistical Software version 4.3.0 and RStudio IDE version 2023.9.1.494 were used for quantitative data analysis to generate the results in the manuscript (Posit team, 2023). Several R packages were used for data processing, analysis, and visualization (Schwalb-Willmann, 2024; Müller, 2023; Pebesma, 2018; Robinson *et al.*, 2023; Spinu *et al.*, 2023; Wickham *et al.*, 2023b, 2023a, 2024).

Raw data and processed data are available as an R package - [Vijay *et al.* \(2024a\)](#). The data analysis code used to generate the figures and tables of this manuscript are contained in a public GitHub repository ([Vijay *et al.*, 2024b](#)).

3 RESULTS AND DISCUSSION

■ Data – Global Health Engineer x +

◀ ▶ ⌂ ghe.ethz.ch/publications/data.html

Blog & News About Us Work with us Research Open Science Courses Publications

Homepage > Publications > Data

Data

2024 2023 2022

Automation from ETH Research Collection

bcsa: Data for source-apportionment of light absorbing carbon in Blantyre, Malawi

Saloni Vijay, Hope Kelvin Chilunga, Lennox Khonje, Jack Kamjombo, Elizabeth Tilley and Lars Schöbitz

Genève: CERN, 2024.

DOI: 10.5281/zenodo.10878607 ↗ Research Collection ↗ Abstract +

GHE GLOBAL
HEALTH
ENGINEERING

Contact

Prof. Dr. Elizabeth Tilley
Associate Professor at the Department of Mechanical and Process Engineering Deputy head of Inst. of Design, Materials a

Automation from Zenodo

bcsa: Data for source-apportionment of light-absorbing carbon in Blantyre, Malawi

zenodo.org/records/12685803

Published July 8, 2024 | Version v0.1.0

Software

Automation from GitHub

313 VIEWS 28 DOWNLOADS

Show more details

Versions

Version v0.1.0	Jul 8, 2024
10.5281/zenodo.12685803	
Version v0.0.1	Mar 26, 2024
10.5281/zenodo.10878884	
Version v0.0.0.9000	Mar 26, 2024
10.5281/zenodo.10878608	

View all 3 versions

Visit the website of this dataset for instructions how to use it: <https://global-health-engineering.github.io/bcsa/>

<https://github.com/Global-Health-Engineering/bcsa>

Global-Health-Engineering / bcsa

Type to search

Code Issues Pull requests Actions Projects Wiki Security Insights

bcsa Public

Watch 2 Fork 0 Star 0

main Go to file Code

larnsce update site 482726 · 4 months ago 66 Commits

.github add github action CMD check 8 months ago

R document df_collocation 8 months ago

data-raw add datasets documentation 8 months ago

data document datasets 8 months ago

docs update site 4 months ago

inst update citation 4 months ago

man document df_collocation 8 months ago

About

The bcsa package provide datasets for source apportionment of light absorbing carbon (LAC) in Blantyre, Malawi. The package contains data on Absorption Angstrom Exponent experiments determination of local pollution sources. The package also contains data on spatial distribution and ambient concentrations of LAC concentrations.

[global-health-engineering.github.io/b...](#)

air-quality open-data malawi

Readme

bcsa 0.0.1 Reference

CONCENTRATIONS. THE MAZOO MEASURES THE LAC CONCENTRATIONS IN REAL-TIME AT FIVE DIFFERENT WAVELENGTHS, THAT ALLOWS FOR SOURCE APPORTIONMENT.

Installation

You can install the development version of `bcsa` from [GitHub](#) with:

```
# install.packages("devtools")
devtools::install_github("Global-Health-Engineering/bcsa")
```

Alternatively, you can download the individual datasets as a CSV or XLSX file from the table below.

dataset	CSV	XLSX
<code>df_aae</code>	Download CSV	Download XLSX
<code>df_collocation</code>	Download CSV	Download XLSX
<code>df_mm</code>	Download CSV	Download XLSX
<code>df_mm_road_type</code>	Download CSV	Download XLSX

Dev status

License CC BY 4.0
R-CMD-check passing
DOI [10.5281/zenodo.12685803](#)

Funding

This project was funded by the [Global Health Engineering group at ETH Zurich](#).



#8 Funding for Open
Research Data exists
existed

Funding schemes

swissuniversities

- 2021 - 2024: swissuniversities - Open Science I
- 2025 - ~~2028~~ 2026: swissuniversities - Open Science II
- Watch: Action Line B5.2 - Professionalisation of ORD specialists and related services
- Newsletter sign-up:
<https://sympa.ethz.ch/sympa/subscribe/isci>

Funding schemes

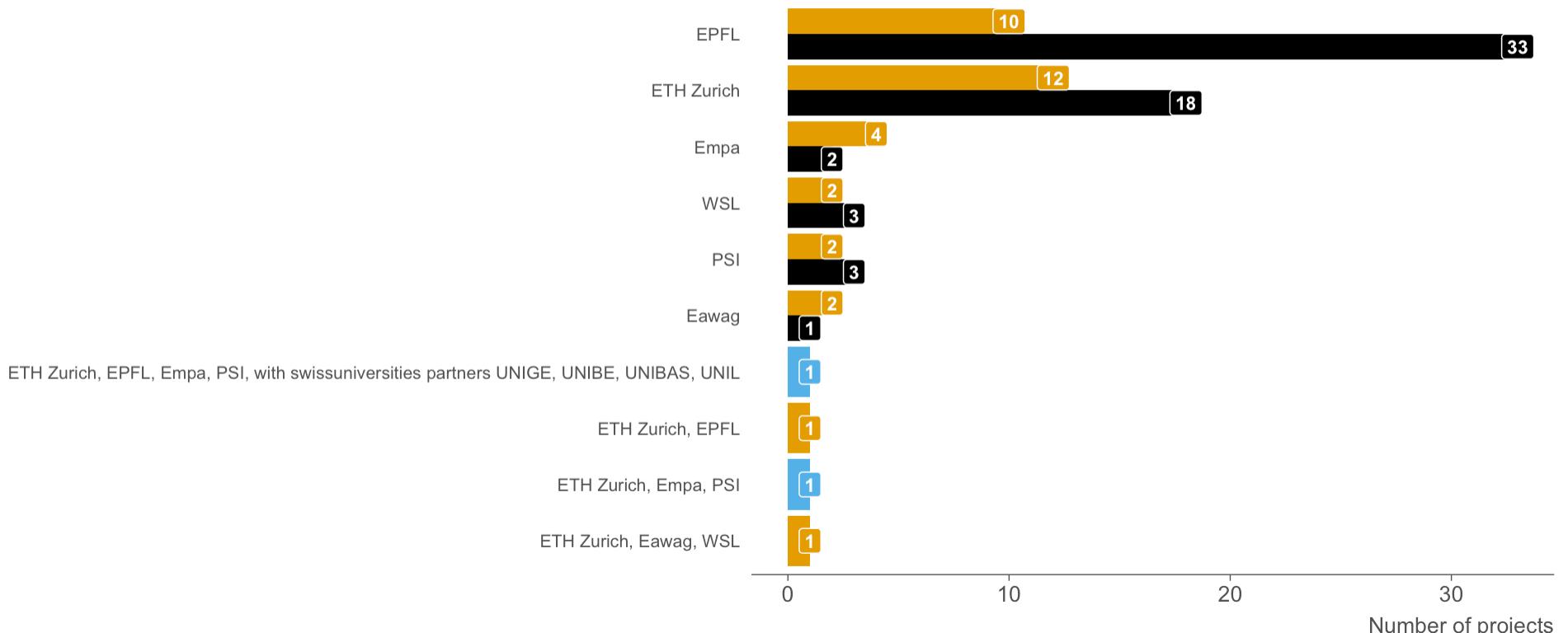
Open Research Data Program of the ETH Board

- 2021 - 2024: ~ 96 projects funded (~ CHF 15 million budget in total)
- Global Health Engineering was awarded 2 Contribute and 3 Explore projects worth 500'000 CHF
- All 96 projects and newsletter sign-up: <https://open-research-data-portal.ch/> (bottom of page)

Open Research Data Program of the ETH Board

Number of funded projects per institution and project category

Project category: ■ Contribute (30k) ■ Explore (150k) ■ Establish (1.5m)



8 take-aways from 30 minutes

- #1 Technology is not on our side
- #2 ETH wants reproducibility
- #3 Data management is project management
- #4 Low IT affinity is not a lack of aptitude
- #5 Data != Data
- #6 Data management is a process, not a checkbox
- #7 Findable: Publish for humans and computers
- #8 Funding for Open Research Data ~~exists~~ existed

Thanks! 🌻

Slides created via revealjs and Quarto: <https://quarto.org/docs/presentations/revealjs/>

Slide background image taken from Danielle Navarro

Access slides as PDF on GitHub or on <https://ghe-open.ch/slides/>



All material is licensed under Creative Commons Attribution Share Alike 4.0 International.

References

Massari, Nicolo, Lars Schöbitz, and Elizabeth Tilley. 2025. “Ethord: ETH Board Open Research Data (ORD) Program Project Metadata and Report Data.”
<https://doi.org/10.5281/zenodo.15554776>.