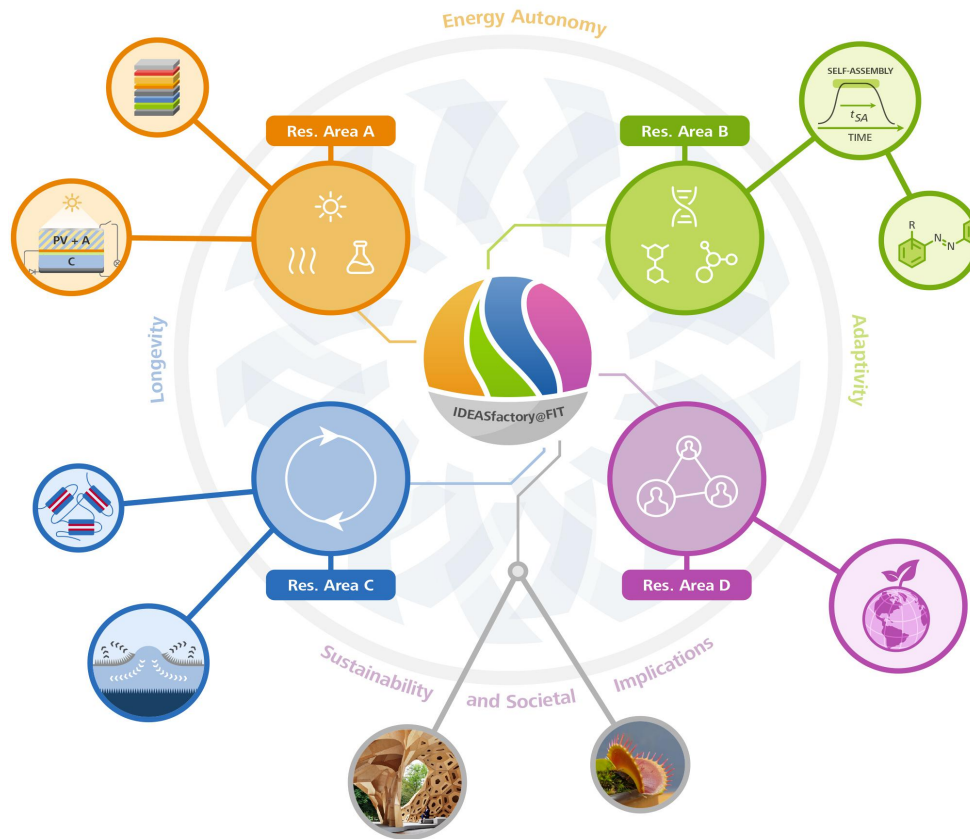


Research Data Management

best practices & *livMatS* state of the art



Johannes L. Hörmann

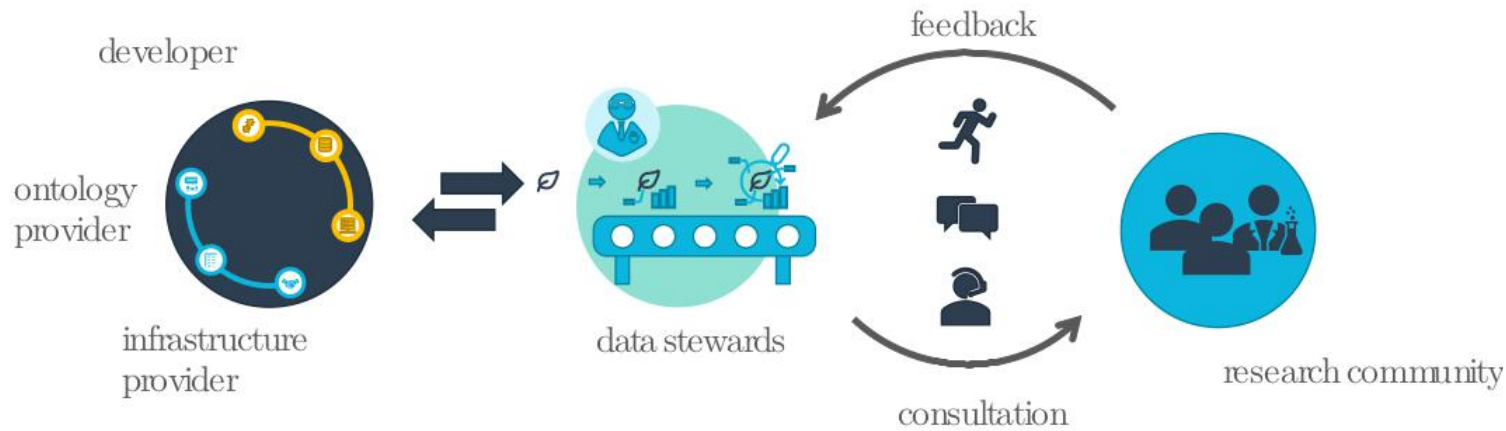
2023-03-02



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



The *livMatS* data steward



Living, Adaptive and Energy-autonomous
Materials Systems

You are here: Home › People › Management › Johannes Hörmann

People



Johannes Hörmann

Data Steward

University of Freiburg
Cluster of Excellence *livMatS* @ FIT - Freiburg Center for Interactive
Materials and Bioinspired Technologies
D-79110 Freiburg

Phone: +49 761 203 95332

Email: data@livmats.uni-freiburg.de

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

Data steward core tasks:

- Draft *livMatS* **RDM policy**
- Develop RDM **training, support**
- Software solution development

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

© Copyright *livMatS* / University of Freiburg



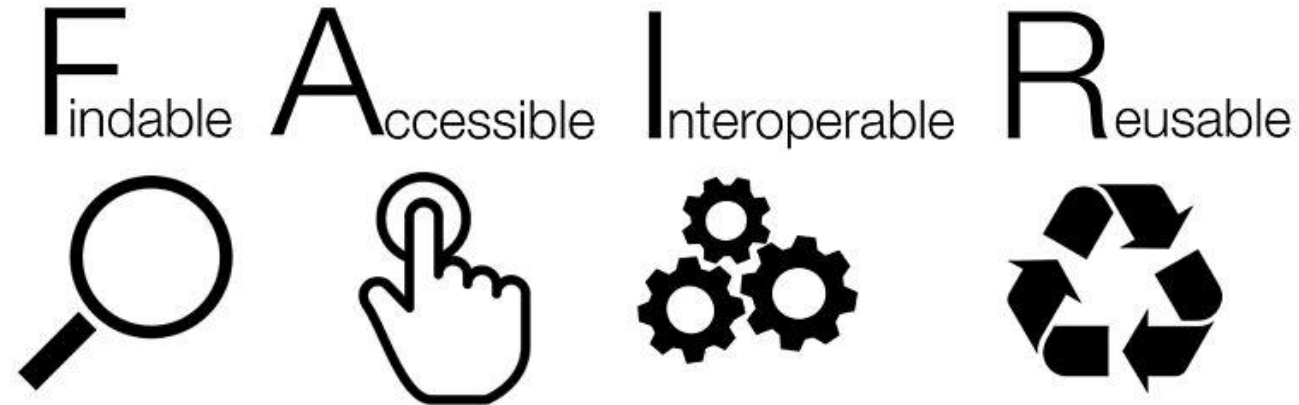
In five years from now, ...

1. will my PI be able to find, understand, and provide all (possibly unpublished) data underlying my research?
2. will administrative staff at *livMatS* and the University of Freiburg be able to find and provide all (possibly unpublished) data underlying my research?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.



Source: Wikipedia



To what end?

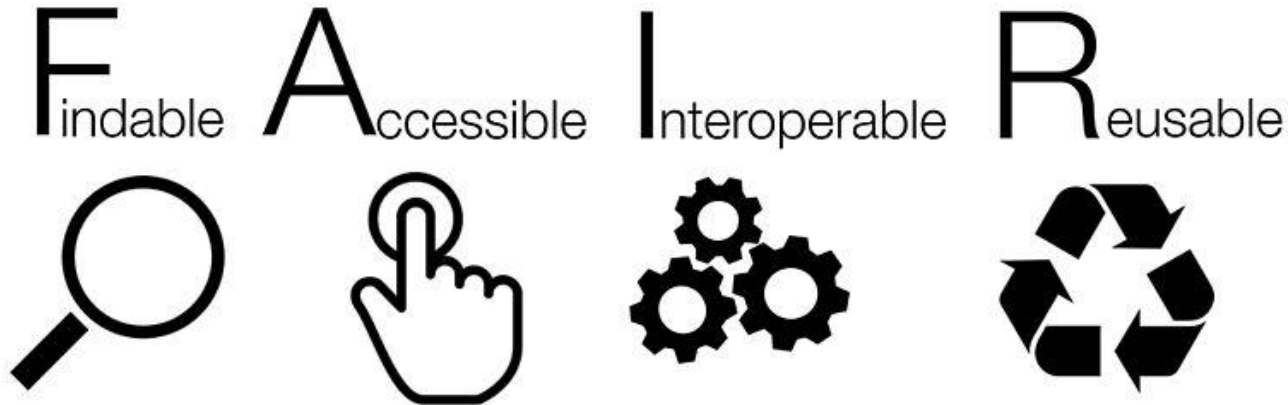
benefit to science and society?

benefit to other data scientists?

personal benefit to the researcher?



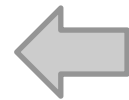
Source: Wikipedia



To what end?

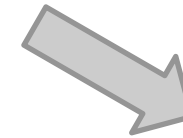
benefit to science and society

- 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



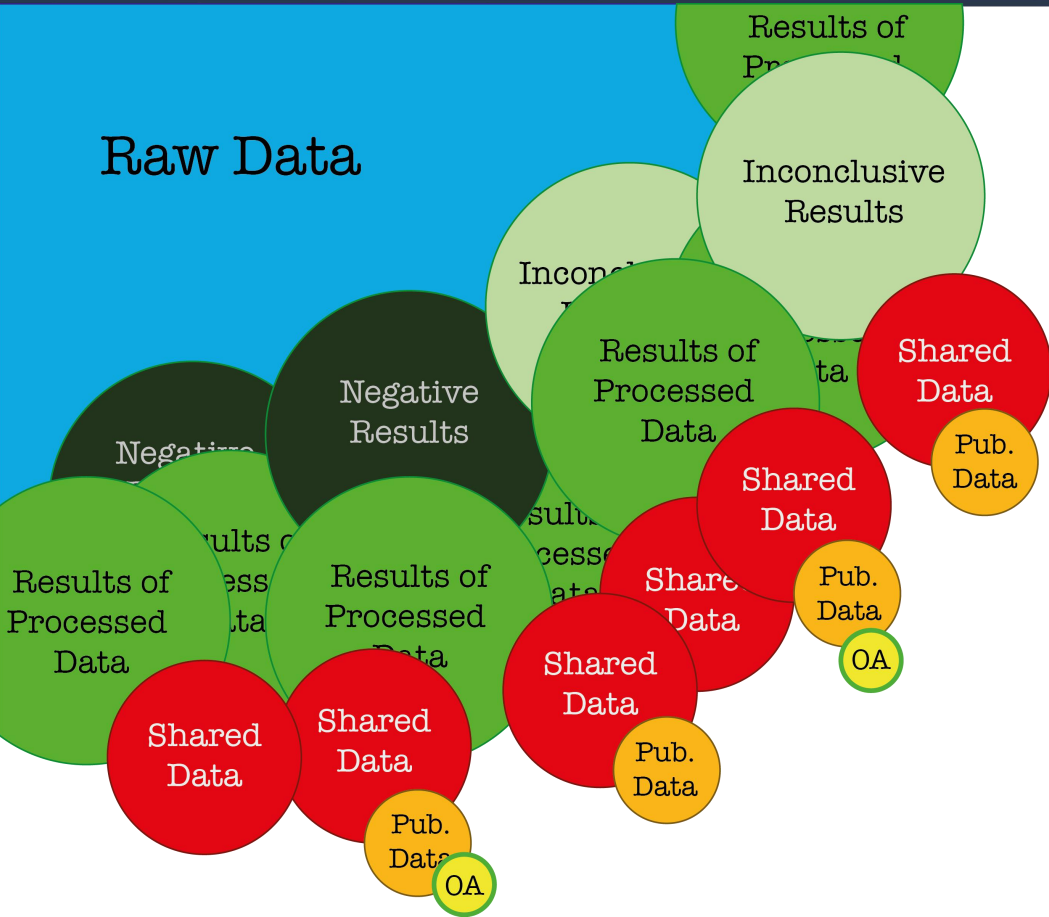
- Why data steward & data management?
- **Basic best practice recommendations**
- *livMatS* RDM examples and services



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.

UNI
FREI



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



Which files can
you delete?

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



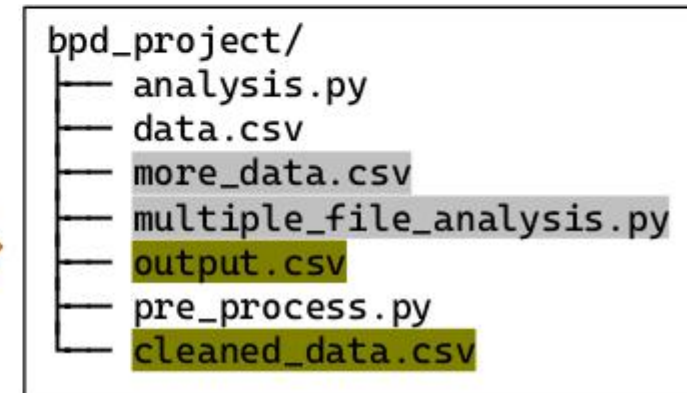
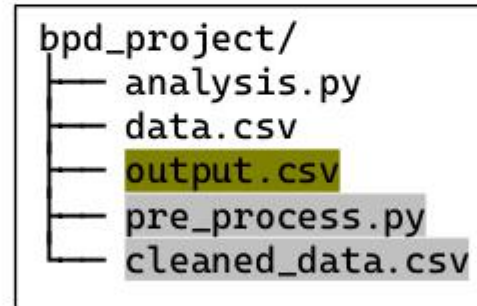
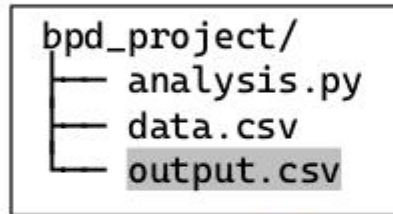
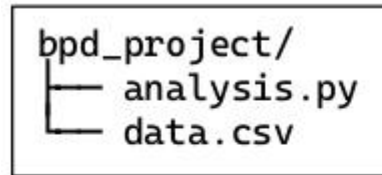
Project directory structure

Suggestion: sort data by provenance.

Legend

New file

Regenerated file



source: Tjelvar Olsson, Four principles for Effective Data Management



Which files can
you delete?

```
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
```

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



Data management in its simplest form: File naming conventions

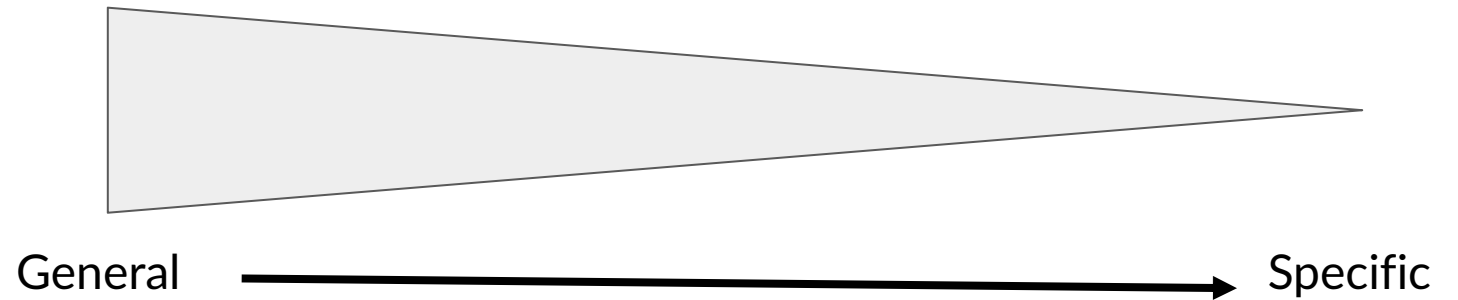


Think about meaningful file names.

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
------	---------	------------	------	----	---------



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

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



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>



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>



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 automated 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.

Reward, salience and agency in event related potentials for appetitive and aversive contexts

0.0B Public 0 ...

Contributors: Harry Stewardson
Date created: 2021-02-01 10:16 AM | Last Updated: 2021-04-12 11:39 AM
Category: Project
Description: Repository for manuscript - 'Reward, salience and agency in event related potentials for appetitive and aversive contexts'.

Name	Modified
OneDrive couldn't load.	OneDrive couldn't load...
OSF Storage (United States)	
README	
OSF Storage (United States)	
OSF README.docx	2021-06-15 01:12 PM
Data	
OSF Storage (United States)	
BVA History Files	
Materials	
OSF Storage (United States)	
Behavioural Task	
Forms	
Stimuli	

ContactEngineering / ce-container-stack Public

Container recipes for well-defined environments providing ContactEngineering tools

MIT license

0 stars 0 forks

Star Unwatch

Code Issues Pull requests 2 Actions Projects

master

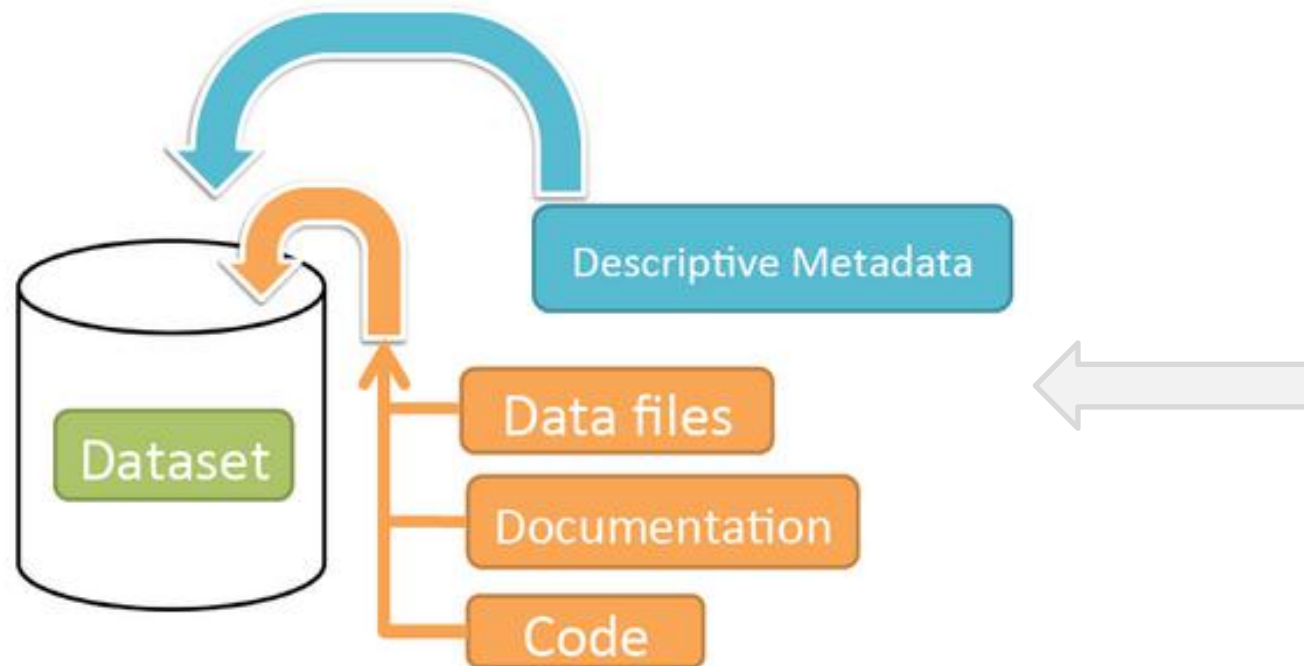
jotelha Merge pull request #10 from Conta... 2 weeks ago 27

.github	3 weeks ago
docker/jupyterlab-SurfaceTopography	2 months ago
maintenance	4 months ago
AUTHORS	4 months ago
LICENSE.md	4 months ago
README.md	2 months ago



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-readable format.



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

```
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



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



Virtue of the *livMatS* RDM policy

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

University's RDM policy

doi: [10.6094/UNIFR/231612](https://doi.org/10.6094/UNIFR/231612)

livMatS RDM policy

livmats.uni-freiburg.de/rdm

PIs and students document the policy's implementation in continuously 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

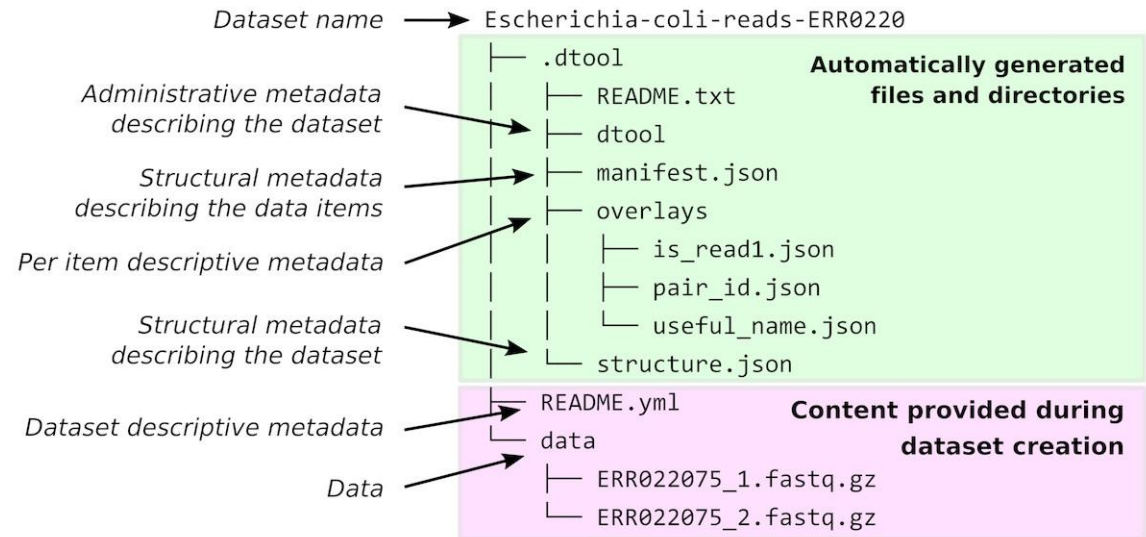
DMP = data management plan

livmats.uni-freiburg.de/rdm/dmp



- package data and metadata 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 dtol," 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

The screenshot shows a dtool GUI window with a search bar containing 'hoermann'. The left pane lists several storage locations: 'Lookup server' (10 datasets, 0.0B), 's3://test-bucket' (1 dataset, 17.0B), 'smb://test-share' (1 dataset, 17.0B), and '/C:/Users/admin/datasets' (1 dataset, 26.0B). The right pane displays a list of datasets with their IDs, names, and creation dates. The selected dataset is '175445e5-d81e-4843-bfa9-399f90a0e34b' (2021-10-01-hoermann-rdm-assessment). Below the list, a 'Dependencies' tab is active, showing a dependency graph with four nodes: a light blue circle at the top left, a light blue circle at the top right, a light blue circle at the bottom center, and a light blue circle at the bottom right. A green square node is also present, connected to the top right circle. Arrows indicate the flow of dependencies: from the top left circle to the top right circle, from the top right circle to the bottom center circle, from the top right circle to the bottom right circle, and from the bottom center circle to the bottom right circle.

The screenshot shows the Windows Start menu search results for 'dtool-lookup-gui'. The search results are categorized into 'Best match', 'Apps', 'Search the web', and 'Folders'. The 'Best match' section shows 'dtool-lookup-gui' as an App. The 'Apps' section shows 'dtool-lookup-gui.exe'. The 'Search the web' section shows results for 'dtool', 'dtools', 'dtoolsshr', and 'dtoolsit'. The 'Folders' section shows 'dtool-lookup-gui-windows', 'dtoolcore', and 'dtool-lookup-gui-windows-installer'. The 'Photos' section shows 'dtool-on-windows'. The search bar at the bottom contains 'dtool-lookup-gui'.

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 & *livMatS* 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](https://www.livmats.uni-freiburg.de) on

www.livmats.uni-freiburg.de/rdm



access for *livMatS* affiliates by registration on

livmats-data.vm.uni-freiburg.de/config,

support by data@livmats.uni-freiburg.de

livMatS 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 \(Attribution 3.0 Unported\)](https://creativecommons.org/licenses/by/3.0/)

Chemotion

All IUPAC, InChI, SMILES, ...

Logged in as Nicole Jung

Collections

- chemotion.net
- Collection 1
- BJOC data**
- Collection 2
- Collection 3
- Test Import
- My shared collections
- Shared with me
- Synchronized with me
- Inbox

219(0) 62(0) 1(0) 1(0) 1(0)

NJu-R243 According to General Procedure 2a

NJu-R242 According to General Procedure 2a

NJu-R241 According to General Procedure 2a

NJu-R240 According to General Procedure 2a

NJu-R239 According to General Procedure 2a

NJu-R238 According to General Procedure 2a

NJu-145 NJu-R24-A

NJu-R243 According to General Procedure 2a

1-0

Scheme Properties References Analyses Green Chemistry Zotero

Starting materials	Ref	T/R	Amount	Conc	Equiv		
NJu-772 2-benzyl-4,5-dihydro-1,3-dithiol-1-...	t	1000	mg	0.00 ml	3.544 mmol	708.9 mmol/l	1.000

Reactants	Reagents	T/R	Amount	Conc	Equiv			
but-3-en-2-one		t	298.1	mg	0.00 ml	4.253 mmol	850.6 mmol/l	1.200

Products	T/R	Amount	Conc	Yield			
NJu-773 NJu-R243-A 5-(1,3-dithiolan-2-ylidene)-5-phen...	r	599.8	mg	0.00 ml	2.268 mmol	453.7 mmol/l	64%

Solvents

Name: According to General Procedure 2a Status: Select... Temperature: Temperature... °C

Role: Parts of GP According to

Show 15

psychologists (Kiesel group) publish data on OSF



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](https://doi.org/10.17605/OSF.IO/UKY3M) ARK c7605/osf.io/uky3m
Category: Project

Files

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



Research Data Management Group

<https://rdmg.uni-freiburg.de/>



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

FreiDok *plus*
Universitätsbibliothek Freiburg

publication platform

<https://freidok.uni-freiburg.de/>



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.

- *livMatS*
 - research data management: livmats.uni-freiburg.de/rdm
 - on github: github.com/livMatS
- Johannes L. Hörmann
 - on LinkedIn: linkedin.com/in/jotelha
 - on github: github.com/jotelha