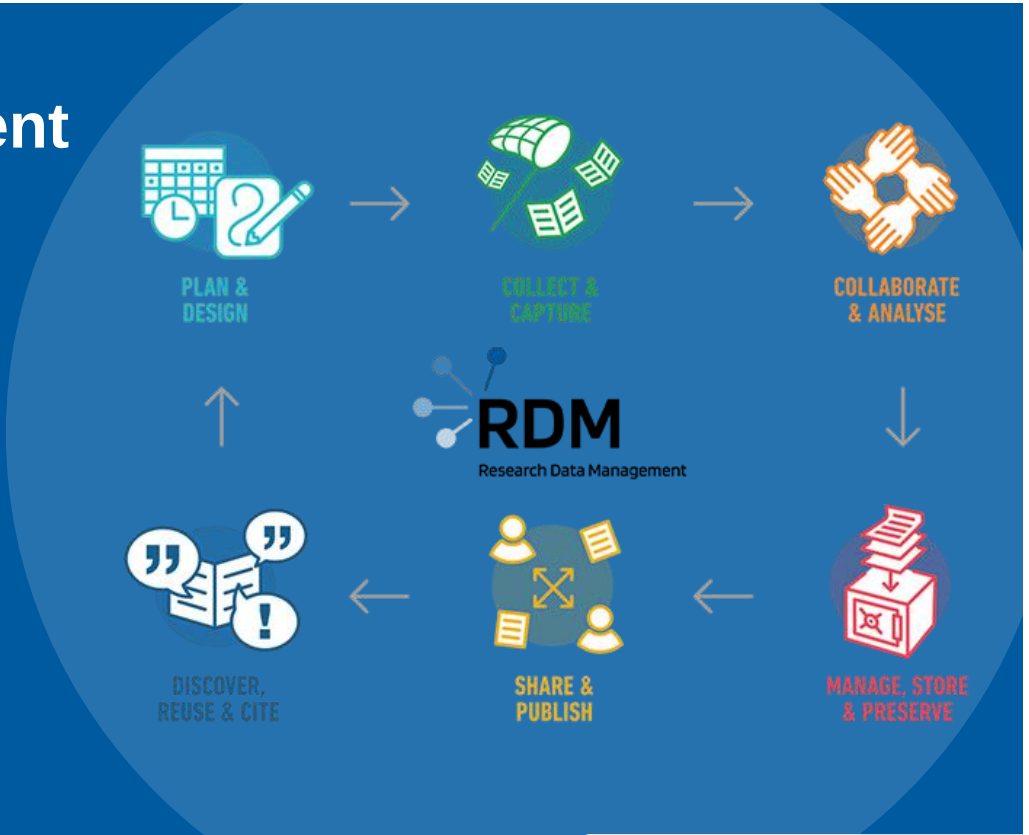
Introduction to Research Data Management

Ronny Gey (RDM)

2025-04-29



Your RDM Team at UFZ

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

Organisation

- Duration: 180 min (including breaks)
- Questions: https://notes.desy.de/pNFzOKZMR7qD1pdc_4vgPg
- Some RDM Theory, best practices, some RDM tools
- (Very) Short introduction:
  - Background?
  - Interest in RDM / workshop?
  - What was the sweetest thing you ate yesterday?

Workshop Agenda

RDM
Research Data Management

1. RDM / FAIR / open
2. Policies / data management plans
3. Metadata
4. Storage / backup / archiving
5. Data publication / reproducibility / legal aspects / discovery & reuse
6. Research software
8. Wrap-Up, Q&A

Research Data Management (RDM)

What are “research data”?

RDM
Research Data Management

“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.”

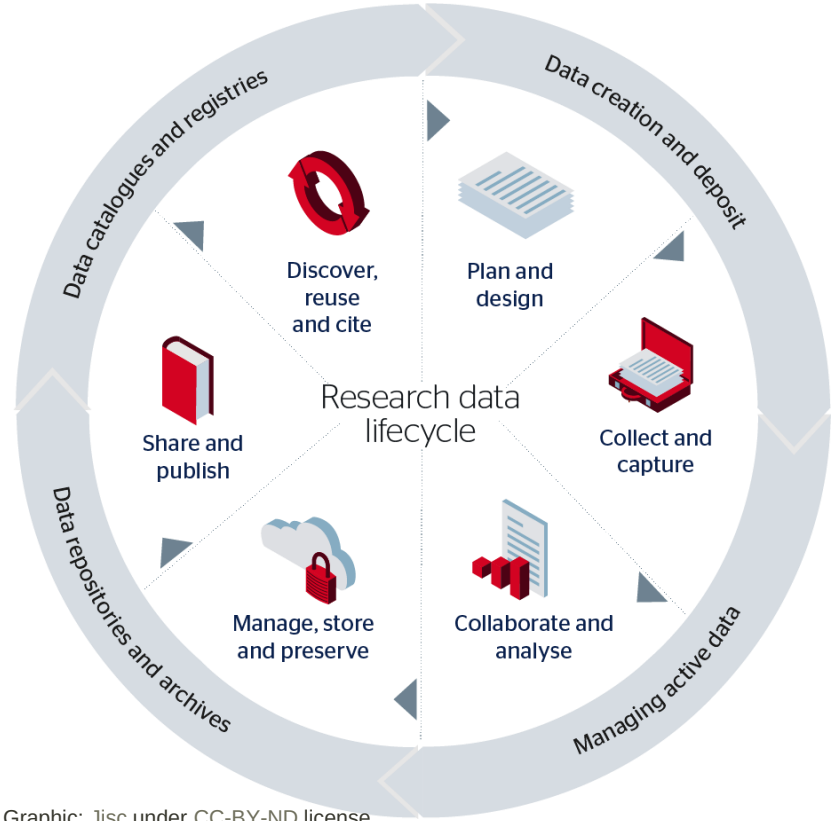
European Open Science Cloud Glossary [1]

“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

Research data lifecycle



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

Research data management

Why is following good RDM practices so important in science?

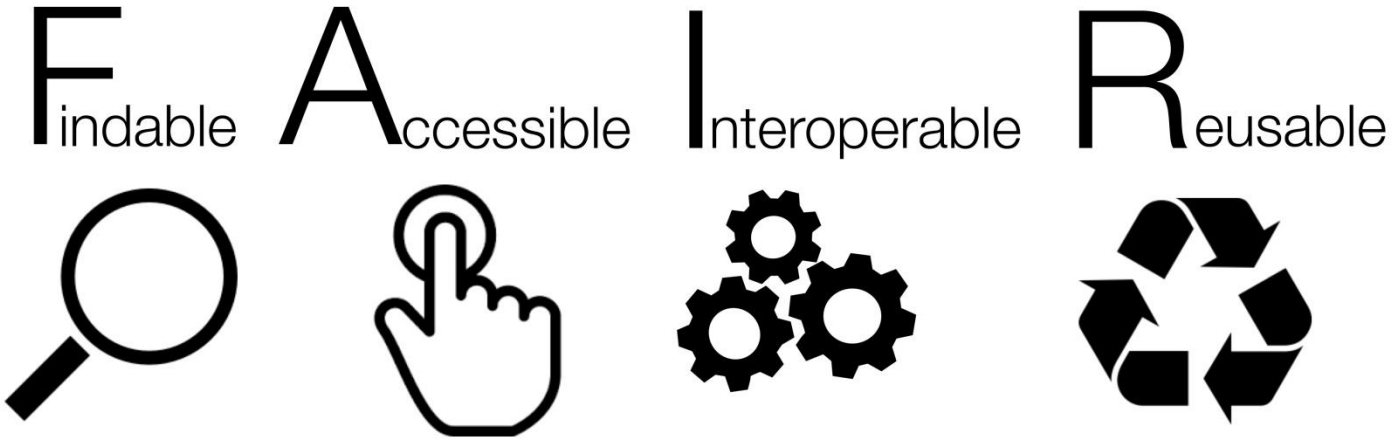
RDM
Research Data Management

- 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 licence, doi: 10.5281/zenodo.3332807

FAIR and OPEN

The FAIR principles



[1]

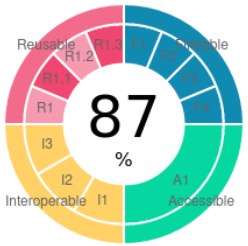
=> enhance the suitability for reuse, **by humans** and at scale **by machines**

=> 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

FAIR – Example
Principle: Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata (defined by R1 below)
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource



	Score earned:	Fair level:
Findable:	7 of 7	advanced
Accessible:	3 of 3	advanced
Interoperable:	4 of 4	advanced
Reusable:	7 of 10	moderate

Dublin Core

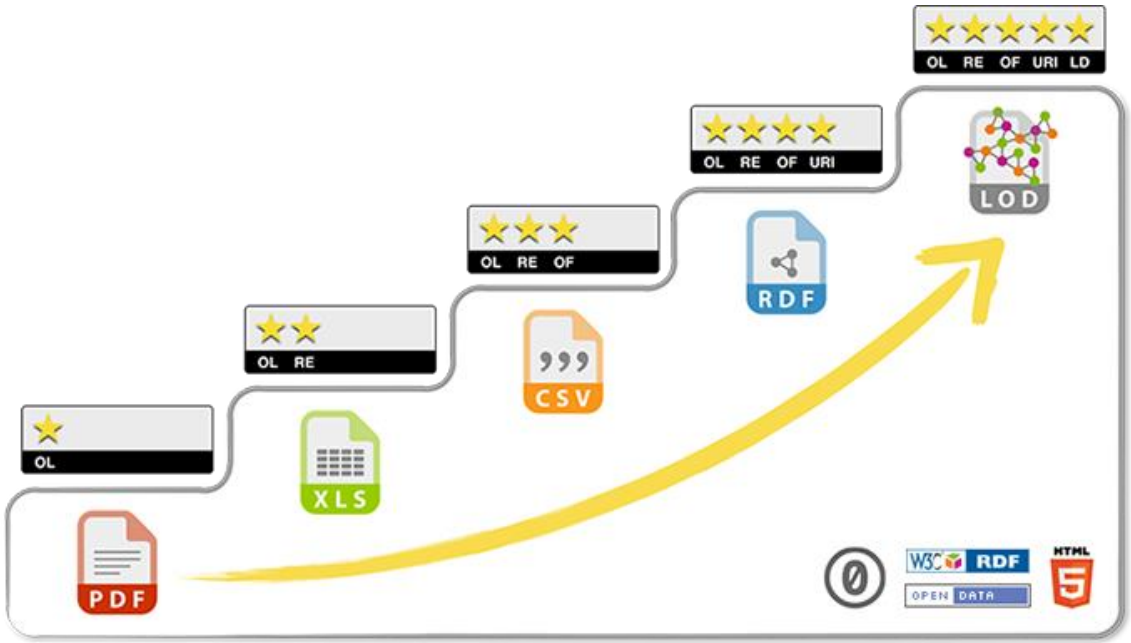
URI	https://www.ufz.de/record/dmp/archive/14903
DOI	10.48758/ufz.14903
Thema	FT-ICR-MS Radiocarbon Sequential extraction (Artificial) Lake sediment Greenhouse gas budget
Veröffenlichungsdatum	31.05.2024
Geltungsbereich	Rappbode predam, October 2021
Datenart	Datensatz (in einem definierten Format strukturierte Daten)
Datenqualität	qualitätsgesicherte Daten
Datenstatus	bearbeitete Daten abgeleitete Daten
Datenformat	csv xlsx
Urheber	Tittel, Jörg Rosenlöcher, Yvonne Dadi, Tallent Lechtenfeld, Oliver Simon, Carsten
Herausgeber	Helmholtz-Zentrum für Umweltforschung
Rechteinhaber	Helmholtz-Zentrum für Umweltforschung
Kontakt	oliver.lechtenfeld@ufz.de carsten.simon@ufz.de joerg.tittel@ufz.de
Zeitpunkt oder -bereich der Erzeugung der Daten	01.11.2021 - 15.05.2024
Sprache des Inhalts	Englisch
Version	1.0
Aktualisierungspolitik	Der Datenbestand wird nicht erweitert
Zugriffsrechte	freier Zugriff

Task: Assess the FAIRness of a data set

- Tool:
  - Fuji
  - https://www.f-uji.net/
- Dataset:
  - your own?
  - Example: https://doi.pangaea.de/10.1594/PANGAEA.946723
- Time: 5min

Open data

- 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)

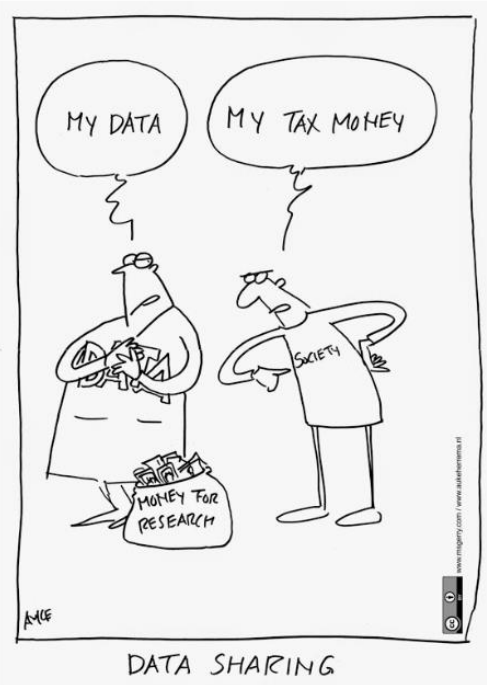
Policies

Policies and guidelines

- journal and publisher policies → **The Transparency and Openness Promotion Guidelines** [1] (The Transparency and Openness Promotion (TOP) Committee)
- institutional 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**"
- project-specific policies
- domain-specific policies → Scientific policies of CESSDA (Consortium of European **Social Science** Data Archives) [2]
  DFG-Guidelines on the Handling of Research Data in **Biodiversity Research**
  OECD Principles of Good **Laboratory Practice**
- funder policies → EC Guidelines on FAIR Data Management in Horizon 2020
  DFG Guidelines on the Handling of Research Data

[1] https://www.science.org/doi/10.1126/science.aab2374
[2] https://www.cessda.eu/About/Documents-and-Policies

Policies and guidelines



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?

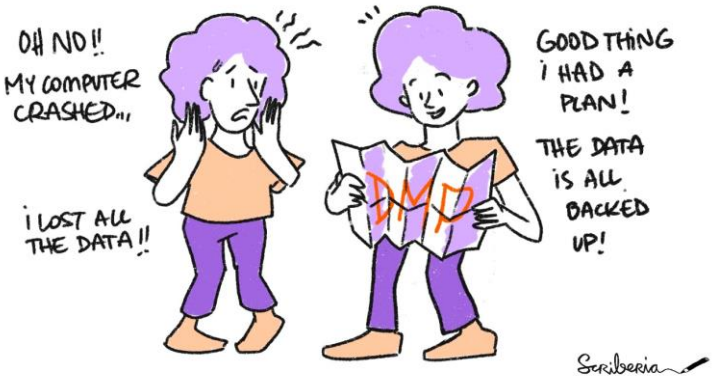
[For more data policies: https://www.forschungsdaten.org/index.php/Data_Policies (accessed 2021-02-02)]
Graphic: CC-BY: https://aukeherrema.nl

Data Management Plan (DMP)

Data management plan (DMP)

Why write a DMP?

- create a binding basis for **uniform handling** of data in the research process
- help **coordination** between project partners
- define **responsibilities** for RDM
- ease knowledge transfer in the event of **personnel changes**
- facilitate the **understanding** of one's own data
- lower the **reuse** barrier of your data
- support **cost estimation** of RDM
- DMPs are requirement of certain **funders**.



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?

RDM
Research Data Management

- Various **tools** available with slightly different features

RDMO

https://rdmo.forschungsdaten.info/

argos

https://argos.openaire.eu

DSW
DATA STEWARDSHIP WIZARD

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

gfbio

https://www.gfbio.org/plan

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

Task: RDMO Login and create a DMP

- Tool:
  - RDMO
  - https://rdmo.forschungsdaten.info / https://rdmo.nfdi4ing.de/
- Task:
  - Create an Account
  - Login
  - Create a DMP stub
  - Explore a bit
- Time: 5 min

Organisation and Versioning

Organisation and Versioning

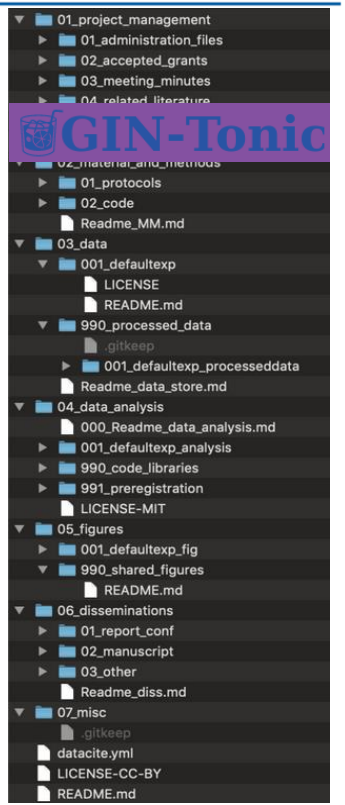
RDM
Research Data Management

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

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

R: https://github.com/jonas-hag/analysistemplates

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



Discussion: Your naming conventions and folder structure

What kind of issues have you experienced with file and folder names, folder structures and restructuring? Why did that happen? How did you solve it?

Organisation and Versioning

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

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

- use cloud storage for collaborative work (files versioned)

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

- use versioning for data, coding and software development

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

Metadata

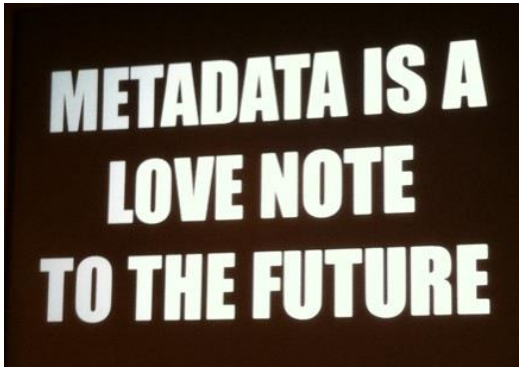
Metadata

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 schemes

- 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

Metadata schemes + terminologies + identifiers

RDM
Research Data Management

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 Management Portal

Persistent Identifiers (PIDs)

Persistent identifiers

Graphic: https://projects.tib.eu/pid-service/persistent-identifiers/persistent-identifiers-pids/, licensed: CC BY 3.0

Persistent identifiers

PIDs support:

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

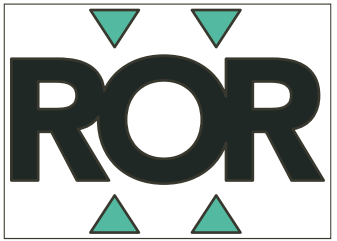
Graphics: Scriberia under a CC-BY license. DOI: 10.5281/zenodo.3332807

Persistent identifiers

RDM
Research Data Management

Different spellings for ‘UFZ’

- Helmholtz Center for Environmental Research
- Helmholtz Centre for Environmental Research
- Helmholtz Center for Environmental Research – UFZ
- Helmholtz Centre for Environmental Research – UFZ
- Helmholtz-Zentrum für Umweltforschung
- Helmholtz-Zentrum für Umweltforschung GmbH
- Helmholtz-Zentrum für Umweltforschung – UFZ
- Helmholtz-Zentrum für Umweltforschung GmbH – UFZ
- Umweltforschungszentrum



https://ror.org/000h6jb29

Task: Create an ORCID profile

- Tool:
  - ORCID – „*the only CV you will ever need*“
  - https://orcid.org/
- Task:
  - Create an Account, Login
  - Start setting up your profile
- Tool
  - https://www.intranet.ufz.de/orcid-app/
  - Connect your ORCID profile
- Time: 15 min (including break)

Storage, Backup & Archiving

Data storage

- **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://rdm.pages.ufz.de/guidelines/RDM-infrastructures/storage-options-ufz

Backup and archiving

RDM
Research Data Management

BACKUP

- backup of **all data**
- regularly **replaced** and **deleted**
- **goal: prevent data loss**
- ideally **automatically**
- In regular **intervals**

ARCHIVING

- preservation of **selected data**
- **long-term** storage
- **goal: preservation**
- **manually**
- on specific **events**
- searchable

Data Publication

Data publication

RDM
Research Data Management

Different options:

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

- **Data repository** = ~~Paper +~~ **Datasets** + Metadata+ PID

Data repositories - interdisciplinary

https://zenodo.org

https://osf.io

https://b2share.eudat.eu

https://datadryad.org

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

Data publication

RDM
Research Data Management

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:

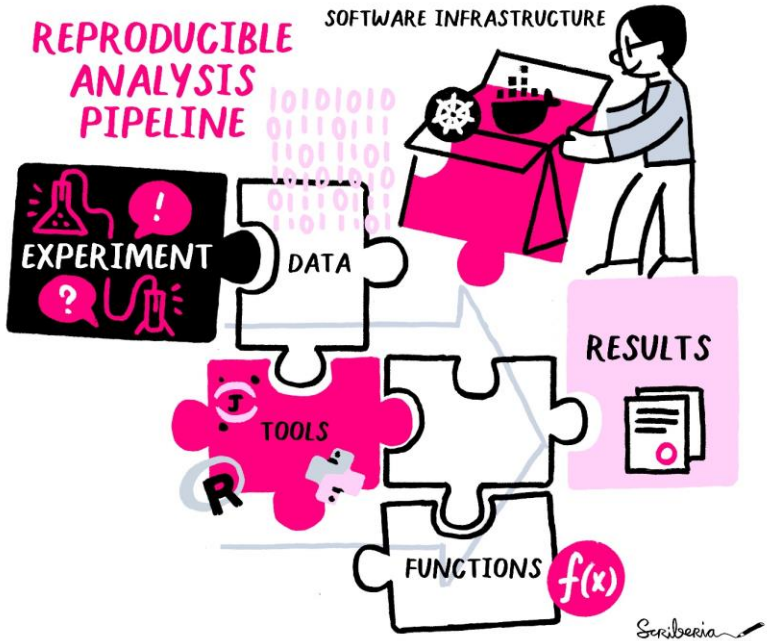
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
https://re3data.org/

Reproducibility

Reproducibility

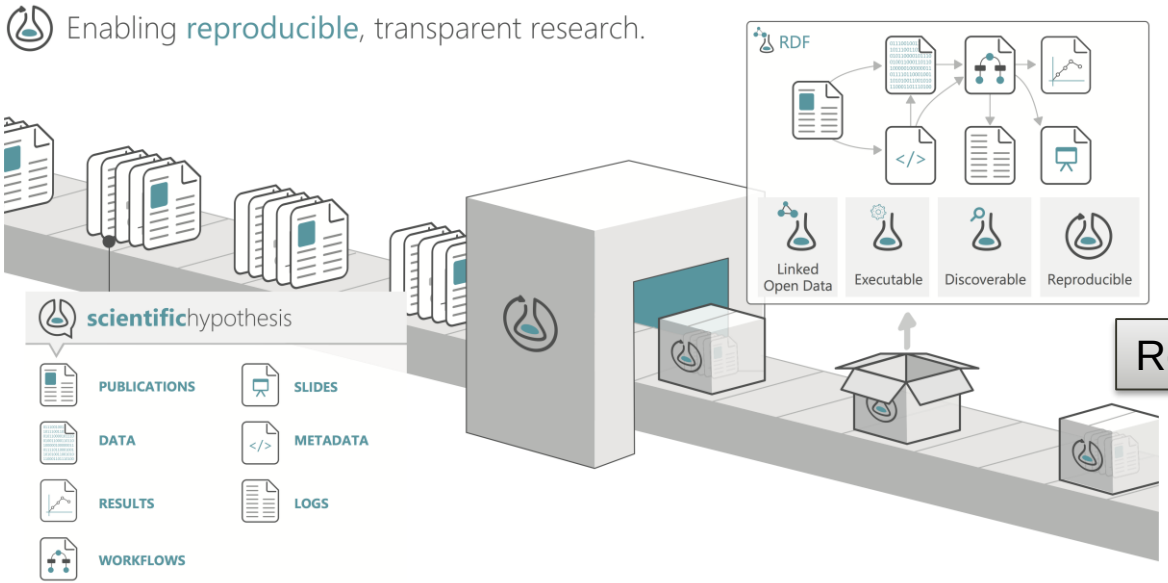
- Going beyond publishing data, publish also:
  - Scripts
  - Tools
  - Methods
  - Documentation
  - Negative results
  - …

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable



Graphic: Scriberia under CC-BY 4.0 licence, doi: 10.5281/zenodo.3332807

Reproducibility



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/

Legal Aspects

Legal aspects

RDM
Research Data Management

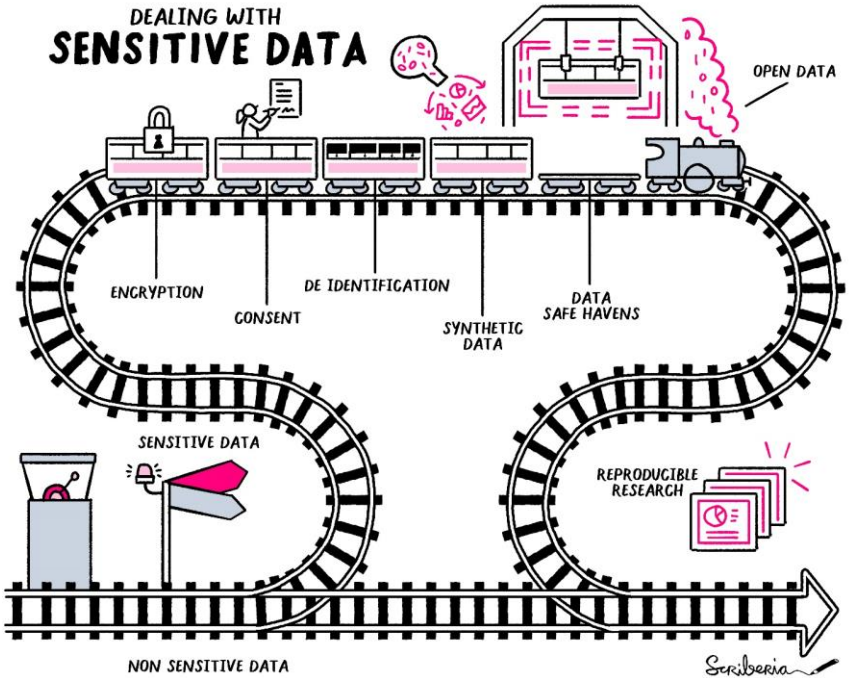
Patent law	Copyright law	Competition law	Data protection
• What has to be considered if RD (can) reach patent maturity?	• Are RD subject to copyright law at all?	• Is data used unfairly in business transactions?	• Which RD is „worthy of protection"?
Science law	**Constitutional rights**	**International law**	**EU law**
• Can licensing and publication requirements for RD be mandated?	• Which constitutional limits have to be considered?	• Which legal regulations exist outside the country?	• What consequences has e.g. the "European Data Economy" for RD?
Contracts	**Labor/service law**	**Funding requirements**	**Policies**
• Are there any agreements on the „intellectual property" of RD?	• Who „owns" the RD that is collected at UFZ?	• Which terms and conditions are set by funders (EU; industry)?	• 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.

Legal aspects - handling sensitive data

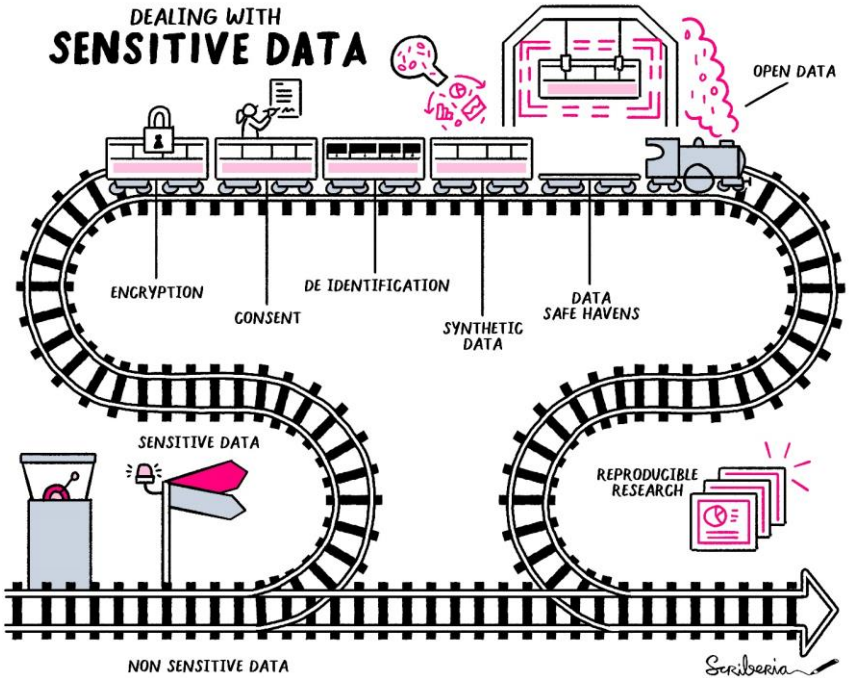
e.g. personal data, biodiversity data, confidential data

- **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“)



Legal aspects - handling sensitive data

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



Legal aspects - 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/

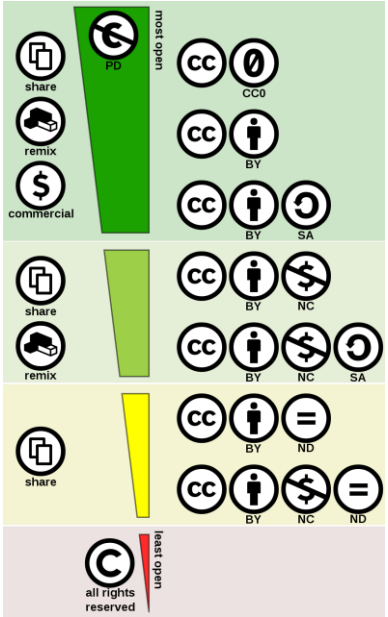
Open Data Commons

https://opendatacommons.org/licenses/

open source initiative Approved License®

https://opensource.org/licenses

creative commons



https://creativecommons.org/about/cclicenses/

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

Data Discovery and Reuse

Data discovery and reuse

RDM
Research Data Management

- **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**, gesisDataSearch** http://datasearch.gesis.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/)

Data discovery and reuse

Data Citation

RDM
Research Data Management

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

> *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):

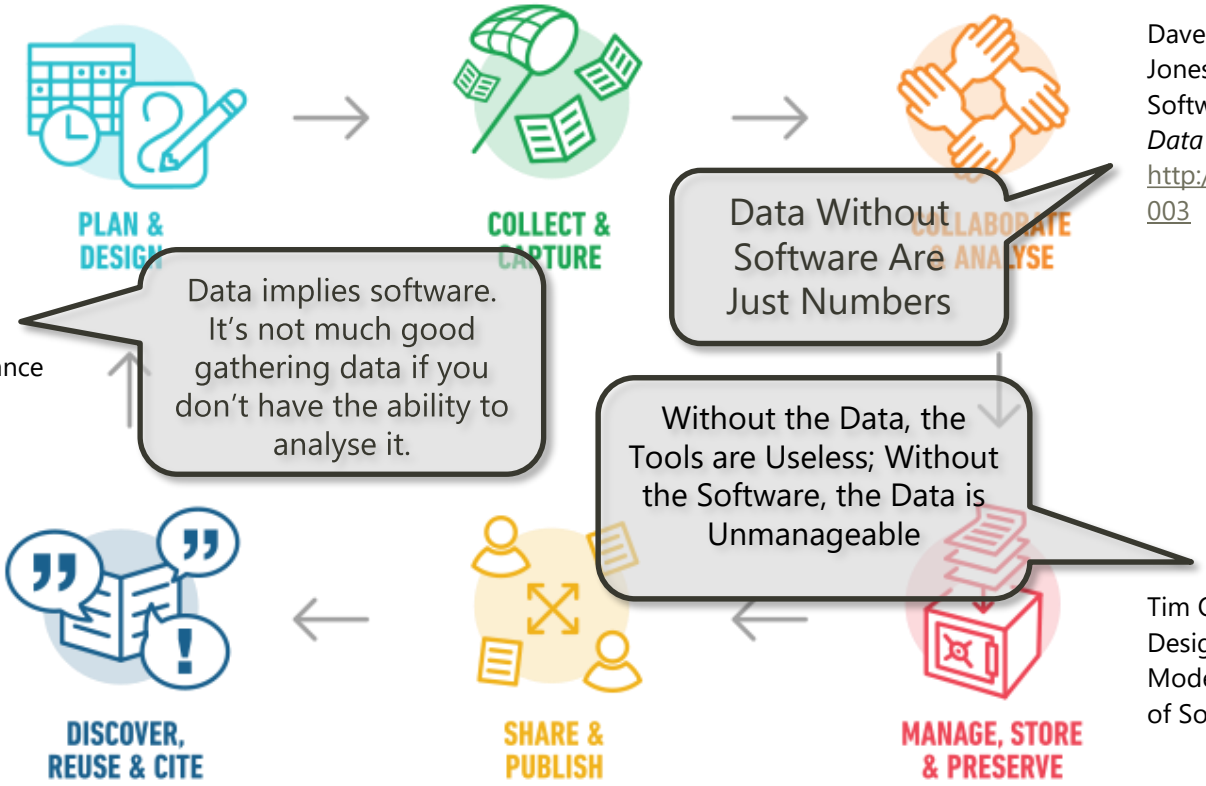
> *Creator (PublicationYear): Title. Version. Publisher. (resourceTypeGeneral). Identifier*

Tools to support **data citation**:

> https://citation.crosscite.org/

Research Software

Motivation



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

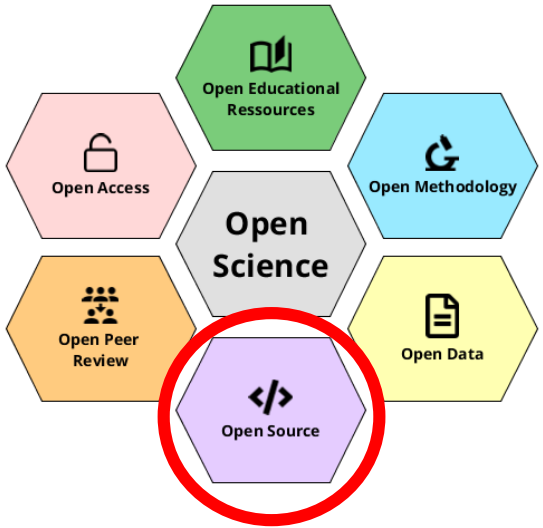
Davenport, J.H., Grant, J. and Jones, C.M., 2020. Data Without Software Are Just Numbers. *Data Science Journal*, 19(1), p.3. http://doi.org/10.5334/dsj-2020-003

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

Motivation

RDM
Research Data Management

- 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

Research Software – HIFIS/HIDA Workshops

RDM
Research Data Management

Foundations of Research Software Publication: https://events.hifis.net/event/2098

RSE Summer Schoo (last years event):
https://events.hifis.net/event/1467/

Task: Explore RDM guidelines & Workshop Pad for infrastructures

- Guidelines: https://rdm.pages.ufz.de/guidelines/RDM-infrastructures/
- Pad: https://notes.desy.de/pNFzOKZMR7qD1pdc_4vgPg
- Explore tools that might be useful for your research
- Time: 5 min

Wrap up

TAKE HOME MESSAGE(S)

RDM
Research Data Management

- **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 your self 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.

TAKE HOME MESSAGE(S)

RDM
Research Data Management

WHY RDM?

- Helps you and others to understand and reuse your data now and in the future
- Increases the reproducibility and credibility of your work
- You need to do it anyway!

HOW to do RDM?

- According to the FAIR data principles
- Perhaps with a little help from your RDM Team? Contact us via rdm-contact@ufz.de

=> Start with small steps that integrate well into your routines
=> Spending a little time upfront, can save a lot of time later on

Need more training, help, advice?

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

RDM community: https://mm.ufz.de/ufz/channels/rdm-community

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

Related:

- **Statistics Support**: https://www.intranet.ufz.de/index.php?en=45894
- **Bioinformatics Service:** https://www.intranet.ufz.de/index.php?en=45898

Q&A

Ronny Gey

wkdv-datamanagement@ufz.de

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

Mattermost:
https://mm.ufz.de/ufz/channels/rdm-community

RDM
Research Data Management