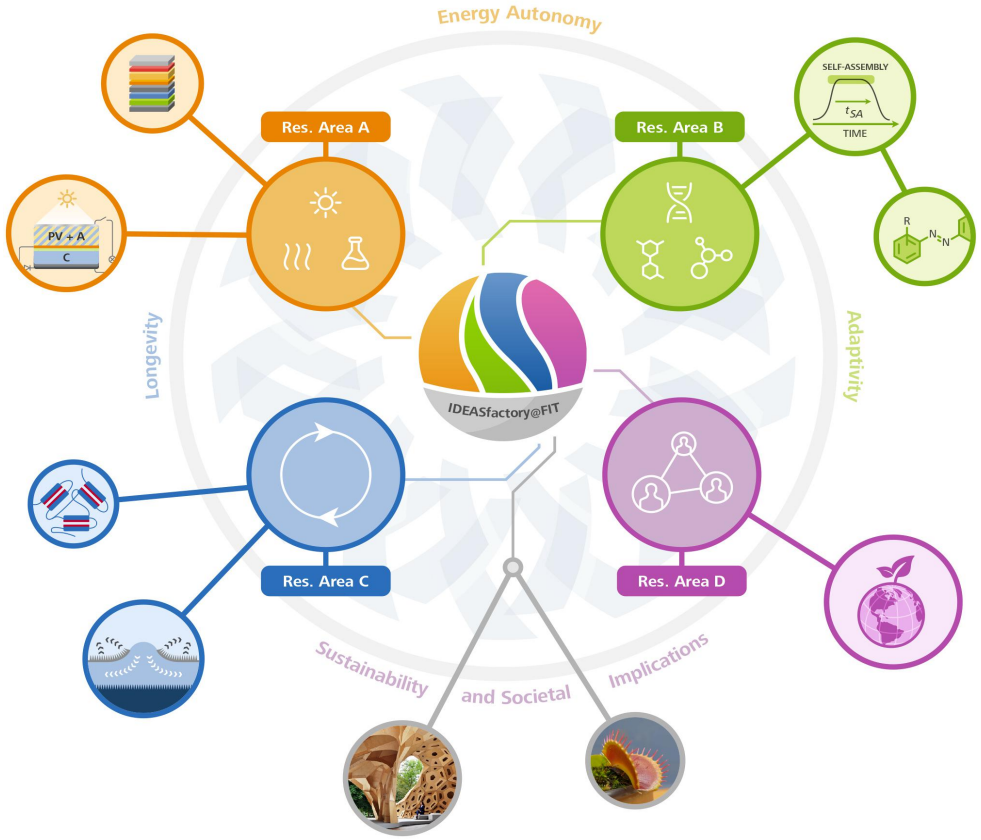
livMatS

UNI FREIBURG

Research Data Management

best practices & *liv*MatS state of the art



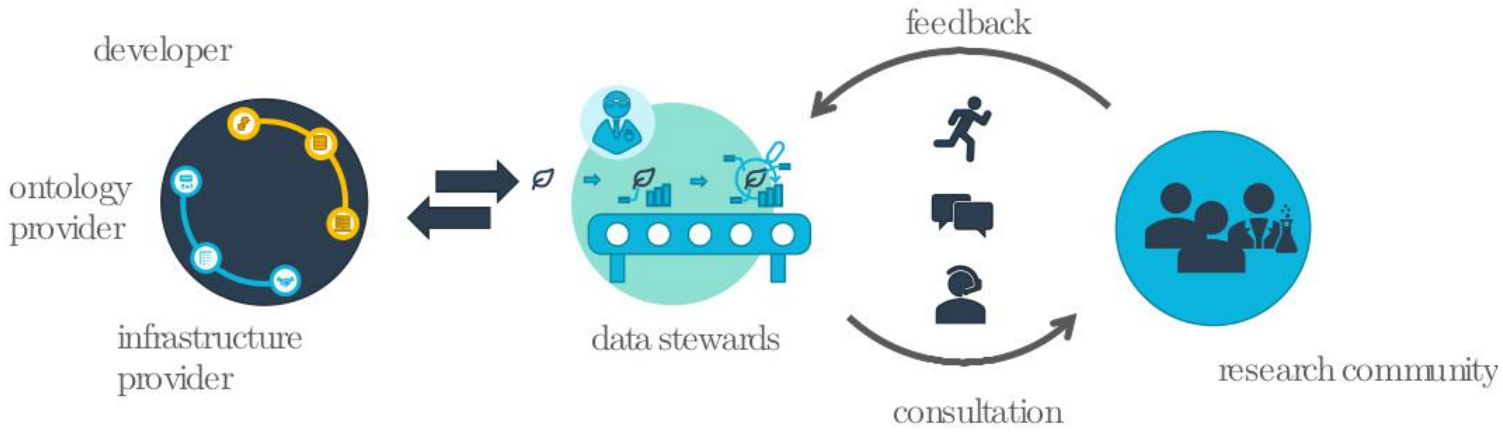
Johannes L. Hörmann

2023-03-02

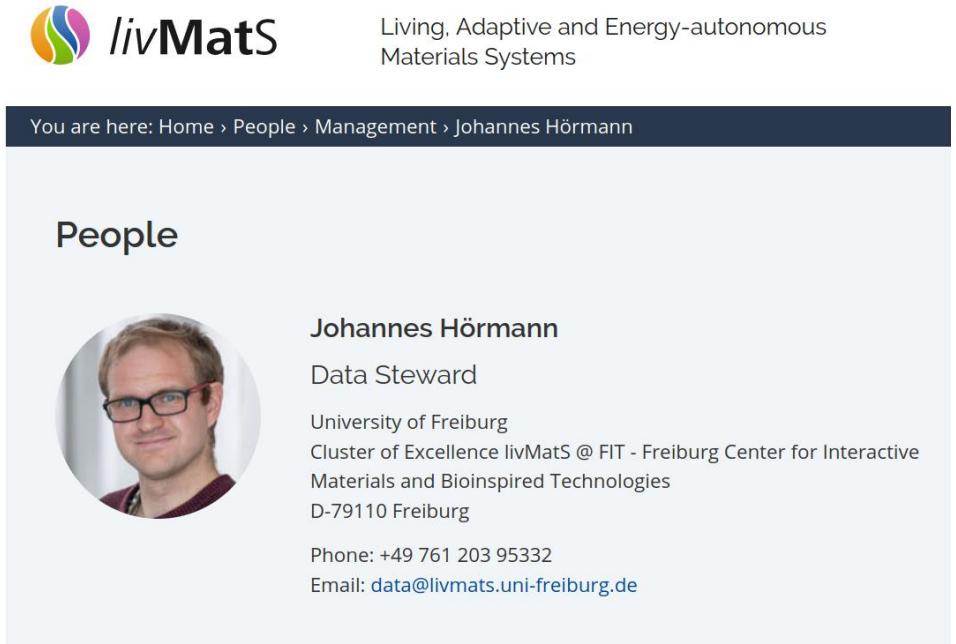
Outline

- **Why data steward & data management?**
- Basic best practice recommendations
- *liv*MatS RDM examples and services

The *liv*MatS data steward



source: von Suchodoletz, "Data Stewards as ambassadors between the NFDI and the community." (2021).



Data steward core tasks:

- Draft *liv*MatS **RDM policy**
- Develop RDM **training, support**
- Sofware solution development

source: https://www.livmats.uni-freiburg.de/en/people/management/johannes-hormann

In five years from now, ...

1. .... will my PI be able to find, understand, and provide all (possibly unpublished) data underlying my research?

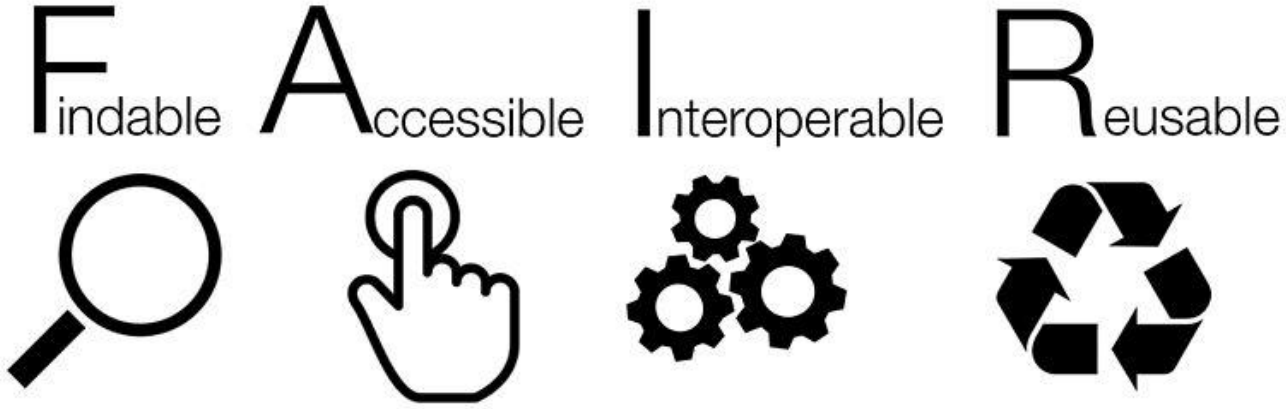
2. .... will administrative staff at *liv*MatS and the University of Freiburg be able to find and provide all (possibly unpublished) data underlying my research?

3. .... will anyone looking up my publications or thesis be able to find and understand the underlying data?

At offboarding, you need to answer these questions positively.

Research data management: Why?

Source: Wikipedia



To what end?

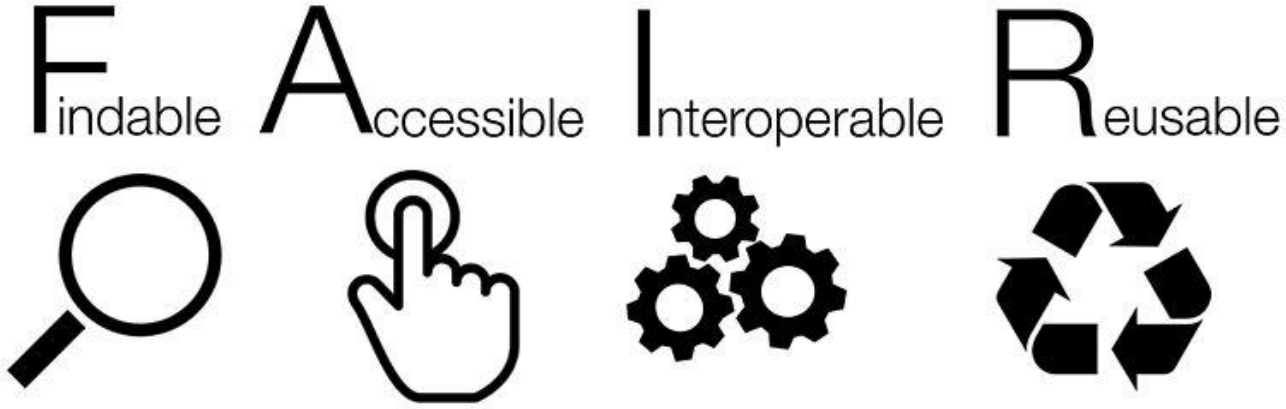
benefit to science and society?

benefit to other data scientists?

personal benefit to the resarcher?

Research data management: Why?

Source: Wikipedia



To what end?

benefit to science and society
- quality control
- reducing redundancy
- accelerating science
- saving costs

benefit to other data scientists

findable and AI-ready

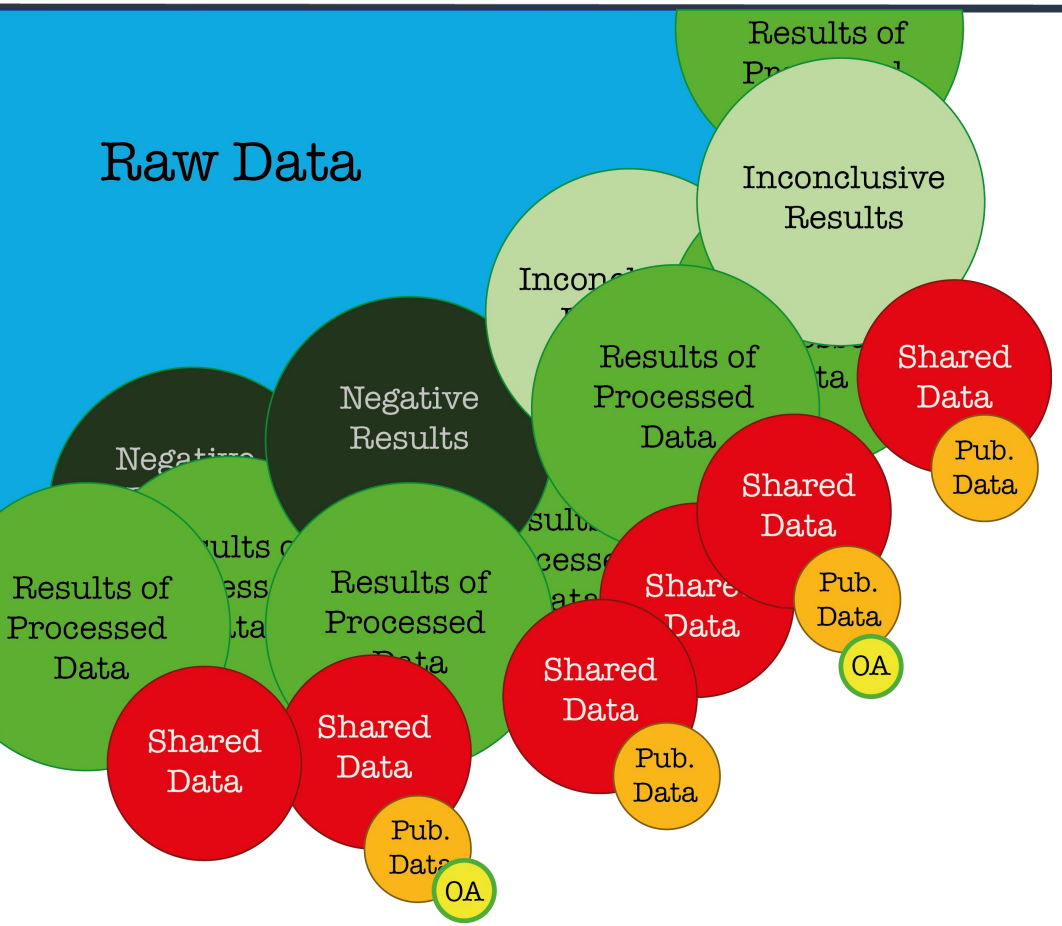
personal benefit to the researcher
- data hygiene, order
- easier collaboration
- increased visibility

Outline

- Why data steward & data management?
- **Basic best practice recommendations**
- *liv*MatS RDM examples and services

Data taxonomy & data life cycle

Data evolves through temporal phases that live one different maturity levels. Try to meet an appropriate cost-benefit ratio when thinking about how to document data at each stage.



Data from Research Processes: from raw data to open access published data, by Raman Ganguly (http://phaidra.univie.ac.at/o:387241)
Creative Commons Attribution-NonCommercial ShareAlike 4.0 International.

https://rdmkit.elixir-europe.org/images/data_life_cycle.svg
Creative Commons Attribution 4.0 International

Project directory structure

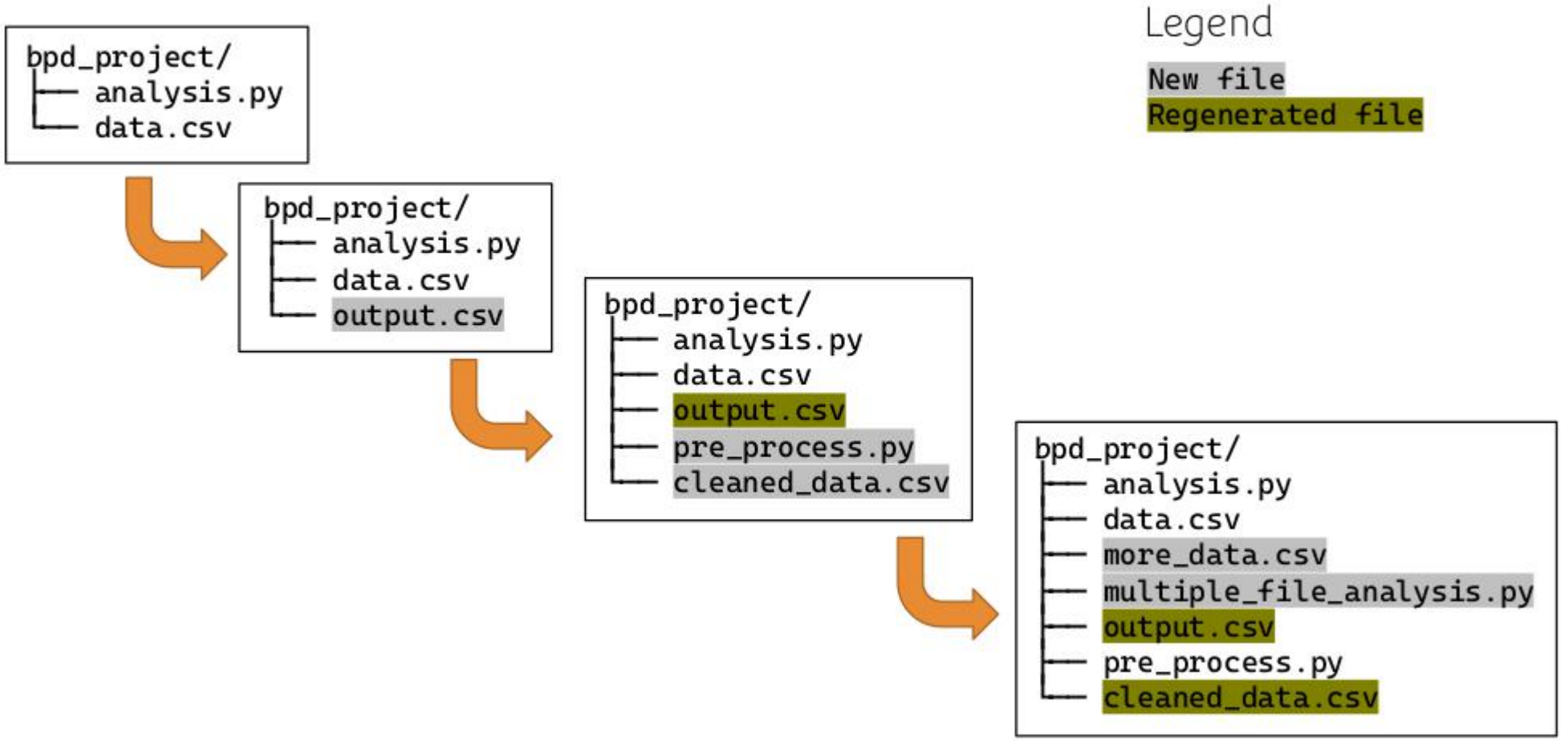
Think about meaningful directory structure.

```
bpd_project/
├── analysis.py
├── data.csv
├── more_data.csv
├── multiple_file_analysis.py
├── output.csv
├── pre_process.py
└── cleaned_data.csv
```

source: Tjelvar Olsson, Four principles for Effective Data Management

Project directory structure

Suggestion: sort data by provenance.



source: Tjelvar Olsson, Four principles for Effective Data Management

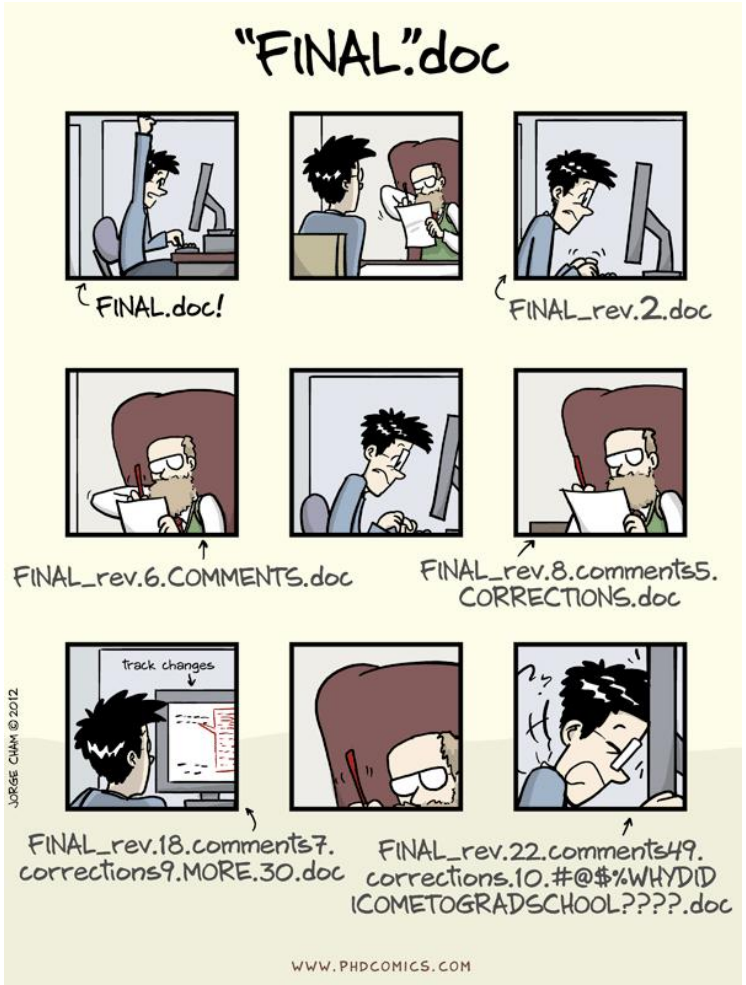
Project directory structure

Result: clear distinctiom between code, raw data, and derived data.

```
bpd_project/
├── final_results
│   └── output.csv
├── intermediate_data
│   └── cleaned_data.csv
├── raw_data
│   ├── data.csv
│   └── more_data.csv
└── scripts
    ├── analysis.py
    ├── multiple_file_analysis.py
    └── pre_process.py
```

source: Tjelvar Olsson, Four principles for Effective Data Management

Data management in its simplest form: File naming conventions



Source: http://phdcomics.com/comics/archive.php?comicid=1531

Think about meaningful file names.

Suggestion: embed metadata in filename, from general to specific.

2013-08-25_DOEProject_Ex1Test1_Data_Gonzalez_v3-03.xlsx

Date **Project** **Experiment** **Type** **ID** **Version**

General → Specific

Source: https://www.repro4everyone.org/resources/intro-fsbyt

Metadata in file names and directories

PROS

- Easy to use

Embedding metadata in file names is a good start but bears disadvantages

CONS

- Risk of losing metadata if renaming a file or reorganising the directory structure
- Loss of metadata if file alone is copied to a different location or sent to a collaborator
- Difficult to get an overview of all the data

source: Tjelvar Olsson, Four principles for Effective Data Management, https://www.youtube.com/watch?v=cq6C6b7MCo8

Metadata in spreadsheets

Suggestion: document data with metadata in accompanying spreadsheet.

File path	Date	Accession	Replicate	Treatment	Experiment type
/data/exp12.czi	2020-01-14	Col-0	1	Control	Microscopy

source: Tjelvar Olsson, Four principles for Effective Data Management, https://www.youtube.com/watch?v=cq6C6b7MCo8

Metadata in spreadsheets

PROS

- Easy to use
- Easy to get an overview of the data

Embedding metadata spreadsheets has its merits, but again bears disadvantages and risks

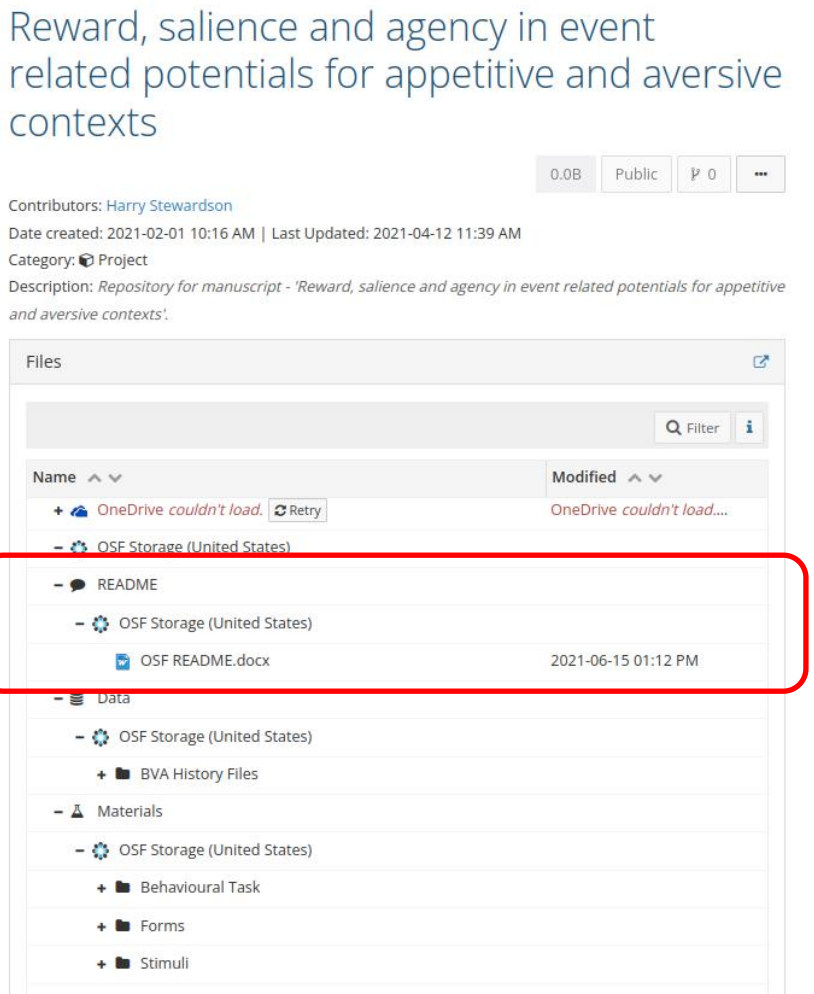
CONS

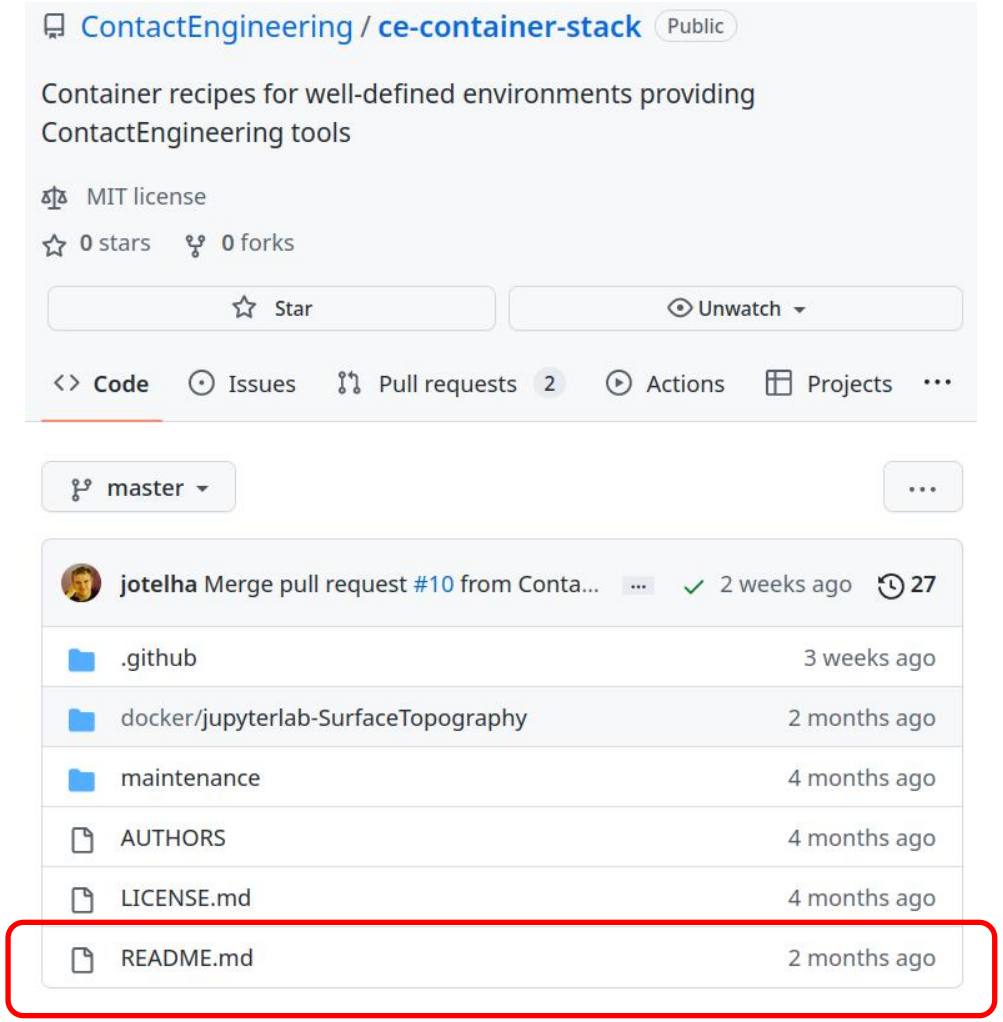
- Very difficult to move files or reorganise
- directory structures without breaking links in
- Loss of metadata if file alone is copied to a different location or sent to a collaborator
- Can be difficult to integrate into automized workflows

source: Tjelvar Olsson, Four principles for Effective Data Management, https://www.youtube.com/watch?v=cq6C6b7MCo8

README files make folders self-descriptive

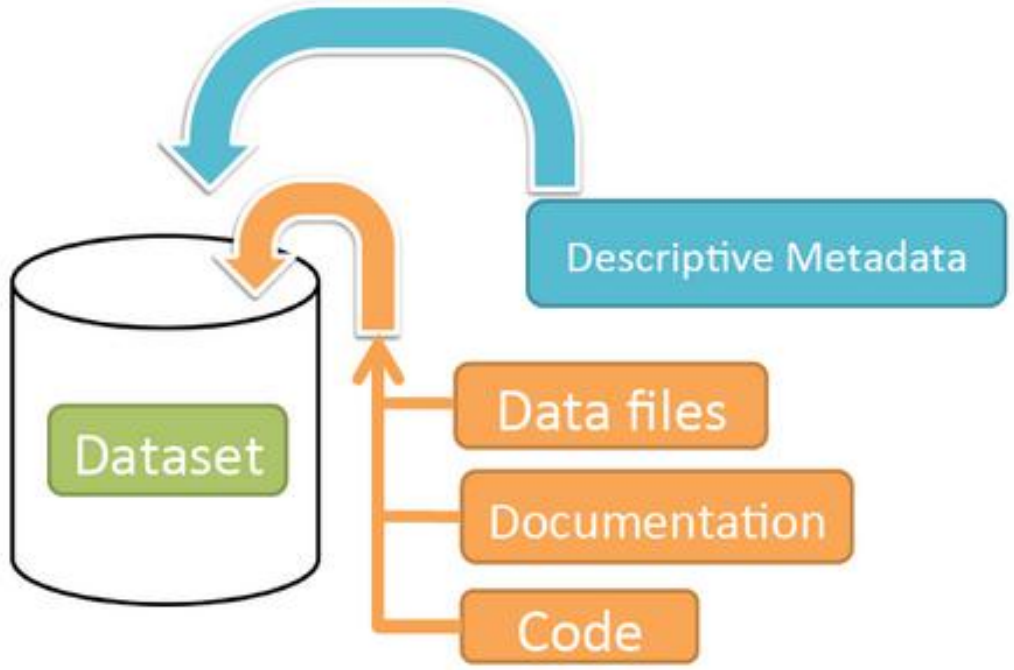
Write free text documentation README file with the aim to make the contents of a folder understandable to anyone who looks at it out-of-context.





Self-descriptive dataset: the common denominator of all RDM approaches

RDM tools and platforms formalize what's been discussed on the previous slides and encourage documentation with metadata in machine-readible format.



Source: https://apps.fz-juelich.de/fdm/staging/6-data-linking/_images/DatasetDiagram.png

```yaml
project: livMatS research data management
description: RDM slides retreat 2021
owners:
  - name: Johannes Hörmann
    email: data@livmats.uni-freiburg.de
    username: jh1130
    orcid: 0000-0001-5867-695X
funders:
  - organization: DFG
    program: livMatS
    code: EXC 2193
creation_date: 2021-11-17
```

metadata

Outline

- Why data steward & data management?
- Basic best practice recommendations
- ***liv*MatS RDM examples and services**

Virtue of the *liv*MatS RDM policy

Policy distributes responsibilities, but does not prescribe any technical pathway.

University's RDM policy

doi: 10.6094/UNIFR/231612

*liv*MatS RDM policy

livmats.uni-freiburg.de/rdm

PIs and students document the policy's implementation in continuosuly evolving per-project DMPs

implements RDM via ELN chemotion

implements RDM via dtool

implements RDM via OSF

...

sample DMP 1

sample DMP 2

sample DMP 3

tabular, per project, ~ 1 to 2 pages

DMP = data management plan

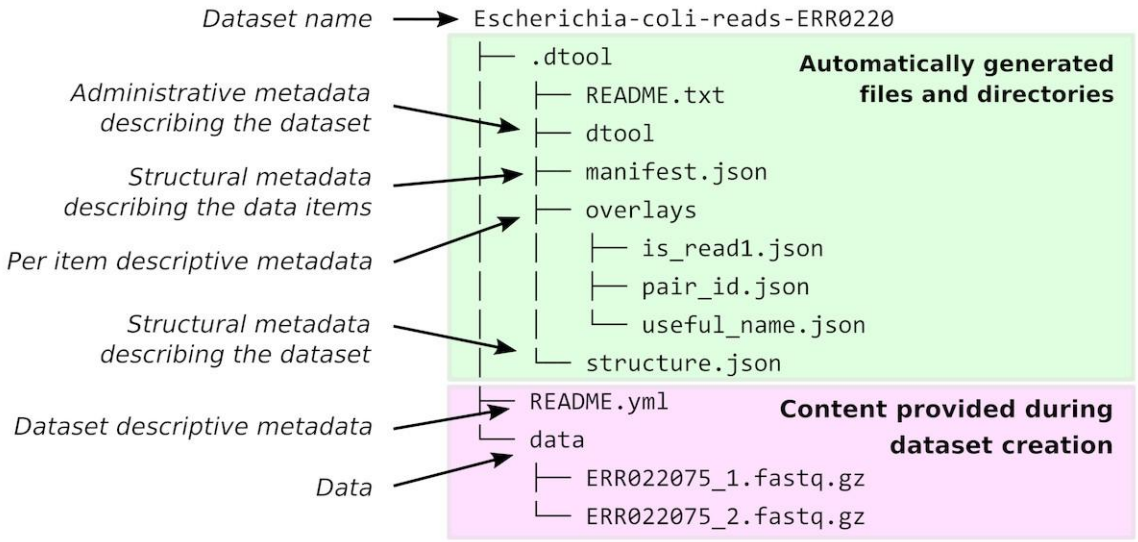
livmats.uni-freiburg.de/rdm/dmp

dtool core features

Basic RDM ecosystem introduced in *liv*MatS RDM workshops.

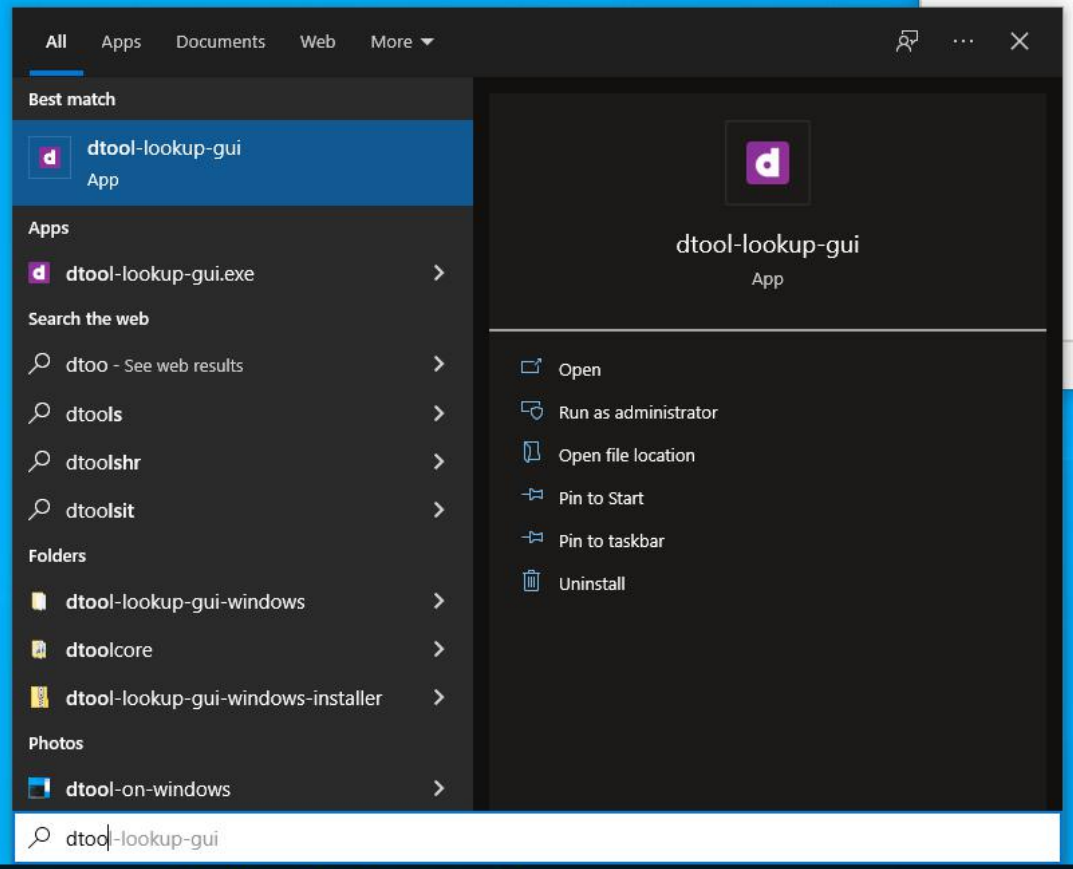
- package data and metdata in **datasets**
- assign **UUID** as unique identifier
- lightweight & **decentralized**
- **modular, implemented in Python, easily extendable**
- standardization via API, not via representation on file system
- underlying storage can be
  - standard file system
  - **S3 object storage**
  - **smb shares**

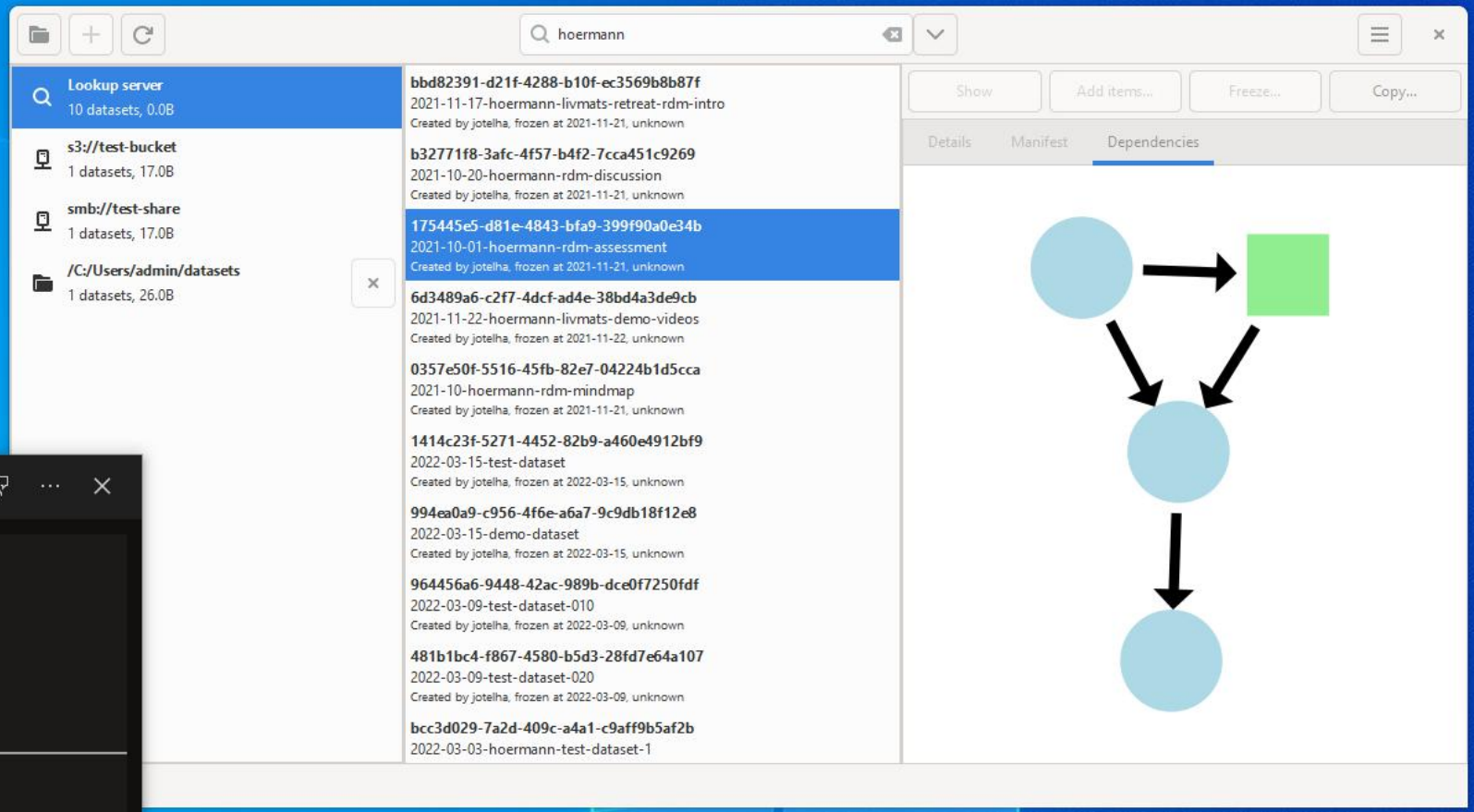
amongst others



T. S. G. Olsson and M. Hartley, "Lightweight data management with dtool," PeerJ, vol. 7, e6562, 2019: https://doi.org/10.7717/peerj.6562/fig-1, Creative Commons Attribution 4.0 International

IMTEK simulation (Pastewka group) uses dtool in production





dtool illustrates neatly:

- keeping data and metadata bundled together in self-contained datasets
- storing them locally and on centralized storage
- searching and sharing them
- deriving new data from them
- keeping track of data provenance

Research Data Management with dtool & *liv*MatS data repository

desktop app at
github.com/livMatS/dtool-lookup-gui

web interface to bwSFS-based data repository at
livmats-data.vm.uni-freiburg.de

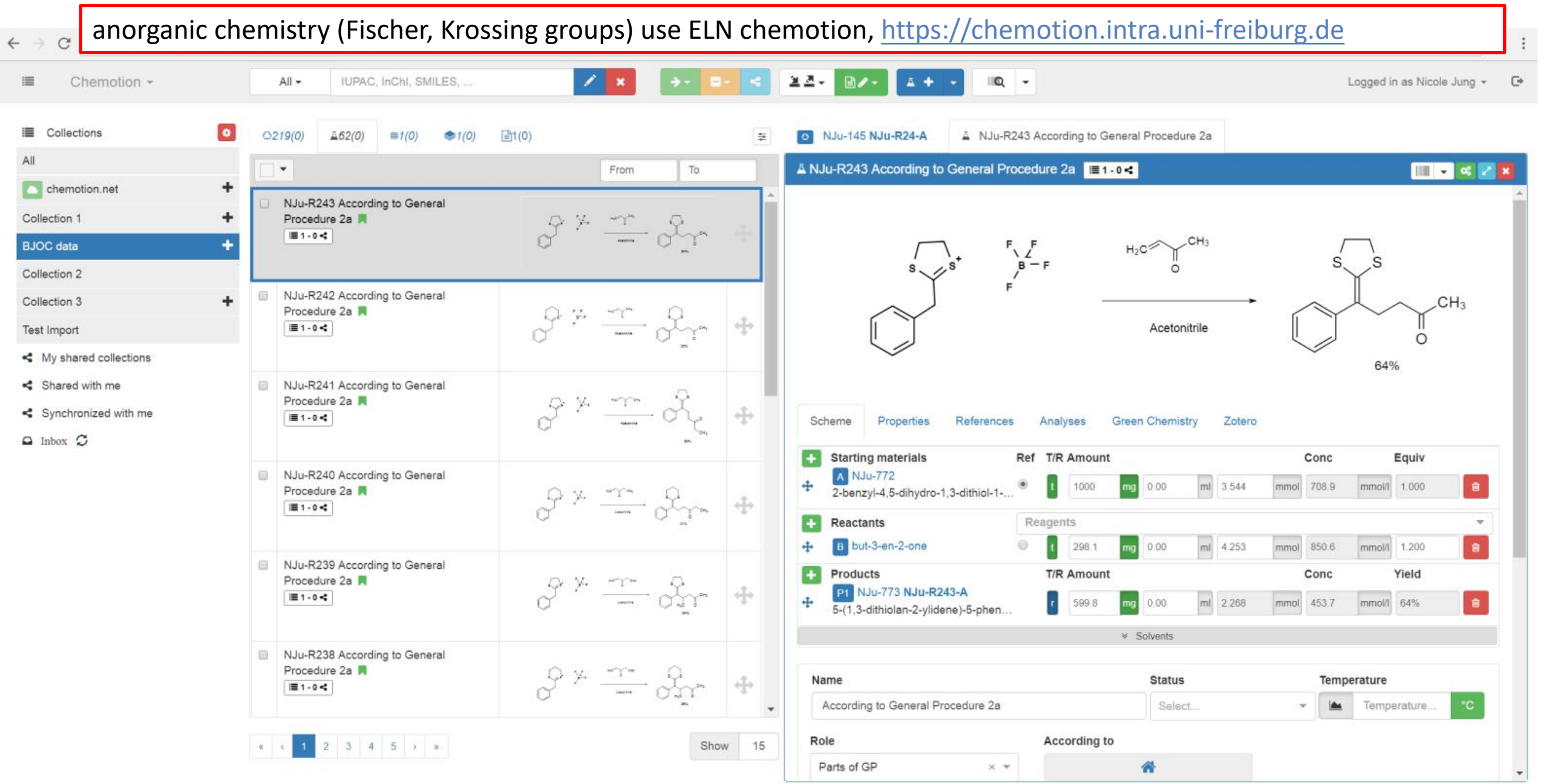
livMatS RDM landing page on
www.livmats.uni-freiburg.de/rdm

access for *liv*MatS affiliates by registration on
livmats-data.vm.uni-freiburg.de/config,
support by data@livmats.uni-freiburg.de

*liv*MatS offers a simple centralized dtool dataset repository based on bwSFS, the Baden-Württemberg Storage for Science S3 object storage infrastructure.

icons source: https://www.iconfinder.com under Creative Commons (Attribution 3.0 Unported)

anorganic chemistry (Fischer, Krossing groups) use ELN chemotion, https://chemotion.intra.uni-freiburg.de



Source: https://os.helmholtz.de/fileadmin/user_upload/os.helmholtz.de/Workshops/eln18hzi_jung.pdf

psychologists (Kiesel group) publish data on OSF

Activities Firefox Web Browser Sep 28 13:49 de

OSF | Anticipatory saccad

https://osf.io/uky3m/

Search

IMTEK Simulation - Zulip | GroupWiki | Overleaf | IMTEK Rooms | Server Groups | dtool-lookup-server/S... | NetApp® StorageGRI... | jic-dtool | bwCloud | pditommaso/awesom... | Weiterbildung und off... | bwsupport | bwHPC Batch

OSF HOME Search Support Donate Sign Up Sign In

Anticipatory saccades - experiments for... Files Wiki Analytics Registrations

Anticipatory saccades - experiments forced-choice & free-choice - Pfeuffer, Kiesel, & Huestegge (2016)

0.0B Public 0 ...

Contributors: Christina U. Pfeuffer, Andrea Kiesel

Date created: 2016-05-23 03:01 PM | Last Updated: 2016-12-23 04:49 PM

Identifiers: DOI 10.17605/OSF.IO/UKY3M ARK c7605/osf.io/uky3m

Category: Project

Files

Filter

Name	Modified
Anticipatory saccades - experiments forced-choice & free-choice - Pf...	
OSF Storage (United States)	
Data - Exp. 1 - forced-choice	
OSF Storage (United States)	
Anticipatory Saccades - Exp. 1 - forced-choice - saccade dat...	2016-06-22 12:00 PM
Anticipatory Saccades - Exp. 1 - forced-choice - trial RT and ...	2016-06-22 12:00 PM
Data - Exp. 2 - free-choice	
OSF Storage (United States)	
Anticipatory Saccades - Exp. 2 - free-choice - saccade data (...	2016-06-22 12:00 PM
Anticipatory Saccades - Exp. 2 - free-choice - trial RT data.txt	2016-06-22 12:00 PM
Data - Exp. 3 - forced-choice	
OSF Storage (United States)	

Citation

Components

- Data - Exp. 1 - forced-choice — Pfeuffer & Kiesel
- Data - Exp. 2 - free-choice — Pfeuffer & Kiesel
- Data - Exp. 3 - forced-choice — Pfeuffer & Kiesel
- Information on Data Files — Pfeuffer & Kiesel

Recent Activity

Uni Freiburg Initiatives & Services

RDMG Research Data Management Group https://rdmg.uni-freiburg.de/

livMatS livmats.uni-freiburg.de/rdm, https://github.com/livMatS/

FreiDok *plus* Universitätsbibliothek Freiburg publication platform https://freidok.uni-freiburg.de/

Conclusion

livMatS

If you know other valuable, well-accepted RDM best practices and tools in your discipline that are not captured within these slides, please get in touch with data@livmats.uni-freiburg.de.

- *liv*MatS
  - research data manegement: livmats.uni-freiburg.de/rdm
  - on github: github.com/livMatS

- Johannes L. Hörmann
  - on LinkedIn: linkedin.com/in/jotelha
  - on github: github.com/jotelha