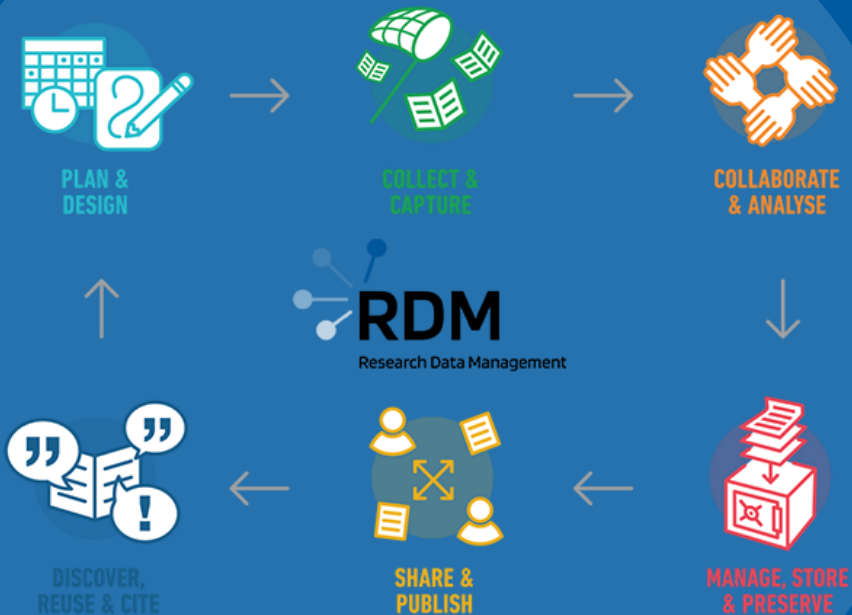


Introduction to Research Data Management

Ronny Gey (RDM)
Desireé Jörke (RDM)

2023-09-21



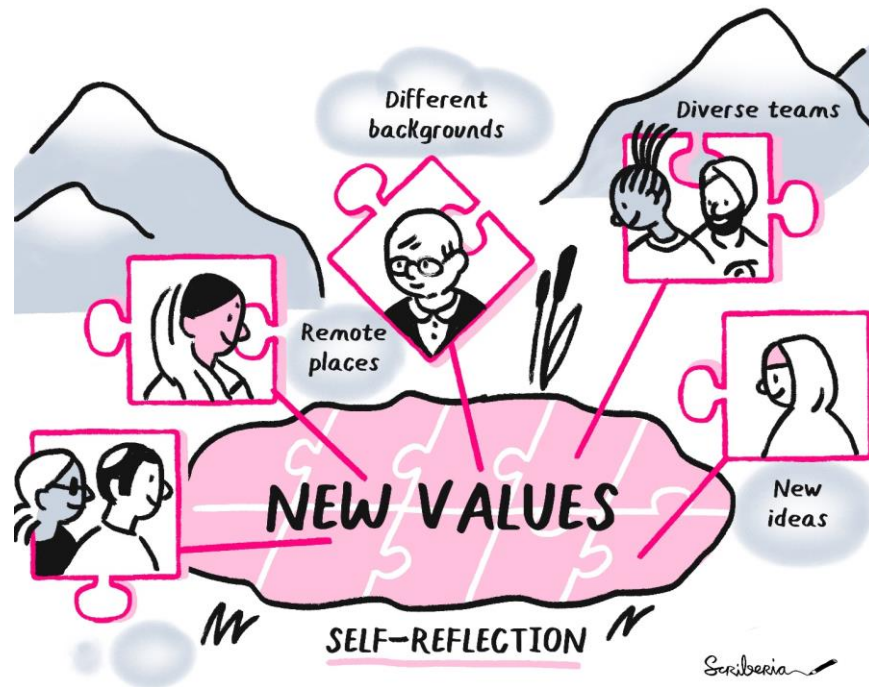
3 main strands:

- **development** and maintenance of data infrastructures, services and tools
- scientific **data processing/** data pipelines
- **consulting**, training and networking



The RDM Team @ UFZ 08/2023 (members missing)

- What are **data** you typically work with?
- What **research methods** do you typically apply?
- What do you want to **learn today**?



Graphic: [Scriberia](#) under [CC-BY 4.0](#)
licence, doi: [10.5281/zenodo.3332807](https://doi.org/10.5281/zenodo.3332807)

1. Welcome and orientation
2. Data lifecycle / RDM / FAIR / open
3. Policies / data management plans

Break

4. Metadata
5. Storage / backup / archiving
6. Data publication / reproducibility / legal aspects / discovery & reuse

Break

7. Research software
8. Institutional infrastructures, services and tools
9. Q&A

Research Data Management (RDM)

What are “research data”?

“Data **collected or produced** in the course of scientific research activities and used as **evidence** in the research process, or commonly accepted in the research community as necessary to **validate** research findings and results.”

European Open Science Cloud Glossary [1]

“Research data might include **measurement data, laboratory values, audiovisual information, texts, survey data, objects from collections, or samples** that were created, developed or evaluated during scientific work. Methodical forms of testing such as **questionnaires, software and simulations** may also produce important results for scientific research and should therefore also be categorised as research data.”

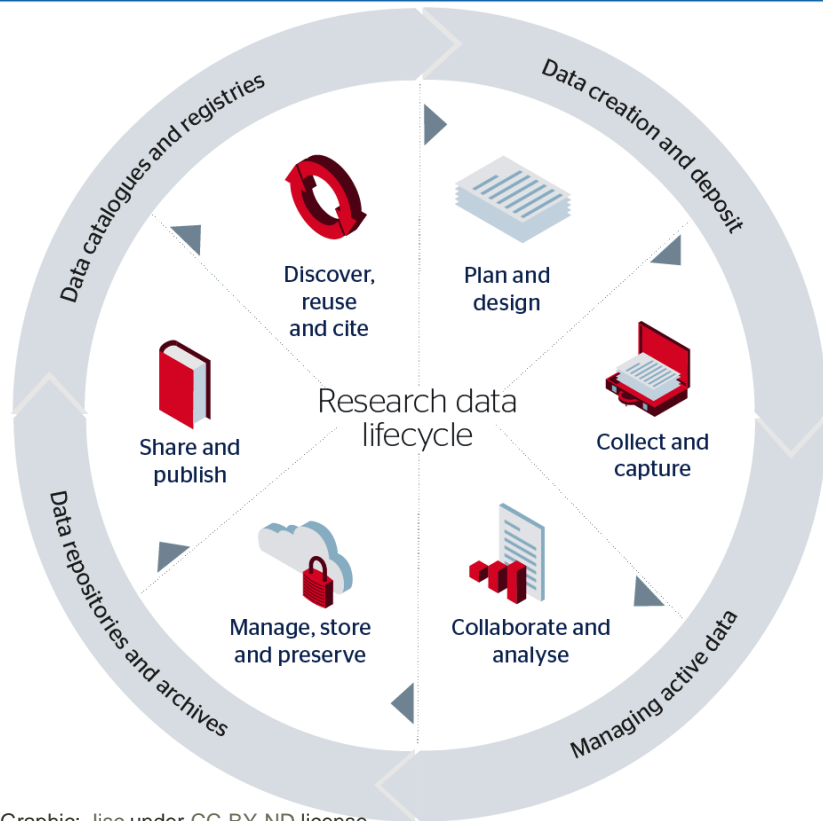
DFG Guidelines on the Handling of Research Data [2]



[1] EOSC Glossary. <https://eosc-portal.eu/glossary>

[2] Deutsche Forschungsgemeinschaft.

https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/guidelines_research_data.pdf



- This includes decisions on, e.g.:
 - Data structure and naming
 - Data transfer and conversion
 - Deployed software, infrastructures and tools
 - Actors and responsibilities
 - Rights and licenses

Why is following good RDM practices so important in science?
(Answers from previous courses)



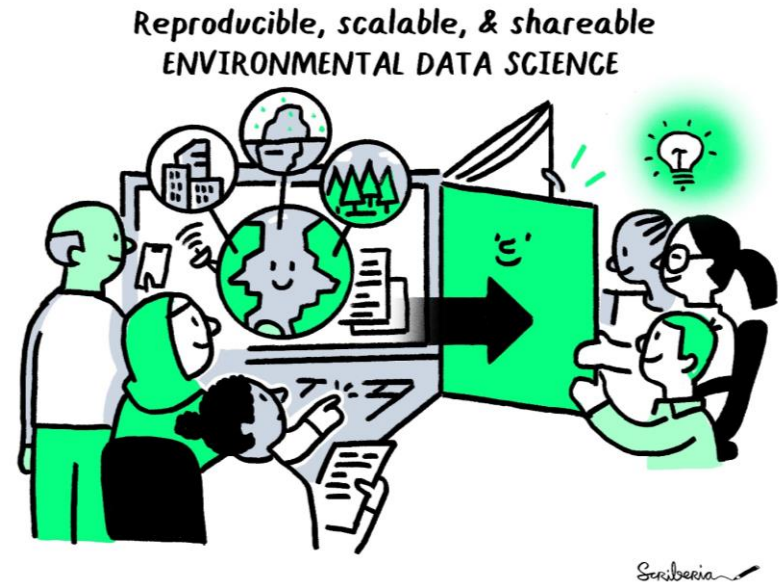
Scriberia

Graphic: [Scriberia](https://doi.org/10.5281/zenodo.3332807) under CC-BY 4.0 licence, doi: [10.5281/zenodo.3332807](https://doi.org/10.5281/zenodo.3332807)

Research data management

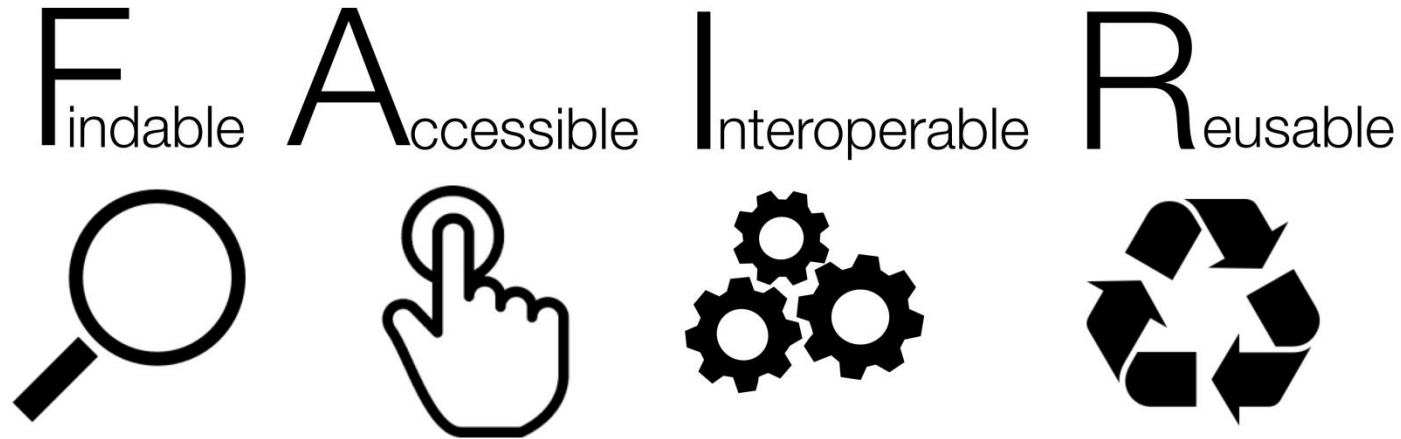
Why is following good RDM practices so important in science?

- faster retrieval of data
- evidence of good scientific practice /audit
- knowledge preservation - independently of individual people, projects or institutions
- transfer of data to future projects
- facilitation of collaboration / research synthesis
- long-term traceability of results, instead of new creation
- prevents loss of data
- (semi-)automatic processing enabled by metadata
- optimized use of resources
- third-party funder requirement
- research data citation
- replicability / reproducibility
- increased relevance by increased visibility



Graphic: Scriberia under [CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/)
licence, doi: [10.5281/zenodo.3332807](https://doi.org/10.5281/zenodo.3332807)

FAIR and OPEN



[1]

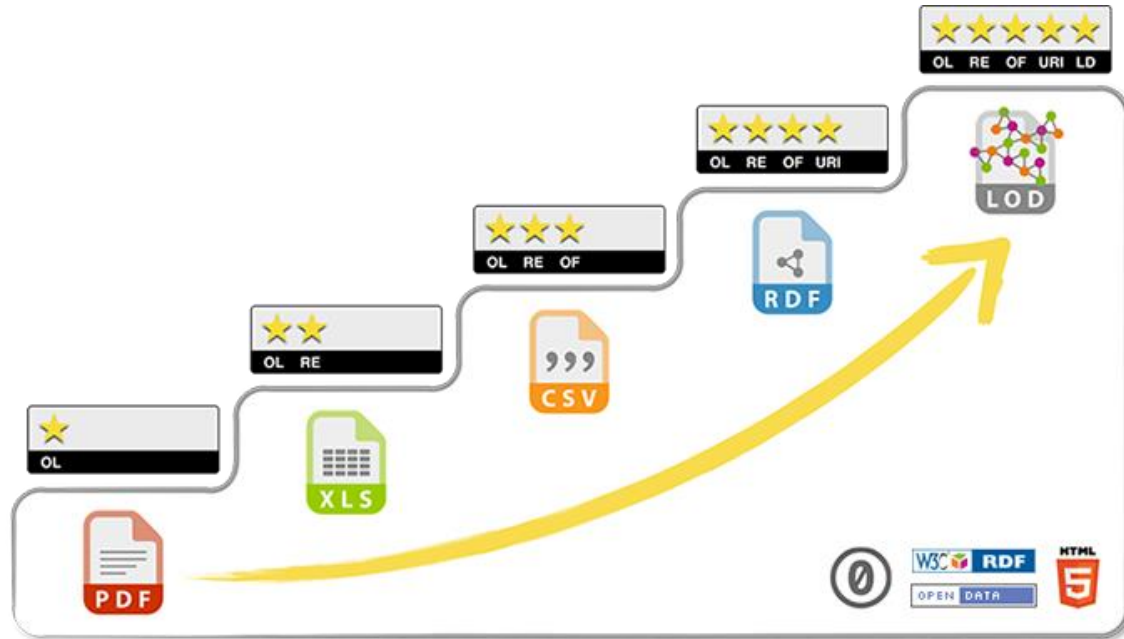
=> enhance the suitability for reuse, **by humans** and at scale **by machines**

=> focus on machine-actionability

[Graphic: [Pundir, Sangya](#) under [CC-BY-SA-4.0](#) license]

[1] Wilkinson, M. D. et al. (2016). <https://doi.org/10.1038/sdata.2016.18>

- making research data publicly available, accessible and reusable with minimal restrictions
 - open license
 - machine-readable
 - non-proprietary format
 - open standards
 - linked to other data
- Tim Berners-Lee's 5-stars of Linked Open Data [1]

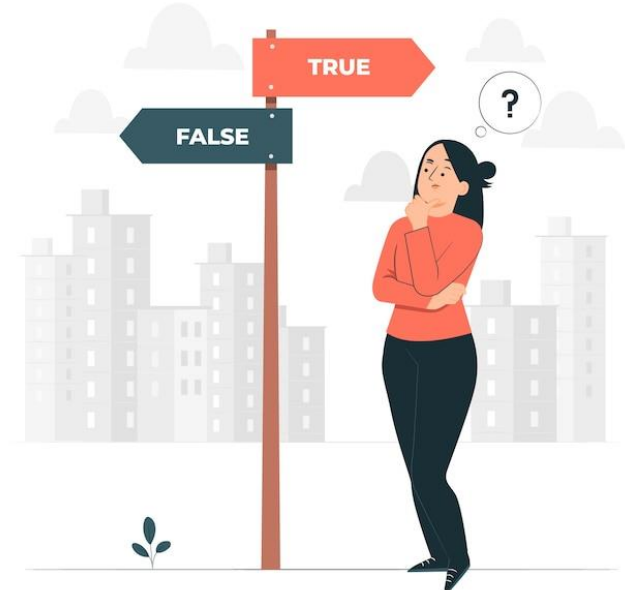


[Graphic: Kim & Hausenblas under CC0 license]

[1] Berners-Lee, B. <https://www.w3.org/DesignIssues/LinkedData.html> (accessed 2021-02-01)

Make a choice:

“Which of the following statements related to the FAIR principles do you agree with?”



[Graphic: Image by storyset on Freepik under [CC-BY 4.0 license](https://www.freepik.com/free-vector/illustration)]

Make a choice: *“Which of the following statements related to the FAIR principles do you agree with?”*

1) Only open data can also be FAIR.

I agree

I disagree

Make a choice: *“Which of the following statements related to the FAIR principles do you agree with?”*

1) Only open data can also be FAIR.

- accessibility does not imply that data should always be open
- access might be constrained due to legitimate concerns around
 - privacy
 - national security
 - commercial interests
- explain why the data is not open and what the conditions for access are



Focus on FAIR when data cannot be open!

Make a choice: *“Which of the following statements related to the FAIR principles do you agree with?”*

2) Good RDM is a prerequisite for data that comply with FAIR principles.

I agree

I disagree

Make a choice: *“Which of the following statements related to the FAIR principles do you agree with?”*

2) Good RDM is a prerequisite for data that comply with FAIR principles.

- Good RDM is a necessary pre-cursor for FAIR and open
- Produce data which is fit for sharing and reuse
- Many decisions affect the potential for data - these can include
 - research project roles and responsibilities
 - consent agreements
 - data ownership and use agreed with partners
 - licences from third-party data owners
 - data format choices
 - metadata schema choices
 - naming conventions
 - and the creation or capture of metadata and data documentation

Make a choice: *“Which of the following statements related to the FAIR principles do you agree with?”*

3) If data is open and FAIR, good data quality is also guaranteed.

I agree

I disagree

Make a choice: *“Which of the following statements related to the FAIR principles do you agree with?”*

3) If data is open and FAIR, good data quality is also guaranteed.

- fair and open is not a guarantee of good quality data
- Data quality also depends on other factors in the collection and analysis

**Neither FAIR nor open data are a reflection of data quality.
Both are simply a measure
of how data have been made available.**

Make a choice: “Which of the following statements related to the FAIR principles do you agree with?”

4) Open data is more useful than FAIR data

I agree

I disagree

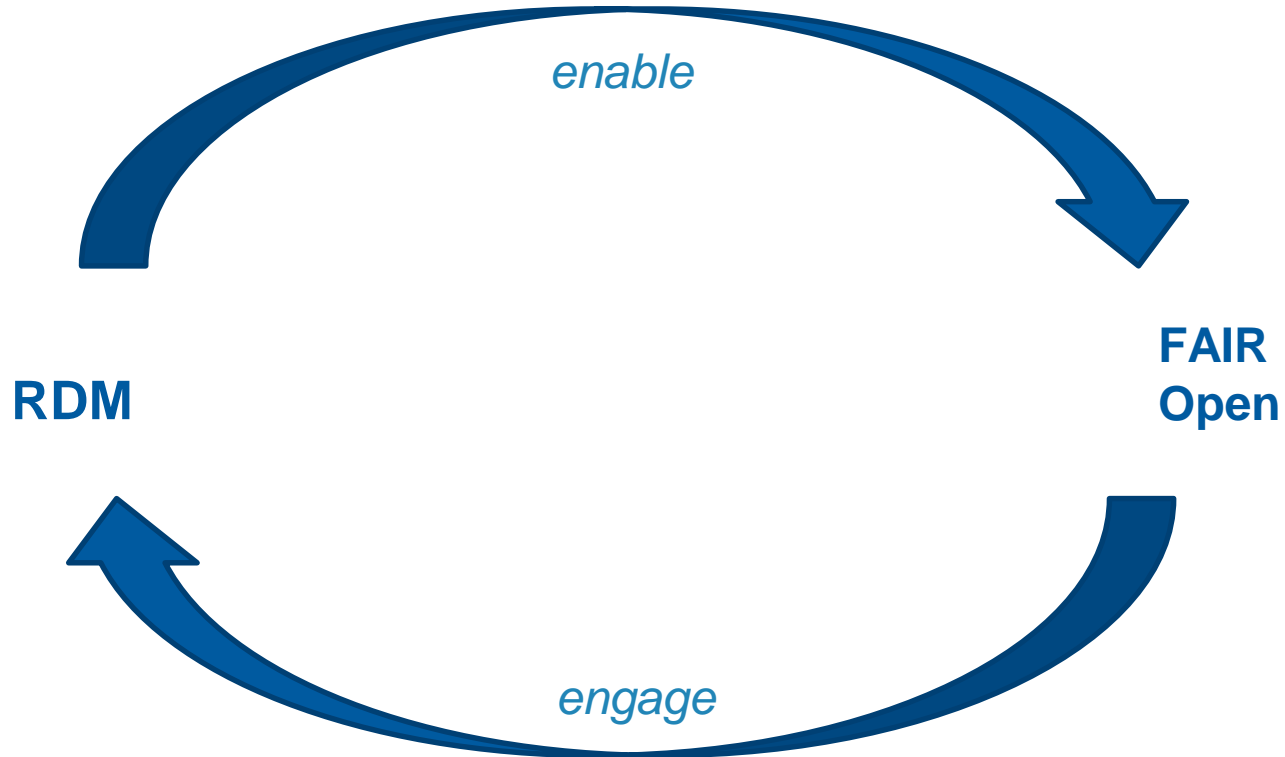
Make a choice: “Which of the following statements related to the FAIR principles do you agree with?”

4) Open data is more useful than FAIR data

- These concepts are not in competition
- the most reusable data will be
 - well documented
 - conform to community standards
 - be as free from restrictions as possible to increase potential reuse



make your data as FAIR and open as possible!



Policies

- journal and publisher policies
- institutional policies
- project-specific policies
- domain-specific policies
- funder policies



- journal and publisher policies
 - institutional policies
 - project-specific policies
 - domain-specific policies
 - funder policies
- **The Transparency and Openness Promotion Guidelines** [1] (The Transparency and Openness Promotion (TOP) Committee)
 - 8 Standards, Level 0 to 3

[1] <https://www.science.org/doi/10.1126/science.aab2374>

- journal and publisher policies
 - institutional policies
 - project-specific policies
 - domain-specific policies
 - funder policies
- UFZ-Regulation | IR-5/18 | “Principles for the Responsible Handling of Research Data at UFZ”
 - UFZ-Guideline| IR-17/12| “Guidelines for safeguarding good scientific practice in the UFZ”

- journal and publisher policies
 - institutional policies
 - project-specific policies
 - domain-specific policies
 - funder policies
- Scientific policies of CESSDA (Consortium of European **Social Science** Data Archives) [1]
 - DFG-Guidelines on the Handling of Research Data in **Biodiversity Research**
 - OECD Principles of Good **Laboratory Practice**

[1] <https://www.cessda.eu/About/Documents-and-Policies>

- journal and publisher policies
- institutional policies
- project-specific policies
- domain-specific policies
- funder policies
 - EC Guidelines on FAIR Data Management in Horizon 2020
 - DFG Guidelines on the Handling of Research Data



[For more data policies: https://www.forschungsdaten.org/index.php/Data_Policies (accessed 2021-02-02)]

Graphic: CC-BY: <https://aukeherrema.nl>

- journal and publisher policies
- institutional policies
- project-specific policies
- domain-specific policies
- funder policies



think from the end!

- ask yourself the Questions:
 - Are you a member of a research institution?
 - Do you apply for a research grant?
 - Are you planning to publish in a specific journal?

Data Management Plan (DMP)

All information that adequately describes and documents the **collection**, **processing**, **storage**, **archiving**, and **publication** of research data as part of a research project.



<https://doi.org/10.5281/zenodo.1421739>

All information that adequately describes and documents the **collection, processing, storage, archiving, and publication** of research data as part of a research project.

- consider different funder requirements
- from paragraphs to multiple pages
- from static to living documents
- differing reporting dates

HIGHLY RECOMMENDED!

Data management plan (DMP)

What should be included in a DMP?

- Administrative information
- Project and data set descriptions
- Data types, formats, scope
- Metadata and standards information
- Data sharing and legal / ethical requirements
- Archiving and backup of data
- Responsibilities
- Costs



Think the process backwards

Consider **where and how** the data will be **archived or published**? These choices require setting the course early in the data management workflow, e.g., formats, standards, metadata, licenses, etc.

Data management plan (DMP)

What tools can assist you?

- Various tools available with slightly different features



<https://rdmo.forschungsdaten.info/>



<https://argos.openaire.eu>



<https://researchers.ds-wizard.org/>



<https://www.gfbio.org/plan>

- Templates very detailed, use as nucleus
- Explore which suites you most, e.g. regarding questionnaire structure

Data management plan (DMP)

Where do I get more information?

RDM Guidelines: <https://rdm.pages.ufz.de/guidelines/>

RDM Consulting: Drop us an Email! wkdv-datamanagement@ufz.de



Time for a  10 min break!

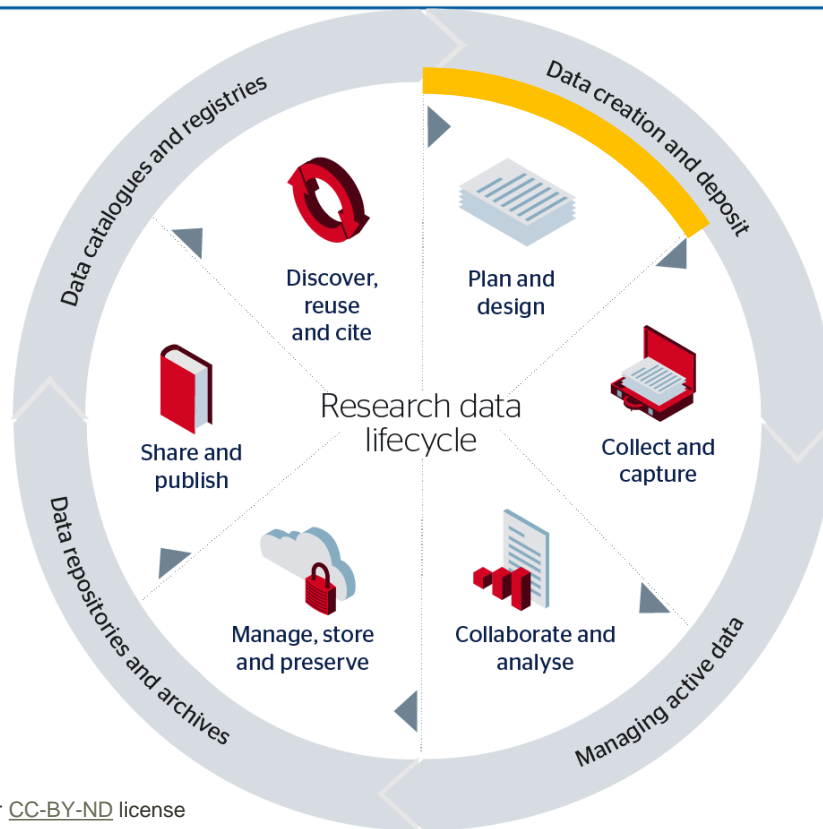
1. Welcome and orientation
2. Data lifecycle / RDM / FAIR /open
3. Policies / data management plans

Break

4. Metadata
5. Storage / backup / archiving
6. Data publication / reproducibility / legal aspects / discovery & reuse

Break

7. Research software
8. Institutional infrastructures, services and tools
9. Q&A



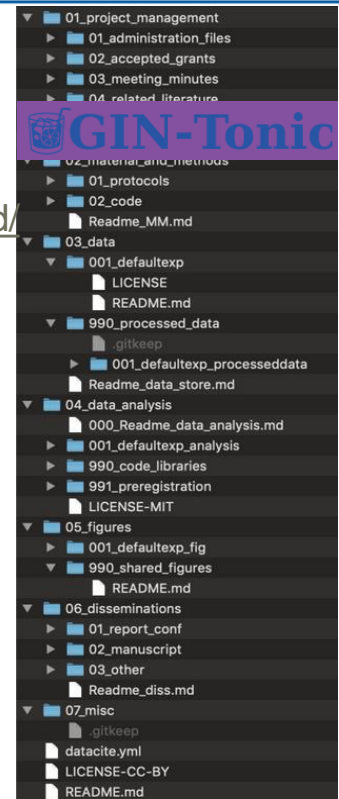
Organisation and Versioning

- consistent folder and file naming conventions in your project/lab

=> take a shot of

<https://gin-tonic.netlify.app/standard/>

```
20160512_Climate_measurement_1_original.jpg  
20160522_Climate_measurement_1_MHU_cutout.jpg  
20160523_Climate_measurement_1_MHU_cutout_edited_color.j  
pg
```



- consistent folder and file naming conventions in your project/lab
- use cloud storage for collaborative work (files versioned)
- use versioning for data, coding and software development
- check HIFIS for federated IT infrastructures and support

=> take a shot of <https://gin-tonic.netlify.app/standard/>



=> e.g. nc.ufz.de or <https://nubes.helmholtz-berlin.de/>



=> <https://git.ufz.de/> or <https://codebase.helmholtz.cloud>



Metadata

is:

- **structured information** about data
 - partial amount of documentation information
 - possibly data itself
 - essential for FAIRness of data
- should be:
- human and machine readable
 - searchable
 - linked
 - standardized



< metadata >

- exist for different scopes
- *domains-agnostic* → Dublin Core, DataCite, PROV, MODS
- *domain specific* → An overview of discipline-specific metadata standards: British Digital Curation Centre and in an overview of the Research Data Alliance. The Helmholtz Metadata Collaboration (HMC): query metadata schemes based on the subject area, e.g. Earth and Environment.
 - EML (ecology)
 - ISO19115 (geoscience)
 - ABCD (specimen collection)
 - DDI (social science)
 - MIxS (genomics)
 - CIM (climatology)



it is good practice to not invent your own schema

often require and support to:

- use terminologies (controlled vocabularies, thesauri, ontologies)
- use standards for names of languages, countries, date/time
- use persistent identifiers to link to information

Data repositories and metadata catalogues support generic and domain-specific standards, e.g.:

=> <https://geonetwork.ufz.de/>

=> <https://bexis.ufz.de>

=> [UFZ Data Archive](#)

Persistent Identifiers (PIDs)

Persistent identifiers

PIDs are different in :

- purpose
- scope
- underlying technology
- governance
- metadata collected
- cost
- extend of use



Wikidata item identifier



PURL



ARK

URI



Persistent identifiers



PIDs support:

- discovery
- disambiguation
- credit
- tracking, linking, connecting
- automating compliance
- reproducibility
- meta science



Graphics: [Scriberia](#) under a CC-BY license. DOI: [10.5281/zenodo.3332807](https://doi.org/10.5281/zenodo.3332807)

ORCID @ UFZ



ORCID
Connecting research and researchers

English

ABOUT FOR RESEARCHERS MEMBERSHIP DOCUMENTATION RESOURCES
NEWS & EVENTS SIGN IN

ORCID ID

Distinguish yourself in a few easy steps

ORCID provides a persistent digital identifier (an ORCID iD) that you own and control. It distinguishes you from others and connects your professional information — including grants, publications, and other work — to a central system, ensuring you get recognition for your contributions, saving you time and hassle, and reducing the risk of errors.

LEARN MORE ABOUT OUR MISSION AND VALUES

<https://orcid.org>



UFZ HELMHOLTZ Centre for Environmental Research

UFZ ORCID Connect

User Name: geyr
Full Name: Ronny Gey




Thank you very much for connecting your ORCID profile.

ORCID ID: <https://orcid.org/0000-0003-1028-1670>

Do you want to add a link to your ORCID profile to your UFZ website?
<https://orcid.org/0000-0003-1028-1670>
Then use the following button to copy the HTML code and add it to a "HTML code" element at your UFZ website.

Feedback
Please send us your ideas about how you use your ORCID profile and how we can support you.
paul.meissner@ufz.de

ORCID Profile vs UFZ Publication Repository

Found in UFZ Repository 2	Found in your ORCID Profile 41	Found in ORCID Profile and UFZ Repository (0)
Found in ORCID Profile only (41)		
10.5281/zenodo.7296954		
Type: other		
Title: Research Software		
		
10.5281/zenodo.7296953		
Type: other		
Title: Research Software		
		

You got no ORCID yet?

Get one in the next break! :)

<https://www.intranet.ufz.de/orcid-app/>



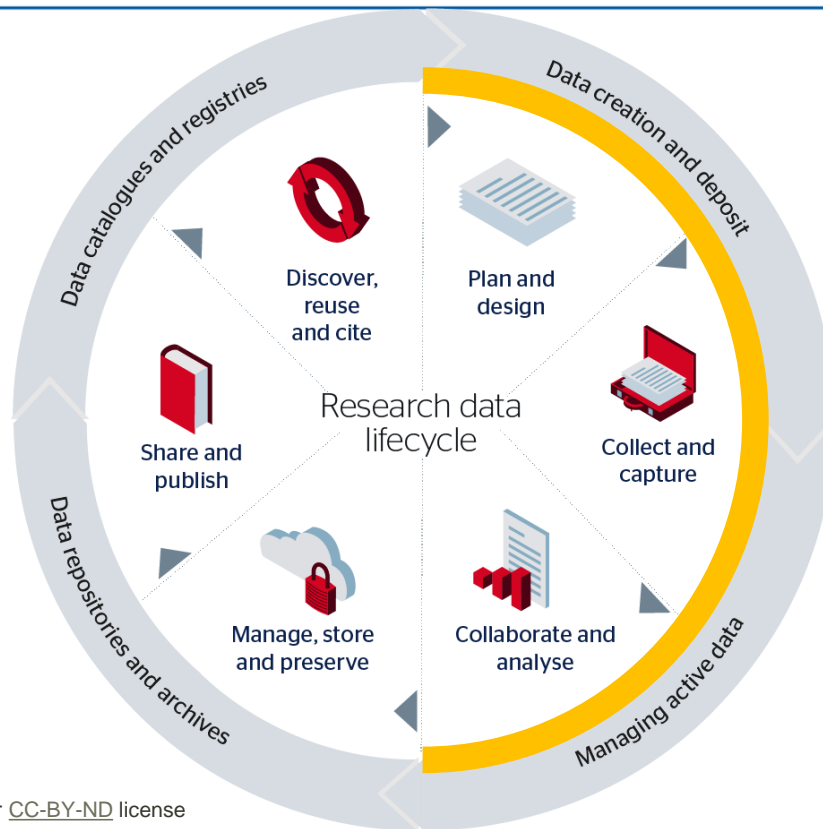
Metadata & PIDs

Where do I get more information?

RDM Guidelines: <https://rdm.pages.ufz.de/guidelines/>

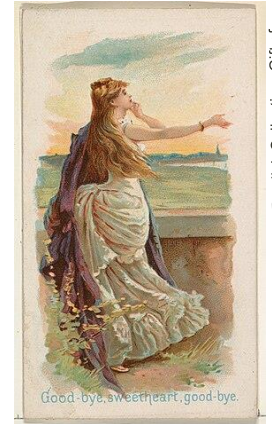
RDM Consulting: Drop us an Email! wkdv-datamanagement@ufz.de





Storage, Backup & Archiving

- **Offboarding!** Develop a strategy to ensure data access when people leave
- use secure options at UFZ [1]:
 - use institutional **network/cloud storage** (“y:/" drive, UFZ Cloud)
 - **databases** (PostgreSQL, Oracle, MySQL)
 - and **file transfer** options (SFTP data transfer)
- use sustainable (preferably **open**) **data/file formats**
- if applicable, calculate costs in project proposals



Graphic: The Jefferson R. Burdick Collection, Gift of Jefferson R. Burdick, CC0 license

[1] More on UFZ storage options: <https://www.intranet.ufz.de/index.php?en=45095>

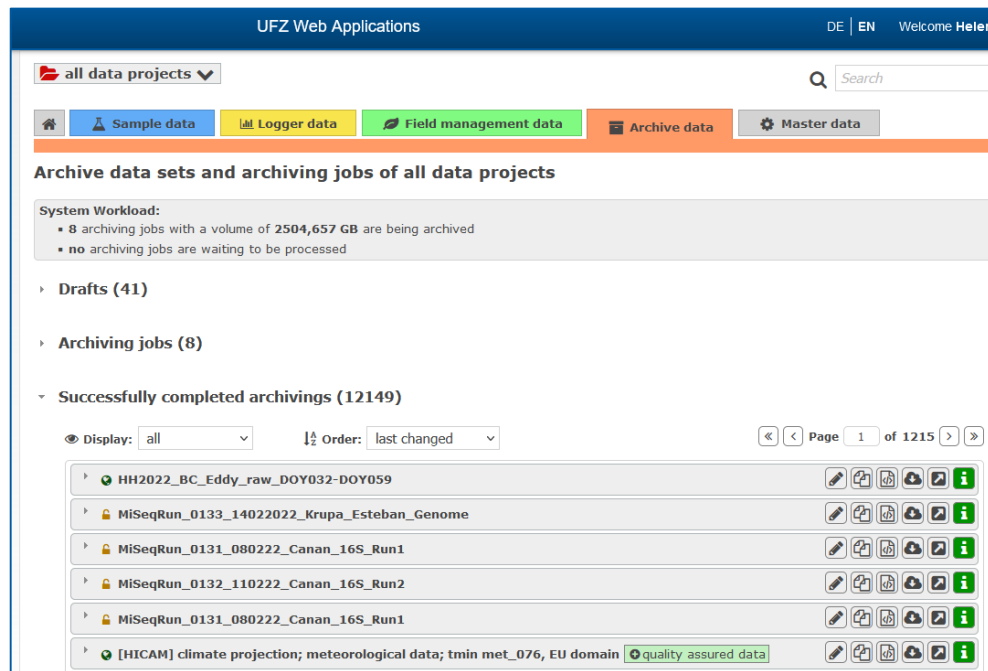
BACKUP

- (automatic) backup of **all data** to prevent data loss (technical, e.g. defective, or human, e.g. accidentally deleted)
- all versions

ARCHIVING

- preservation of **selected data** to keep it for the long term
- integrity backup
- long-term storage
- searchable

How long is “long-term”?



UFZ Web Applications DE | EN Welcome Helen

all data projects

Sample data Logger data Field management data Archive data Master data

Archive data sets and archiving jobs of all data projects

System Workload:

























- 8 archiving jobs with a volume of 2504,657 GB are being archived
- no archiving jobs are waiting to be processed

› Drafts (41)

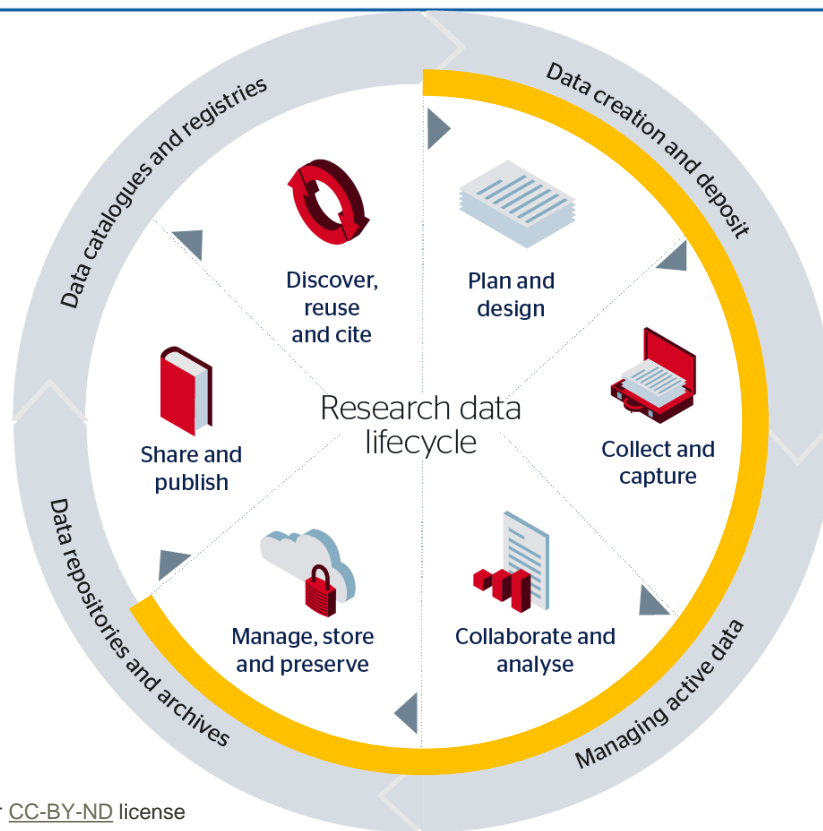
› Archiving jobs (8)

› Successfully completed archivings (12149)

Display: all Order: last changed Page 1 of 1215

› HH2022_BC_Eddy_raw_DOY032-DOY059	   
› MiSeqRun_0133_14022022_Krupa_Esteban_Genome	   
› MiSeqRun_0131_080222_Canan_16S_Run1	   
› MiSeqRun_0132_110222_Canan_16S_Run2	   
› MiSeqRun_0131_080222_Canan_16S_Run1	   
› [HICAM] climate projection; meteorological data; tmin met_076, EU domain quality assured data	   

UFZ Data Archive: <https://www.intranet.ufz.de/dmp>



Graphic: [Jisc](#) under [CC-BY-ND](#) license

Data Publication

Different options:

- Journal = Paper + PID + [supplement]
- Data Journal = Paper + Datasets + Metadata + PID



- Data repository = Datasets + Metadata+ PID

Data repositories - interdisciplinary



<https://zenodo.org/>



<https://osf.io/>

—> Comparison of generic repositories:
<https://doi.org/10.5281/zenodo.3946720>



<https://b2share.eudat.eu/>



<https://datadryad.org>

Data repositories - domain specific, e.g.

- PANGAEA (<https://pangaea.de>)
for earth science/ environmental data
- EarthChem (<http://www.earthchem.org/>)
for geochemical, petrological data
- World Data Center for Climate
(<https://www.dkrz.de/up/systems/wdcc>) for climate data
- SowiDataNet (<https://data.gesis.org/sharing>)
for social and economic data
- NORMAN databases (<https://www.norman-network.com/>)
for substances in the environment
- gfbio data centres (<https://www.gfbio.org/>)

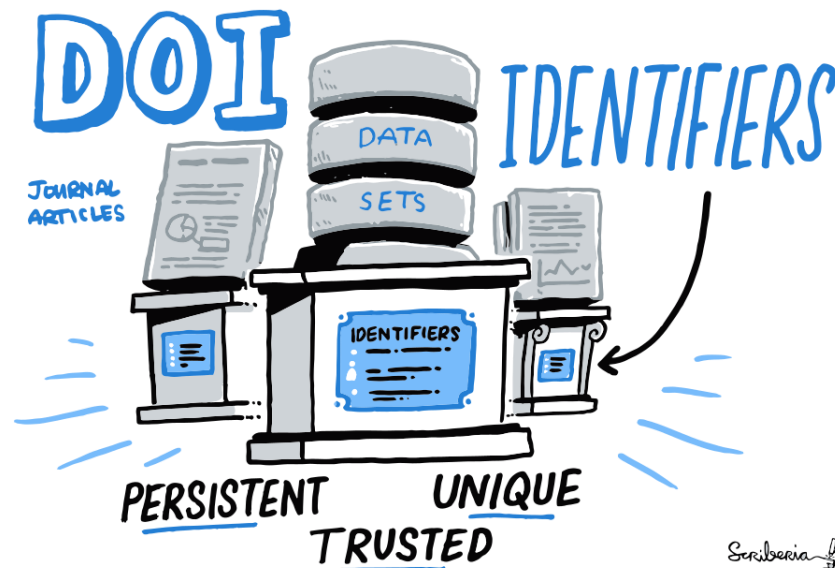
Find a suitable one at:



re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

<https://re3data.org/>

- Digital Object Identifier
- Semantic relationships between research artefacts
- 10.1234/abc123
- 10.48440/OS.HELMHOLTZ.031
- 10.48758/ufz.10591
- 10.5281/zenodo.5888548
- 10.1371/journal.pcbi.1008549
- 10.1007/978-3-030-50436-6_33



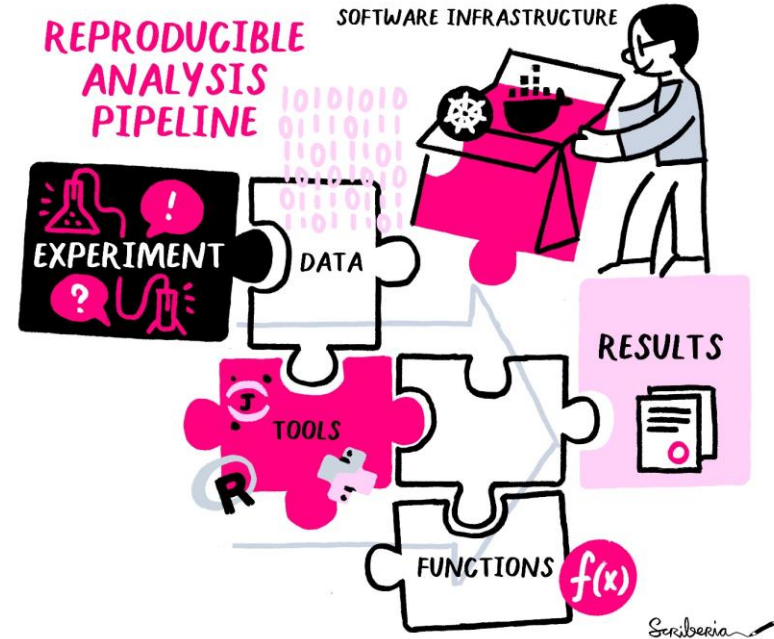
Acknowledge contributors to data and software



Graphic: [Scriberia](#) under [CC-BY 4.0](#) licence, doi: [10.5281/zenodo.3332807](https://doi.org/10.5281/zenodo.3332807)

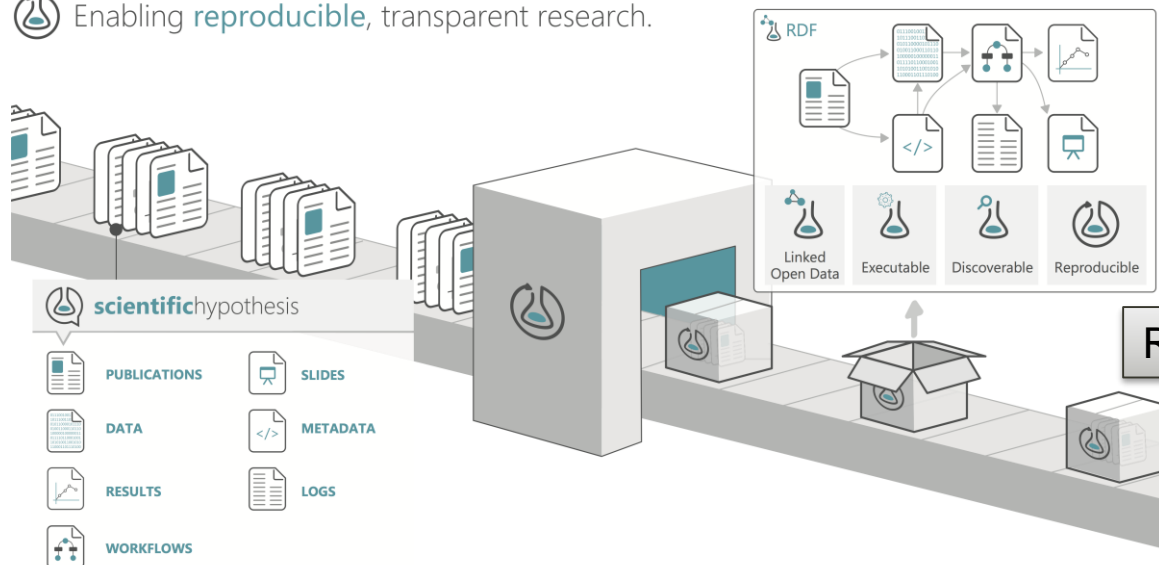
- Going beyond publishing data, publish also:
 - Scripts
 - Tools
 - Methods
 - Documentation
 - Negative results
 - ...

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable



Graphic: Scriberia under [CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/) licence, doi: [10.5281/zenodo.3332807](https://doi.org/10.5281/zenodo.3332807)

Enabling **reproducible**, transparent research.



Reproducible workflows

Containers

Provenance

Research bundles

Reproducible Environments

- Diverse approaches exist to target reproducibility, e.g. RO-Crate
- A recommended read: Guide for Reproducible Research

GRN – German Reproducibility Network
<https://reproducibilitynetwork.de/>

Data publication, DOI and reproducibility

Where do I get more information?

RDM Guidelines: <https://rdm.pages.ufz.de/guidelines/>

RDM Consulting: Drop us an Email! wkdv-datamanagement@ufz.de



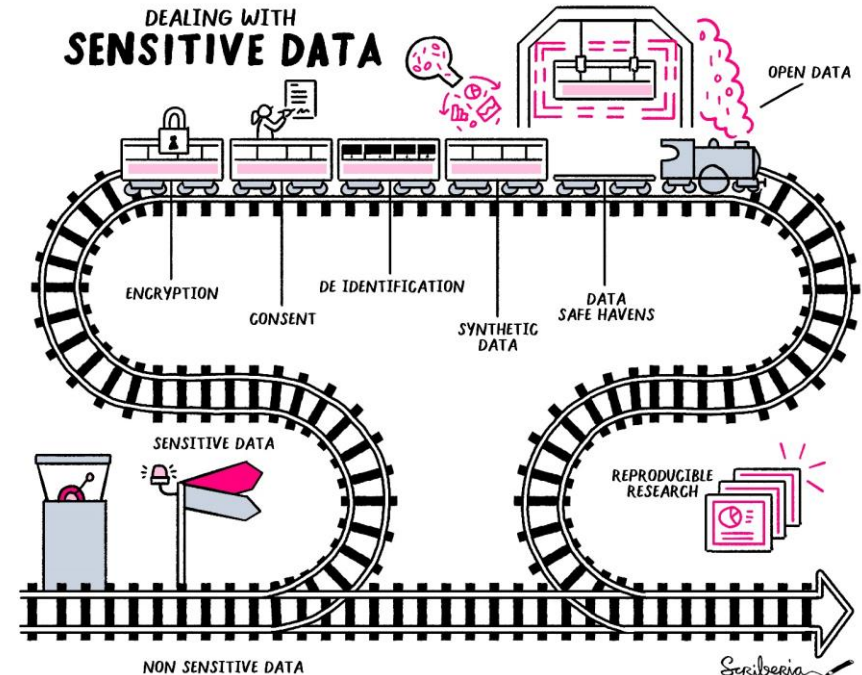
Legal Aspects

Patent law <ul style="list-style-type: none">• What has to be considered if RD (can) reach patent maturity?	Copyright law <ul style="list-style-type: none">• Are RD subject to copyright law at all?	Competition law <ul style="list-style-type: none">• Is data used unfairly in business transactions?	Data protection <ul style="list-style-type: none">• Which RD is „worthy of protection“?
Science law <ul style="list-style-type: none">• Can licensing and publication requirements for RD be mandated?	Constitutional rights <ul style="list-style-type: none">• Which constitutional limits have to be considered?	International law <ul style="list-style-type: none">• Which legal regulations exist outside the country?	EU law <ul style="list-style-type: none">• What consequences has e.g. the "European Data Economy" for RD?
Contracts <ul style="list-style-type: none">• Are there any agreements on the „intellectual property“ of RD?	Labor/service law <ul style="list-style-type: none">• Who „owns“ the RD that is collected at UFZ?	Funding requirements <ul style="list-style-type: none">• Which terms and conditions are set by funders (EU; industry)?	Policies <ul style="list-style-type: none">• Which legal obligations can policies develop?

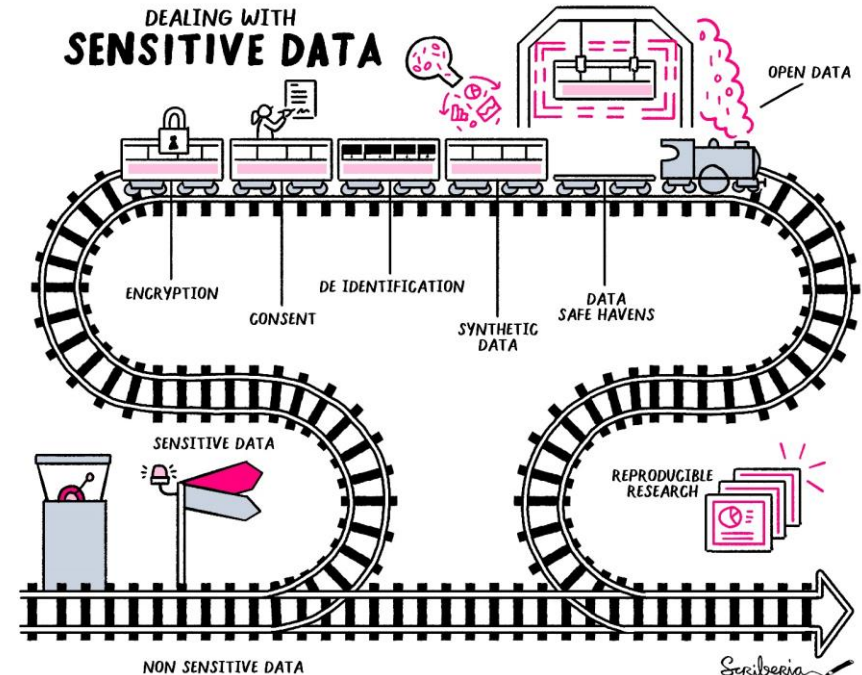
Source: translation of Hartmann, Thomas. (2019). *Rechtsfragen: Institutioneller Rahmen und Handlungsoptionen für universitäres FDM*. Zenodo. <https://www.doi.org/10.5281/zenodo.2654306>.

e.g. personal data, biodiversity data, confidential data

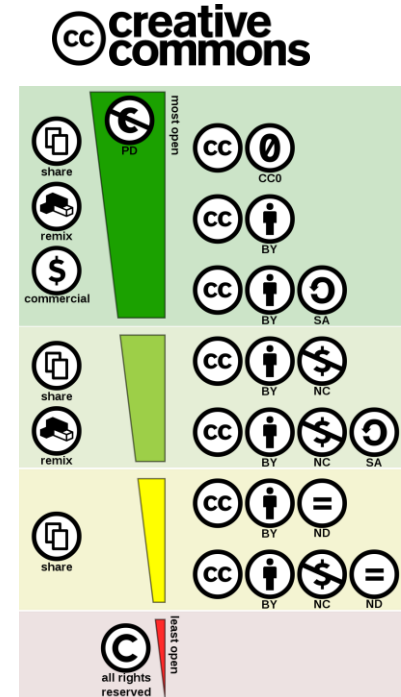
- **Encryption** (Storage, Cloud, Email)
- **Anonymisation & Pseudonymisation**
- **Informed consent:** Participants must be informed about what happens with their data
- Frequent problem when archiving sensitive data:
 - (no informed consent or)
 - formulation in consent form is too strict („data will be deleted after project end“)



- Share and publish only in trusted research environments
- **Access restriction** (physical, legal)
 - Password protection
 - Encryption
 - Access rights /Licenses



- Research data may be:
 - Automatically protected by the law
 - Regulated by contract
- Keep license declarations consistent
- Find further info at, but consider professional advice:
 - <https://www.openaire.eu/how-do-i-license-my-research-data>
 - <https://choosealicense.com/>



<https://creativecommons.org/about/ccllicenses/>



Expert advice on legal aspects can only be given by professionals, e.g. the UFZ legal department: recht@ufz.de

Legal aspects

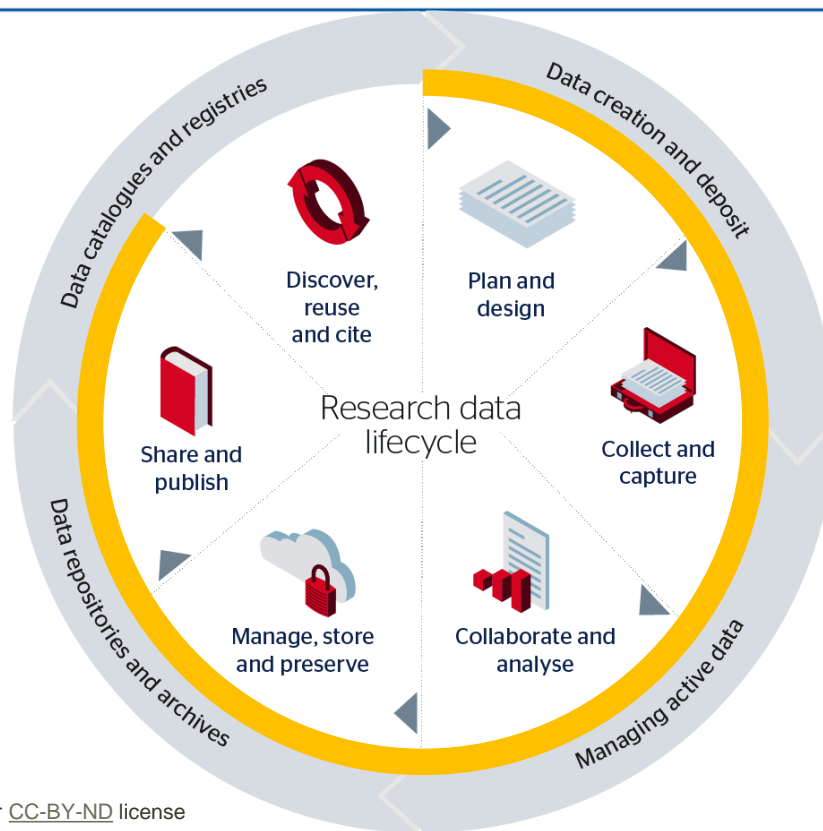
Where do I get more information?

RDM Guidelines: <https://rdm.pages.ufz.de/guidelines/>

RDM Consulting: Drop us an Email! wkdv-datamanagement@ufz.de

! Legal advice from the legal department only !





Graphic: [Jisc](#) under [CC-BY-ND](#) license

- **Various entrance points. Find research data** e.g. in:
 - Directly in subject-specific or generic repositories
 - Via meta search engines (e.g. **B2FIND** <http://b2find.eudat.eu>, **geisDataSearch** <http://datasearch.geis.org/start>, **DataOne** <https://www.dataone.org/>, **DataCite Search** <https://search.datacite.org/>)
 - Search in library search engines (e.g. **BASE** <https://www.base-search.net/Search/Advanced>)
 - Google: keyword and „data set“ or **Google Dataset Search** (<https://datasetsearch.research.google.com/>)



According to **FORCE11** recommendation (<https://doi.org/10.25490/a97f-egykh>):

Author(s), Year, Data set title, Data repository or archive, Version, Global persistent identifier (preferably as link)

According to **DataCite 2019** (<https://doi.org/10.14454/7xq3-zf69>):

Creator (PublicationYear): Title. Version. Publisher. (resourceTypeGeneral). Identifier

Tools to support **data citation**:

<https://citation.crosscite.org/> (lets try with e.g. <https://doi.org/10.1594/PANGAEA.931767>)

Time for a  15 min break!

1. Welcome and orientation
2. Data lifecycle / RDM / FAIR /open
3. Policies / data management plans

Break

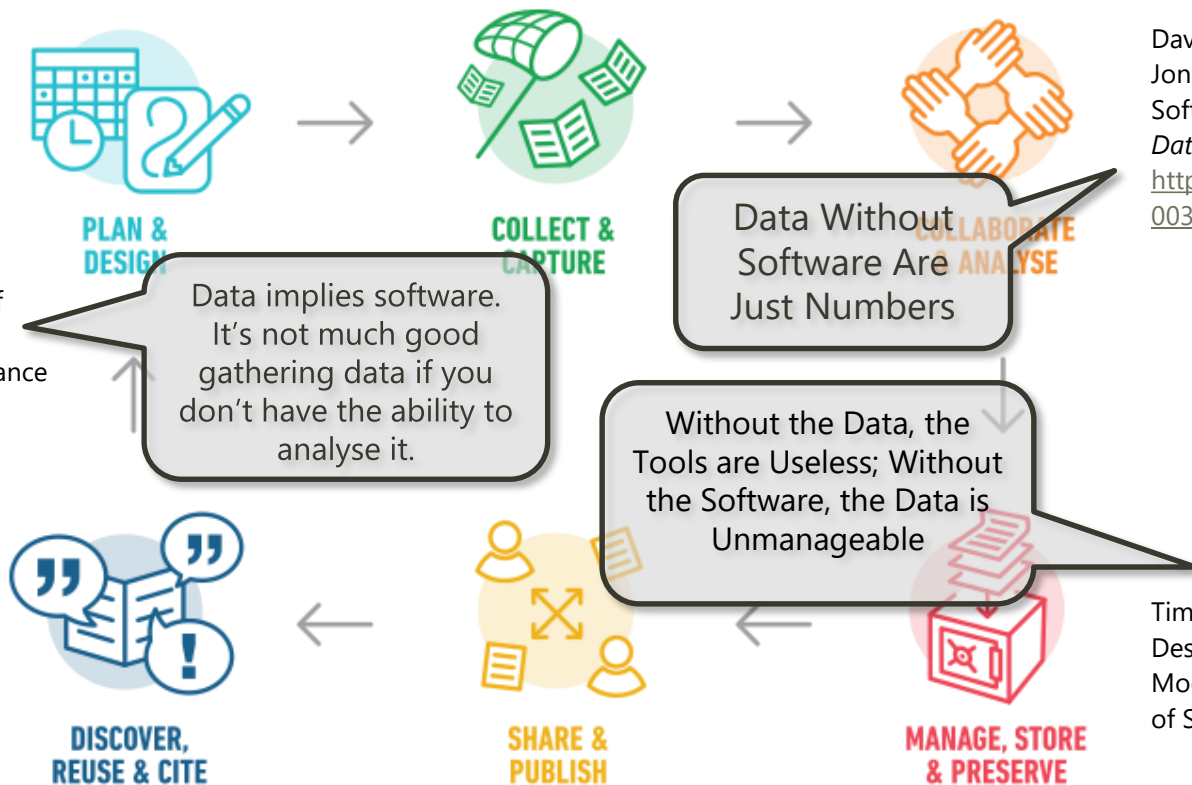
4. Metadata
5. Storage / backup / archiving
6. Data publication / reproducibility / legal aspects / discovery & reuse

Break

7. Research software
8. Institutional infrastructures, services and tools
9. Q&A

Research Software

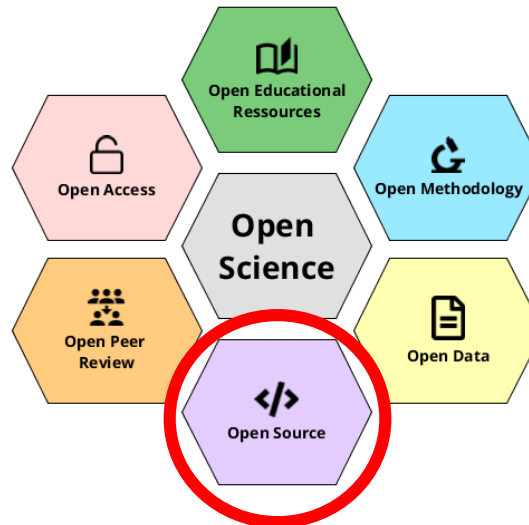
Titus Brown, Associate Professor, University of California, Davis, USA, Research Software Alliance



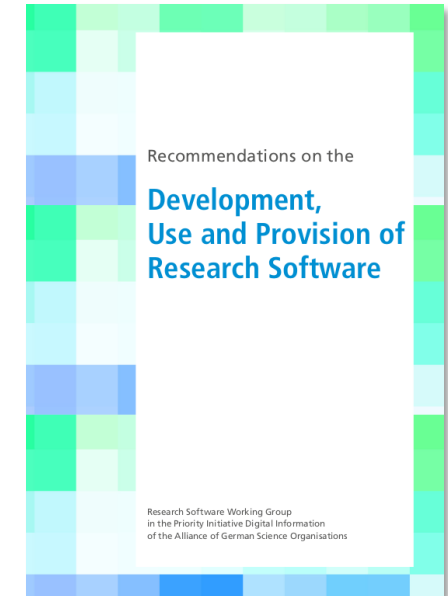
Davenport, J.H., Grant, J. and Jones, C.M., 2020. Data Without Software Are Just Numbers. *Data Science Journal*, 19(1), p.3. <http://doi.org/10.5334/dsj-2020-003>

Tim O'Reilly, 'What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software.'

- Always think data and software together
- Requirements of research funders
- Open research software is a pillar of Open Science
- Software is a scientific contribution
- FAIR for research software



DFG Deutsche
Forschungsgemeinschaft



<https://doi.org/10.5281/zenodo.1172988>

- Keep track of changes and contributions
- Release history - which software version was used in a specific research process?
- Use git (or sth. similar, e.g. SVN)








<https://git.ufz.de>
<https://codebase.helmholtz.cloud>



<https://github.com/Helmholtz-UFZ/>

Name	Last commit	Last update
docs	Chapter Infrastructure finished	2 days ago
includes	finalized chapter legal, new chapter infrastru...	5 days ago
site	Chapter Infrastructure finished	2 days ago
.gitignore	Initial commit	7 months ago
.gitlab-ci.yml	Initial commit	7 months ago
CONTRIBUTING.md	structure added	1 month ago
LICENSE	Initial commit	7 months ago
README.md	updated structure and introduction chapter	1 month ago

26 Jan, 2022 1 commit	
 updated structure and introduction chapter Ronny Gey authored 1 month ago	Verified f4dec4ad
25 Jan, 2022 2 commits	
 Update README.md Ronny Gey authored 1 month ago	5a181990
 structure added Ronny Gey authored 1 month ago	8fb618d1
17 Dec, 2021 2 commits	
 typos fixed Ronny Gey authored 2 months ago	d4c6d860
 Update README.md Ronny Gey authored 2 months ago	fec839f5

Metadata

- user and developer!
- Better use, better contributions!
- readthedocs.org

- [CodeMeta](#)
- [Citation File Format \(cff\)](#)
- [CiteAs](#)

Project Documentation

Software is Data! (see data archiving)



Software Heritage

<https://archive.softwareheritage.org/>

Content

A significant amount of source code has already been ingested in the Software Heritage archive. It notably includes the following software origins.

Regular crawling

These software origins get continuously discovered and archived using the listers implemented by Software Heritage.

 Bitbucket

2,054,062 origins



22,042 origins



19,762 origins

 debian

126,635 origins



5,869 origins

 GitHub

128,849,883 origins

 GitLab

3,476,909 origins

 Guix

11,735 origins

 GNU

354 origins

 heptapod

1,028 origins

 launchpad

20,414 origins

 NixOS

30,942 origins

 npm

1,802,899 origins



4,083 origins



447,968 origins

 SOURCEFORGE

313,566 origins

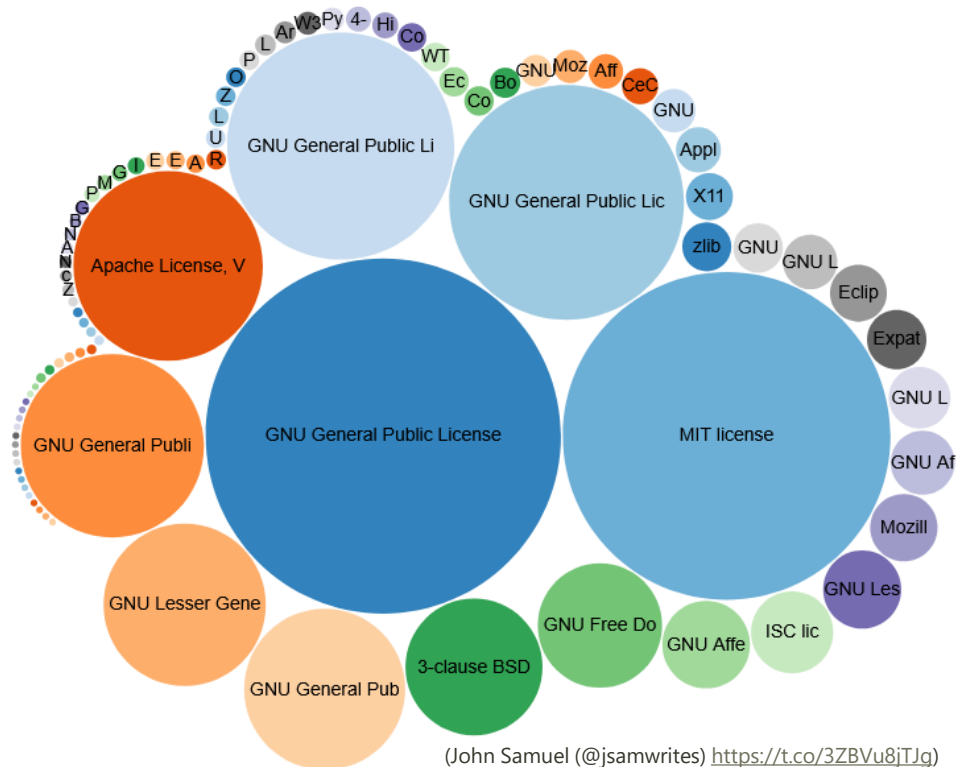
- Open Source? Open Source License!
- Open Science? Open Source!
- No License? Unclear state!
- Funders requirements? Open Source!



Public Money

Public Code

publiccode.eu



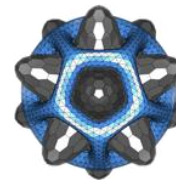
(John Samuel (@jsamwrites) <https://t.co/3ZBVu8jTJg>)



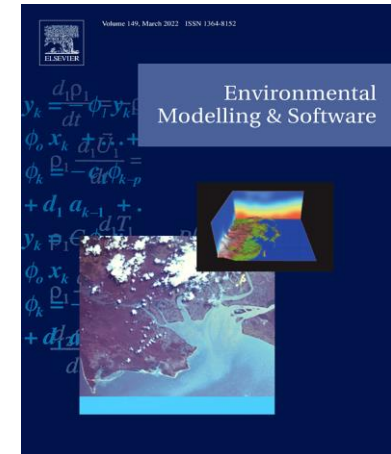
In which journal should I publish my software?

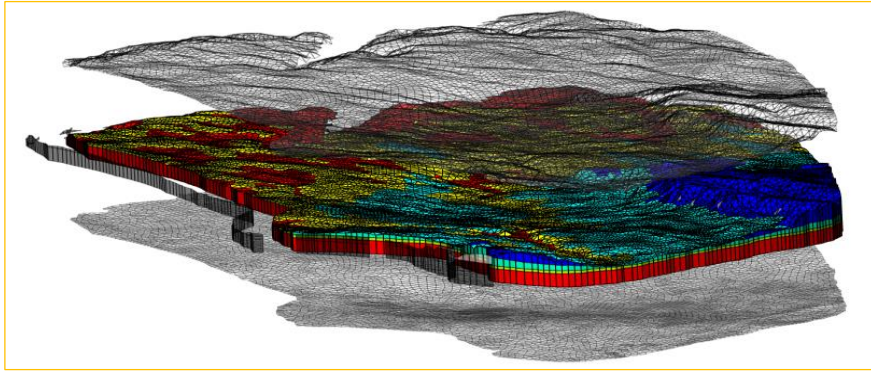
- Software Publication
- Software Meta Journal
- Journal accepting Articles about Software
- Journal Research Article + Supplements

- Publish your software and get a DOI!
- Zenodo: GitHub-Integration (in progress: GitLab-Integration)
- Alternatives: [Figshare](#), [CodeOcean](#)



The Journal of
Open Source Software



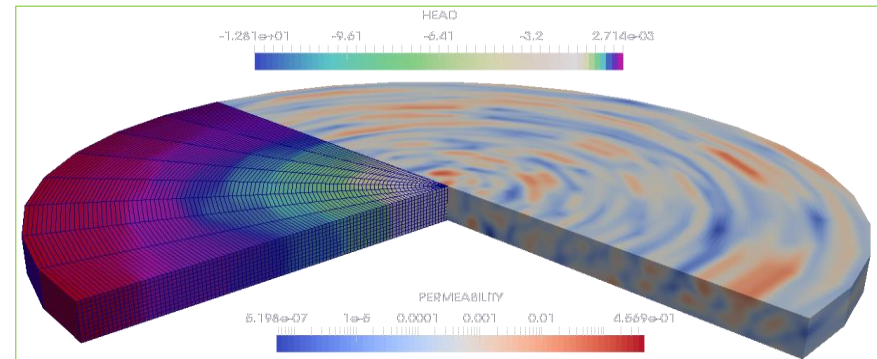


OpenGeoSys

(<https://www.opengeosys.org>)

GeoStat Framework

(<https://geostat-framework.org>)



Where to find software?



Software Catalog/Registries

Code Repositories



Search Engines: [BASE](#), [DataCite](#)

Software Citation?

Final act in the reproducibility chain!



Apache Software Foundation. Hadoop. Version 0.20.2. Feb. 19, 2010. URL:

<https://hadoop.apache.org>

K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop Distributed File System," 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, 2010, pp. 1-10, doi: 10.1109/MSST.2010.5496972.

- [Awesome Research Software Registries](#)



- Domain specific
 - [bio.tools](#) - Software tools, databases and services for bioinformatics and the life sciences.
 - [biii.eu](#) - Bioimaging
 - [geodynamics.org](#) - Variety of software packages covering crustal dynamics, tectonics, mantle convection, seismology, spherical harmonics, and computational science
 - [geospace.ontosoft.org](#) - Software from the Geospace domain, annotated and searchable
 - [sbgrid.org](#) - Structural biology softwares and datasets.
 - [scicrunch.org](#) - Biomedical resources

Research software

Where do I get more information?

RDM Guidelines: <https://rdm.pages.ufz.de/guidelines/>

RDM Consulting: Drop us an Email! wkdv-datamanagement@ufz.de

TBD: Contact point research software



Foundations of Research Software Publication

Overview We will provide you with actionable advice about how to prepare your research code before publishing it or submitting it alongside a research publication. You learn to apply the presented strategies using different practical examples

01 November 2023 - 02 November 2023

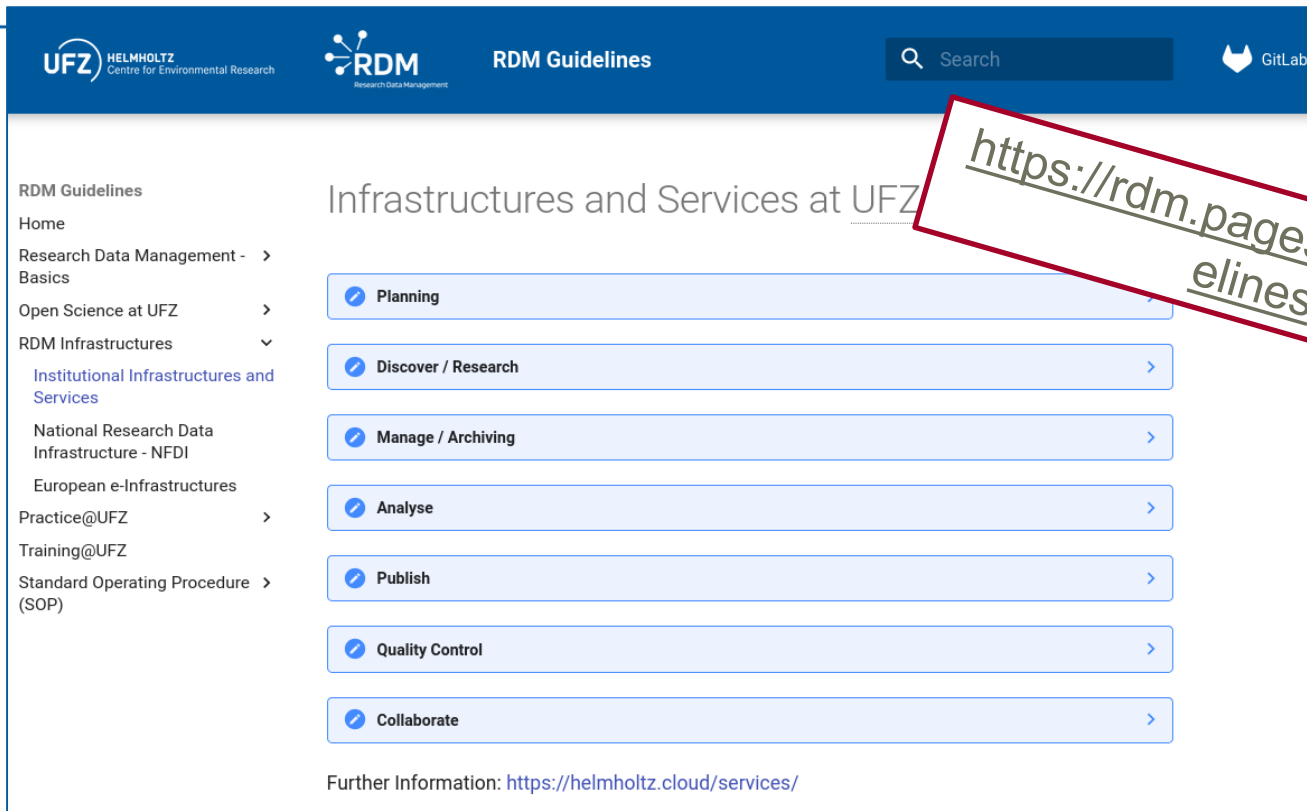
Online

HIFIS and Helmholtz Events » HIFIS - Helmholtz Federated IT Services » Education

<https://events.hifis.net/event/948/>



Infrastructures























The screenshot shows the RDM Guidelines website. The header includes the UFZ Helmholtz Centre for Environmental Research logo, the RDM Research Data Management logo, the text 'RDM Guidelines', a search bar, and a GitLab icon. The main content area is titled 'Infrastructures and Services at UFZ'. On the left is a navigation menu with items like 'Home', 'Research Data Management - Basics', 'Open Science at UFZ', 'RDM Infrastructures' (expanded to show 'Institutional Infrastructures and Services', 'National Research Data Infrastructure - NFDI', and 'European e-Infrastructures'), 'Practice@UFZ', 'Training@UFZ', and 'Standard Operating Procedure (SOP)'. The main content area lists seven categories: Planning, Discover / Research, Manage / Archiving, Analyse, Publish, Quality Control, and Collaborate, each with a right-pointing arrow. A red-bordered box is overlaid on the right side of the page, containing the URL <https://rdm.pages.ufz.de/guidelines>.

Helmholtz Cloud Services

<https://helmholtz.cloud/services>

Search Services: Provider: Sort by:

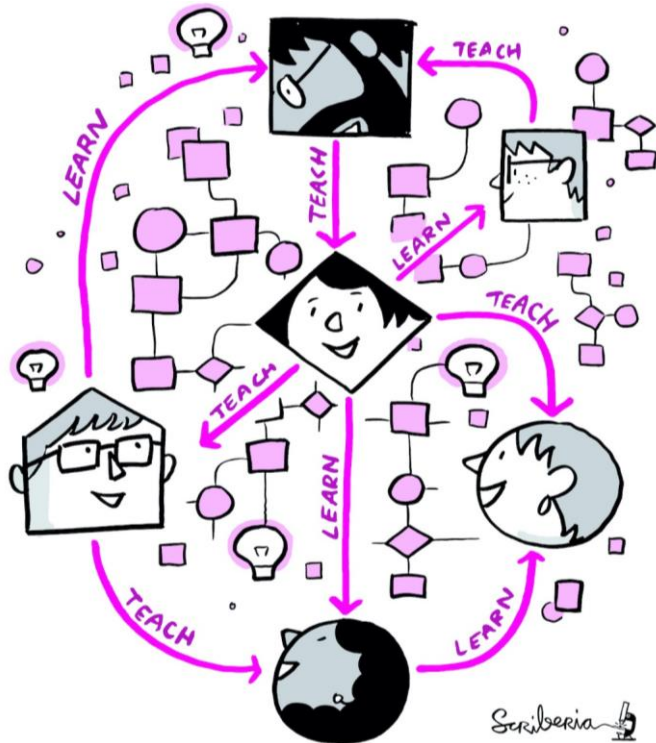
B2Share B2Share Research Data (+ corresponding Metadata) Publishing service. by:  Go to service	Compute Projects Applications for Computing Time Apply for Computing Time at Jülich Supercomputing Centre (JSC). by:  Go to service	GitLab GitLab A web-based DevOps lifecycle tool that provides a Git-repository manager. by:  Go to service	HAICORE Haicore Dedicated CComputing REsources for the Helmholtz AI community. by:  Go to service	HAICORE Haicore Dedicated CComputing REsources for the Helmholtz AI community. by:  Go to service
Notes HedgeDoc A collaborative platform to write and share markdown based documents. by:  Go to service	HIFIS Events Indico An Events Management service for everyone within Helmholtz and their partners, based on Indico. by:  Go to service	Jupyter Jupyter Open-source software and service for interactive computing. by:  Go to service	Jupyter JupyterHub Interactive supercomputing in a browser. by:  Go to service	Jupyter on HAICORE JupyterHub Jupyter enables interactive supercomputing on HPC resources. by:  Go to service
LimeSurvey LimeSurvey Community Edition An online survey tool offered by DKFZ to everyone within Helmholtz group. by:  Go to service	LimeSurvey LimeSurvey Community Edition An Open source on-line statistical survey web application. by:  Go to service	Mattermost Mattermost A hosted chat service for everyone within Helmholtz based on Mattermost. by:  Go to service	bwSync&Share Nextcloud File Sync and Share. Groupware-Functionalities: Files, Fotos, Calendar, etc. by:  Go to service	nubes Nextcloud Sync&Share based on Nextcloud with OnlyOffice and Calendar function. by:  Go to service
Sync & Share Nextcloud, dCache File Sync and Share. Collaborative Editing using OnlyOffice. by:  Go to service	OpenStack (HDF Cloud) OpenStack The Service allows provisioning of user-controlled VMs with Linux OS. by:  Go to service	Singularity SingularityCE Container runtime environment on HPC systems at FZJ / Jülich Supercomputing Centre (JSC). by:  Go to service	Data Projects (HDF) Storage Resources Apply for data projects at FZJ / Jülich Supercomputing Centre (JSC). by:  Go to service	HIFIS Helpdesk Zammad HIFIS Helpdesk Ticketing System based on Zammad. by:  Go to service

Wrap up

TAKE HOME MESSAGE(S)

- **Do not touch the raw data.** Back it up somewhere reasonable and keep a read-only copy.
- **Have a plan!** Decide where your data is going to be stored, what it is called, when/if it needs to be deleted BEFORE you start collecting it and note it down in a data management plan. If you collect sensitive data, plan for consent for sharing from the start!
- **Document everything.** You know who the worst person is at replying to emails about what the sampling frequency of channel X was? Nope not him, it's actually yourself from a year ago. Keep the documentation with the data!
- **Create the data you want to see in the world.** Imagine someone was about to give you a dataset that you needed to analyse well in order to get that job you have been dreaming about. What would you want it to look like? That is how yours should look.
- **Try not to re-invent the wheel.** Before you start creating some crazy new schema, storage format or naming protocol, have a quick google or ask your colleagues. Something that is already being used is likely to be better in the long run, even if you think there is a better solution.

How to connect and engage?



- Motivate your colleagues!
- Share your own best practices!
- Set an example!

Need more training, help, advice?

RDM guidelines: <https://rdm.pages.ufz.de/guidelines/> 

RDM training: <https://rdm.pages.ufz.de/guidelines/training/>

Helmholtz training: <https://www.helmholtz-hida.de/course-catalog/>

RDM Consulting: Drop us an Email! wkdv-datamanagement@ufz.de

HIFIS Consulting: <https://hifis.net/services/software/consulting.html>

Data Representatives: <https://www.intranet.ufz.de/index.php?en=50387>



UFZ - Helmholtz-Zentrum für Umweltforschung
Data and software publications for open science @ UFZ ONLINE EVENT
durch Frau Julia Sperfeldt, Frau Katharina Klemm, Herr Paul Meißner, Ronny Gey (Helmholtz Centre for Environmental Research)
Mittwoch 29.11.2023, 09:30 → 11:30 Europe/Berlin
Online

<https://events.hifis.net/event/1004/>

Q&A



Graphic: Scriberia under CC-BY 4.0 licence, doi: [10.5281/zenodo.3332807](https://doi.org/10.5281/zenodo.3332807) Scriberia



Ronny Gey, Desiree Jörke



wkdv-datamanagement@ufz.de



<https://www.intranet.ufz.de/rdm>



Mattermost:

<https://mm.ufz.de/ufz/channels/rdm--wkdv-news>

