
What's on in research data management?

Roman Gerlach

29/01/2020

Research Data Management Helpdesk

- established 2015
- Central contact point for all RDM aspects at FSU Jena
- Most recent information on our website:

www.researchdata.uni-jena.de



Kontaktstelle Forschungsdatenmanagement

In vielen Bereichen der Wissenschaft sind Forschungsdaten gleichermaßen Grundlage und Ergebnis der wissenschaftlichen Arbeit. Forschungsdaten sind gemäß der [Leitlinie zum Umgang mit Forschungsdaten der DFG](#) Daten, die während eines Forschungsprozesses entstehen oder Teil der Forschungsergebnisse sind. Ein verantwortungsvoller, offener Umgang mit Forschungsdaten trägt maßgeblich zur Sicherung guter wissenschaftlicher Praxis bei, fördert die Transparenz und Reproduzierbarkeit von Ergebnissen und kommt daher nicht zuletzt dem Forschenden selbst zugute.

Helpdesk Services & Activities

Support

- On all data life cycle aspects
- Data management planning
- Data quality assurance & control
- Data publication
- Data preservation

Training

- Intro to RDM in existing lectures
- Module in M.Sc. „Computational and Data Science“
- Graduate Academy Courses: 1 - 2 day

Information

- Website (Infos and Links)
- Awareness material and events (Video, Flyer, Poster etc.)

Research & Development

- BMBF Project eeFDM (2017-19)
- Participating in strategic developments at FSU and beyond
- Contributing to the RDM community

Outline

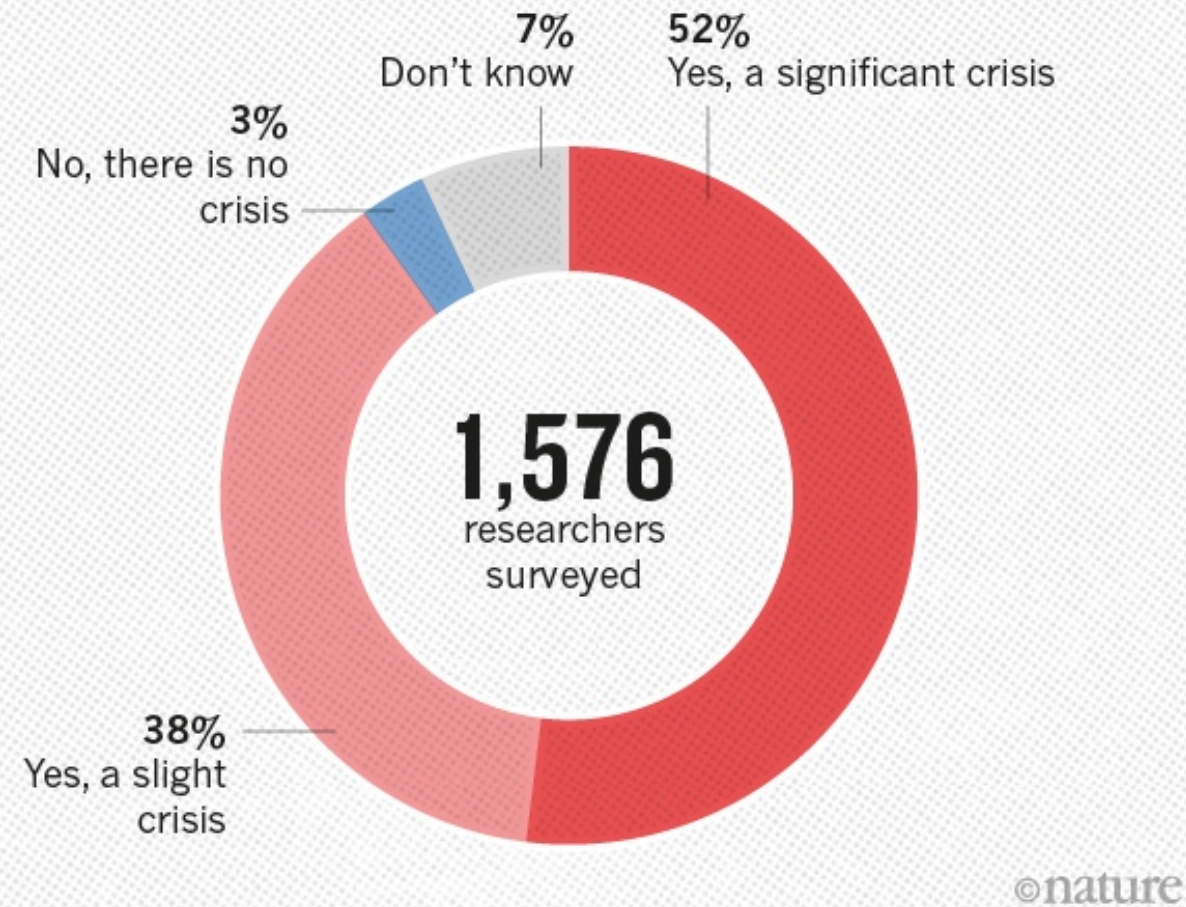
1. Introduction
2. Why Research Data Management?
3. What is Research Data Management?
4. Data/software publishing
5. Nationale Forschungsdateninfrastruktur (NFDI)



Abb. aus: Grasse, Marleen; López, Ania; Winter, Nina: Musterleitlinie für Forschungsdatenmanagement (FDM) an Hochschulen und Forschungseinrichtungen, 2018. <https://doi.org/10.5281/zenodo.1149133>. License: Other (Open)

Why *research data management*?

IS THERE A REPRODUCIBILITY CRISIS?



Baker (2016) Nature 533, 452–454 , doi:10.1038/533452a

Benefits

- Ensuring research integrity and reproducibility.
- Ensuring research data and records are accurate, complete, authentic and reliable.
- Increasing your research efficiency.
- Saving time and resources in the long run.
- Enhancing data security and minimising the risk of data loss.
- Meeting funding body grant requirements.
- Complying with practices conducted in industry and commerce.
- Preventing duplication of effort by enabling others to use your data.

(MANTRA 2013)

Data Policies

- international organisations
- funding agencies
- institutions
- publishers





- **DFG Denkschrift “Sicherung guter wissenschaftlicher Praxis”, 1998**
- Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, 2003
- OECD: “Principles and Guidelines for Access to Research Data from Public Funding”, 2007
- **Allianz: “Grundsätze zum Umgang mit Forschungsdaten”, 2010**
- G8: Science Ministers Statement , 2013
- **EC: Open Research Data Pilot Horizon 2020, 2014**
- **DFG: Leitlinien zum Umgang mit Forschungsdaten, 2015**
- Allianz: Positionspapier “Research data at your fingertips”, 2015
- OECD: Making Open Science a Reality, 2015
- **FORCE11: FAIR Principles, 2016**
- **FSU Jena: Leitlinie und Handlungsempfehlung zum Umgang mit Forschungsdaten, 2016**
- GWK: Beschluss zum Aufbau einer Nationalen Forschungsdateninfrastruktur, 2018
- **DFG: Kodex “Leitlinien zur Sicherung guter wissenschaftlicher Praxis”, 2019**

Example: DFG

- Guidelines on the Handling Research Data (2015):

2. Accessibility

Assuming that the publication of research data from a DFG-funded project does not conflict with the rights of third parties (in particular data protection or copyright), research data should be made available as soon as possible. Data should be made accessible at a stage of processing that allows it to be usefully reused by third parties (raw data or structured data). To make sure this is the case, it must be ensured that access to the data is still guaranteed when, through publication, the rights of use relating to research data are transferred to a third party, usually a publishing house.

3. Long-term archiving

In accordance with the rules of good scientific practice, research data should be archived in the researcher's own institution or an appropriate nationwide infrastructure for at least 10 years.

Example: European Commission

The screenshot shows the 'RESEARCH & INNOVATION Participant Portal H2020 Online Manual' website. The header includes the European Commission logo and navigation links. The main content area is titled 'Data management' and contains a section 'Background - Extension of the Open Research Data Pilot in Horizon 2020'. A red box highlights a warning icon and text explaining the principle of the ORD pilot. The left sidebar lists various user account and roles options, and the bottom section mentions the 'Extension Of The Open Research Data Pilot In Horizon 2020'.

RESEARCH & INNOVATION
Participant Portal H2020 Online Manual

> H2020 Online Manual > Cross-cutting Issues > Open access & Data management >

Open access Data management

Data management

Background - Extension of the Open Research Data Pilot in Horizon 2020

Please note the distinction between open access to scientific peer-reviewed **publications** and open access to **research data**:

Warning: Participating in the ORD Pilot does not necessarily mean opening up all your research data. Rather, the ORD pilot follows the principle "**as open as possible, as closed as necessary**" and focuses on encouraging sound data management as an essential part of research best practice.

Note that these guidelines do not apply to their full extent to actions funded by the ERC. For information and guidance concerning Open Access and the Open Research Data Pilot at the ERC, please see [this specific guidance](#).

Extension Of The Open Research Data Pilot In Horizon 2020

The Commission is running a flexible pilot under Horizon 2020 called the **Open Research Data Pilot** (ORD pilot). The **ORD pilot aims** to improve and maximise access to and re-use of research data generated by Horizon 2020 projects

http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

FAIR Principles

Findable
Accessible
Interoperable
Re-usable

<https://www.force11.org/group/fairgroup/fairprinciples>

Publishers Perspective



„[...] authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications.“



“PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception.“

NPG (2013). Availability of data and materials. Retrieved from <http://www.nature.com/authors/policies/availability.html>
PLOS (2014). PLOS Editorial and Publishing Policies. Retrieved from <http://www.plosone.org/static/policies.action>



What is *research data management*?

Aspects of Research Data Management

- Planning (incl. cost)
- Order and structuring
- Documentation and metadata
- Data security
- Storage and backup
- Publication
- Long-term preservation / archiving
- Legal aspects

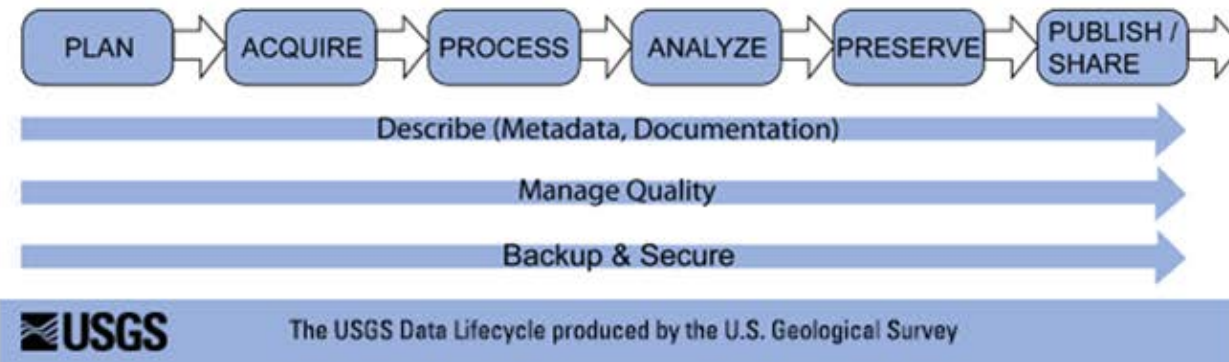
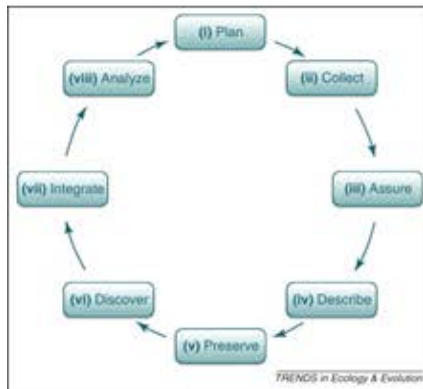
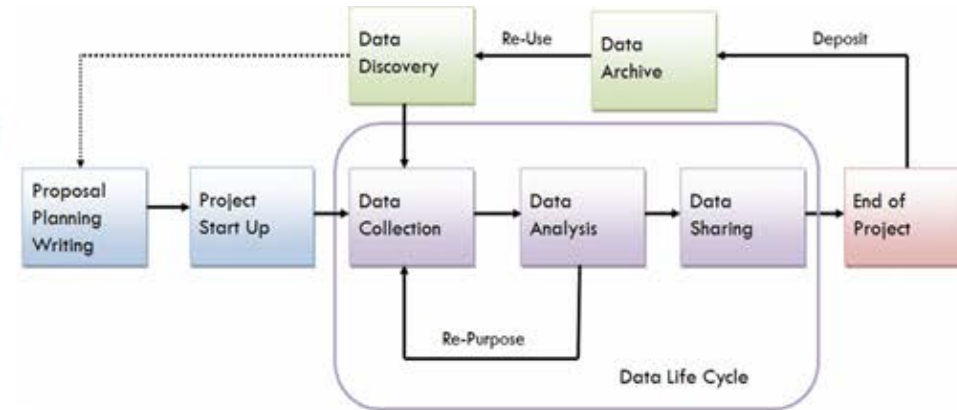
Biernacka et al. (2018)

Research Data Life Cycle



<https://www.ukdataservice.ac.uk/manage-data/lifecycle>

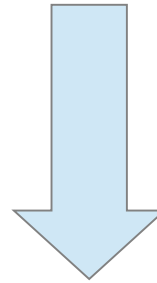
there is many data life cycles ...



Data management refers to all aspects of creating, housing, delivering, maintaining, and archiving and preserving data. It is one of the essential areas of responsible conduct of research.

(MANTRA 2017: 5)

**To ensure that no aspect of research data
management is forgotten**



**Research Data Management Plan
(DMP)**

What is a DMP?

- (formal) document that outlines how you will handle your data both during your project, and after the project is completed
- Information about all stages of the data life cycle
- It's a *living document*

Data Management Plan for Post-Graduate Research Projects	
Researcher:	
Project Title:	
Project Duration:	
Project Context:	
1. What Data will be Produced?	
2. How will the Data be Documented and Described?	
3. Has a 'File Structure/Naming Form' been completed?	
4. Deposition of E-Thesis:	
5. What are the plans for data sharing and access after submission of the thesis?	
6. What are the plans for long-term archiving of the digital data supporting the thesis?	
Signed:	Version:
Date Created:	Date Amended:

Why is a DMP needed?

- Provides a guideline during the project
- Defines responsibilities
- more and more funding agencies request it as part of a grant proposal (e.g. EU)



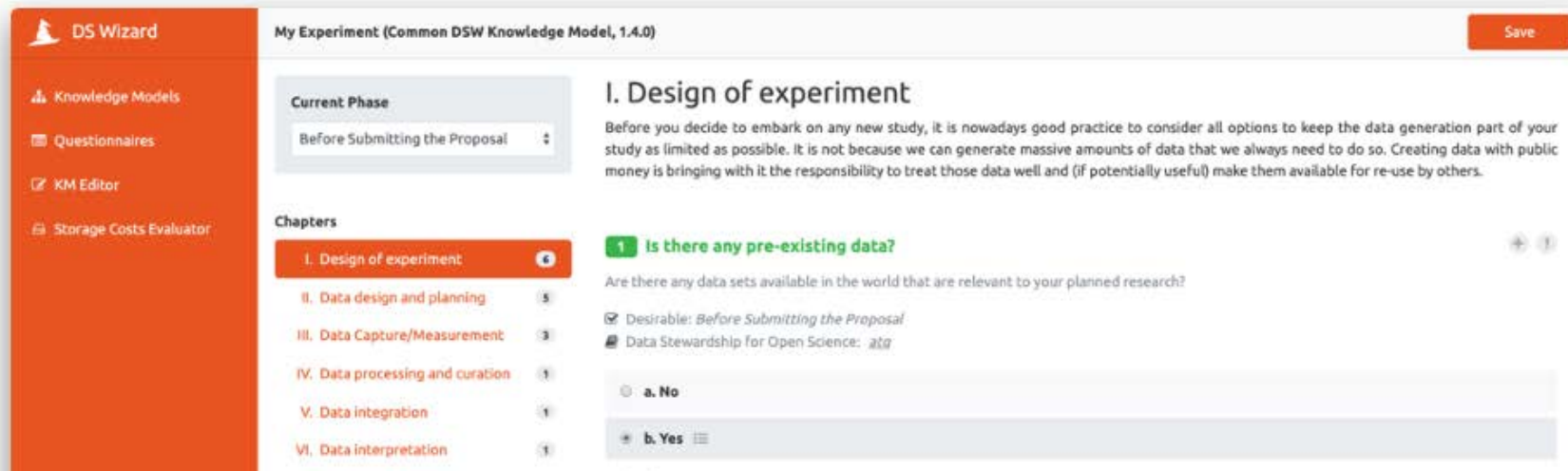
DMP for your working group/department?

Data Stewardship Wizard

Create Smart Data Management Plans
for FAIR Open Science

[Get started](#)

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



How will the data be managed **during** the project?

How will the data be managed **during** the project?

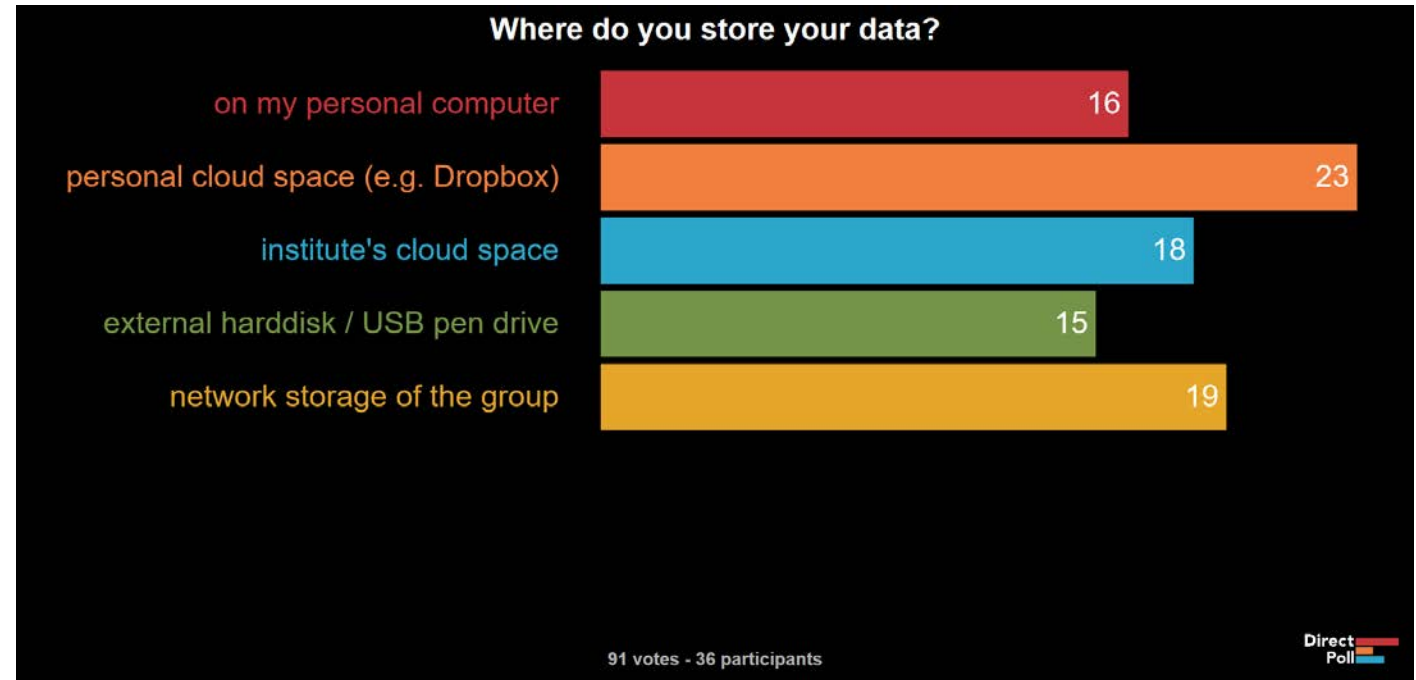
- storage capacity & back-ups (frequency, media)
- data organisation and structure (files, database)
- naming convention for files and folders?
- metadata creation
- version control (e.g. Git)
- Security & access control
- Who will be responsible?

Where do you store your data?



Live Poll:

<http://etc.ch/2CfY>



What are **pro's** and **con's** of these options?

Criteria for data storage

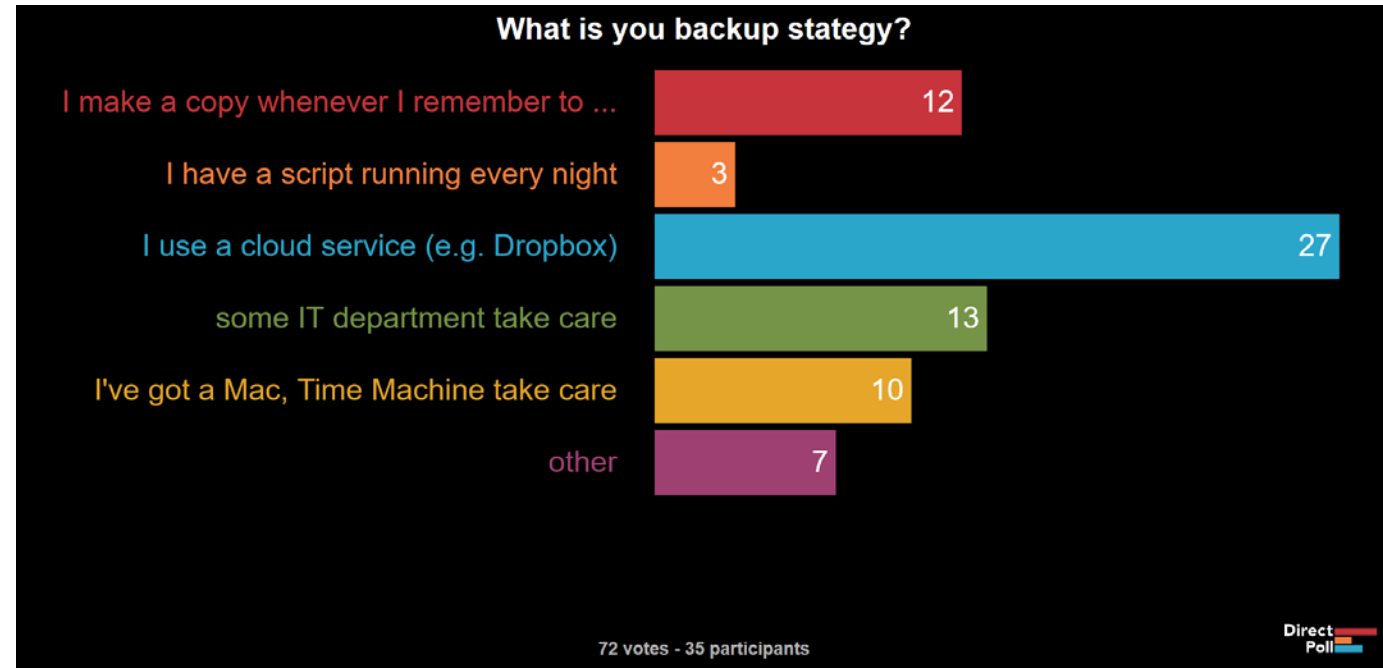
- Storage space
- network speed
- Access frequency (e.g. 24/7 online?)
- Share access (e.g. collaborators, home office?)
- Protection against manipulation and theft
- Legal constraints (e.g. data privacy)
- Backup plan

What is your backup strategy?



Live Poll (#2):

<http://etc.ch/2CfY>



What are **pro's** and **con's** of these options?

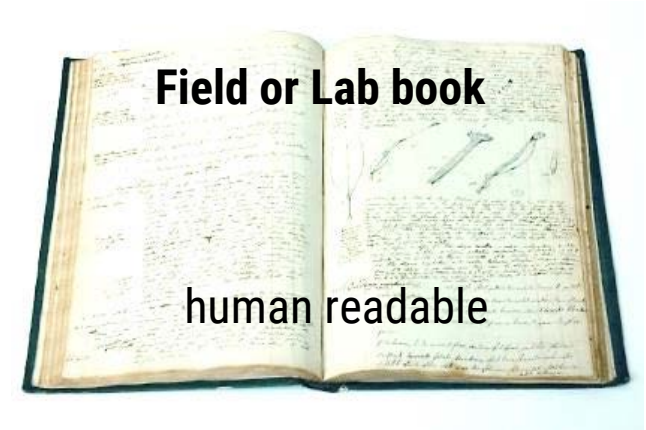
Backup

- 3-2-1 rule (3 copies, 2 types of hardware, 1 offsite)
- Prevent data loss due to hardware/software errors, theft, human error, viruses, accidents, ...
- Frequent backups (e.g., when something was changed, daily, weekly)
- Save at least last three versions
- Restricted Lifetime → every 2-5 years new hardware
- (USB stick 10-30 years, hard drive 2-10 years, CD 5 -25 years)¹

¹ <https://www.zeit.de/wissen/2013-10/s37-infografik-speichermedien.pdf>

Metadata

- Document your work (data), while you work
 - WHO created the data?
 - WHAT is the content of the data?
 - WHEN were the data created?
 - WHERE is it (from)?
 - HOW were the data collected?
 - WHY were the data collected?



```
<?xml version="1.0"?>
<dwr:DarwinRecordSet
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.guid.org/sites/arg/127%2Fdwrc_dwc_classes.xsd"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:dwc="http://rs.tdwg.org/dwc/terms/"
  xmlns:dwr="http://rs.tdwg.org/dwc/terms/"
  <dcterms:Location>
    <dwc:locationID>http://guid.dwz.org/sites/arg/127%2Fdwrc_locationID%
    <dwc:country>Argentina</dwc:country>
    <dwc:countryCode>AR</dwc:countryCode>
    <dwc:stateProvince>Neuquén</dwc:stateProvince>
    <dwc:locality>25 km al NE de Bariles, Ruta 40 (=237)</dwc:locality>
  </dcterms:Location>
  <dwc:Occurrence>
    <dcterms:type>Physical Specimen</dcterms:type>
    <dcterms:modified>2000-01-01</dcterms:modified>
    <dcterms:rightsHolder>Museum of Vertebrate Zoology</dcterms:rightsHolder>
    <dcterms:rights>Creative Commons License</dcterms:rights>
  </dwc:Occurrence>

```


Dublin Core Standard

created by
https://nsteffel.github.io/dublin_core_generator/generator_nq.html

```
<dc:title>The Northwind and the Sun/Thuringia</dc:title>
<dc:creator>Psychology Team</dc:creator>
<dc:subject>Aesop's Fable</dc:subject>
<dc:description>Please see the publication XY for more
information.</dc:description>
<dc:publisher>FSU Jena, Dep. of Psy</dc:publisher>
<dc:contributor>Schweinberger/Zäske</dc:contributor>
<dc:date>3-2-2019</dc:date>
<dc:type>sound</dc:type>
<dc:format>recordings saved as .wav</dc:format>
<dc:identifier>doi.org/123456</dc:identifier>
<dc:source>f</dc:source>
<dc:language>deu</dc:language>
<dc:relation>referenced by DOI xyz</dc:relation>
<dc:coverage>East Thuringia, 2017-19</dc:coverage>
<dc:rights>https://creativecommons.org/XYZ</dc:rights>
```

Metadata

- Use standards whenever possible
 - ISO 8601 Codes for date and time:
e.g. 2014-06-19T13:15:30Z; 2014-06-19T13:16:30+01:00
http://en.wikipedia.org/wiki/ISO_8601
 - Consistent terms/definitions/nomenclature
 - Domain-specific controlled vocabulary
- **20 year rule:** The metadata accompanying a data set should be written for a user 20 years into the future.
- But at least have a README.TXT file.



Metadata Concept Map by Amanda Tarbet is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.


Data Versioning

- **Why it is important?**
 - Reproducibility (identify and retrieve the exact same resource that the author used)
 - Keep track of changes/updates (e.g. “dynamic” data from sensors)
 - Hint: use timestamps instead of version numbers
-
- **For more details:**
 - ANDS: <http://www.ands.org.au/working-with-data/data-management/data-versioning>
 - DCC: <http://www.dcc.ac.uk/resources/how-guides/cite-datasets#sec:versions>

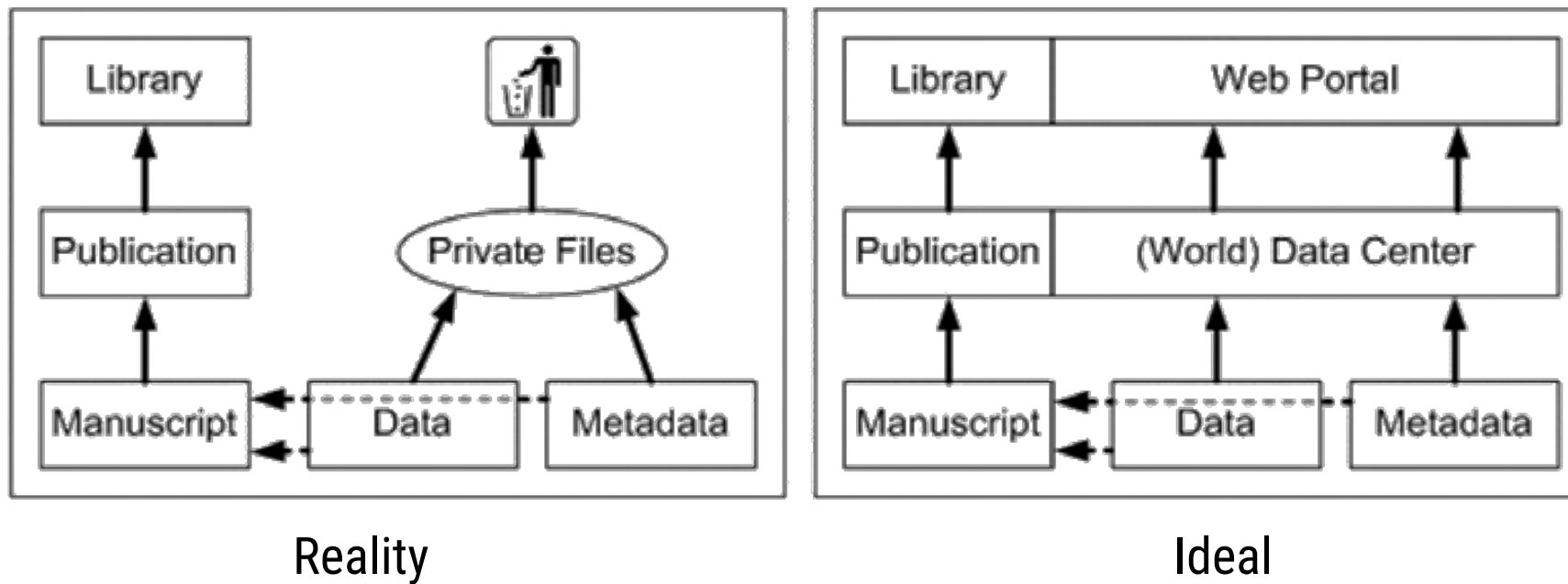
How will the data be managed **after** the project?

How will the data be managed **after** the project?

- Data Publication vs. data archiving?
- Where can I publish my data?
Isn't my website enough?
- Maintenance of existing technical infrastructure?
- Migrating data into a (public) data repository?
- Embargo periods?

 Funding organisation encourage/request data publication,
not only data archiving

Reality Check?!



(Rümpel 2011:26)

Data Publication

How can I publish my data?

- journal + supplement
- data journal
- data repository

Journal

Journal = Paper + PID + [supplement]

The screenshot displays the Science journal website. The top navigation bar includes the Science logo, AAAS.ORG, FEEDBACK, HELP, LIBRARIANS, and a search bar. Below this is a red navigation bar with links to NEWS, SCIENCE JOURNALS, CAREERS, BLOGS & COMMUNITIES, MULTIMEDIA, and COLLECTIONS. The main content area shows the article title "Essential Biodiversity Variables" by H. M. Pereira et al. The article's DOI, 10.1126/science.1229931, is circled in blue. Below the article title, the section "Materials/Methods, Supplementary Text, Tables, Figures, and/or References" is also circled in blue, containing links to "Download Supplement", "Supplementary Text", "Fig. S1", and "Full References".

DOI

Supplement

Data Journal

Data Journal = Paper + Datasets + PID



An 18-yr long (1993–2011) snow and meteorological dataset from a mid-altitude mountain site (Col de Porte, France, 1325 m alt.) for driving and evaluating snowpack models
S. Morin, Y. Lejeune, B. Lesaffre, J.-M. Panel, D. Poncet, P. David, and M. Sudul
Page(s) 13-21

▢ [Abstract](#) ▢ [Final Revised Paper](#) (PDF, 2559 KB) ▢ [Supplement](#) (14997 KB) ▢ [Discussion Paper](#) (ESSDD)



List of Data Journals: <https://www.researchdata.uni-jena.de/Information/Datenpublikation.html>

Data Repositories

Data repository = Datasets + Metadata + PID



Search Engine for Repositories



Home Search Browse Suggest FAQ About Schema API Contact Imprint

Search for Repositories (1223 Reviewed Repositories)

Subject: Content Type: Country (of the responsible institutions):

☒ Biology

☐ Certificates ☒ Open Access ☐ Persistent Identifier

421 results (filtered) (1 – 25)

« 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 »

Sort by weight

3TU.Datacentrum

3TU.DC



Subjects: **Biology** Agriculture, Forestry, Horticulture and Veterinary Medicine Agriculture, Forestry, Horticulture and Veterinary Medicine
Analytical Chemistry, Method Development (Chemistry) Basic Biological and Medical Research Bioinformatics and Theoretical Biology

ZENODO & GitHub

- Push your GitHub repo to Zenodo and get a DOI = make it citable and reusable

The screenshot shows the Zenodo interface for a repository named 'cldf/cldf: CLDF 1.0.1'. The header is blue with the Zenodo logo, a search bar, and links for 'Upload' and 'Communities'. On the right, there are 'Log in' and 'Sign up' buttons. Below the header, the repository name is displayed, along with the date 'May 24, 2018' and a 'Software' tag. The authors listed are Robert Forkel, Johann-Mattis List, Michael Cysouw, Christoph Rzymiski, and Simon J Greenhill. A note mentions 'Incorporated feedback from reviews.' The main content area is divided into two sections: 'Preview' and 'Files'. The 'Preview' section shows a directory tree with files like 'sources.bib', 'test_data.py', 'faq.md', 'modules' (containing 'Dictionary', 'Generic', 'ParallelText', 'StructureDataset', 'Wordlist'), and 'scripts' (containing 'make_defaults.py'). The 'Files' section shows a table with one file: 'cldf/cldf-v1.0.1.zip' (429.3 kB). To the right of the main content, there are statistics: 356 views, 48 downloads, and a 'Blogged by 1' badge. Below these are logos for 'Available in GitHub' and 'Indexed in OpenAIRE'. At the bottom right, there is a 'Publication date' section showing 'May 24, 2018', a 'DOI' of '10.5281/zenodo.1252097', 'Keyword(s)' including 'historical linguistics', 'computer-assisted language comparison', and 'standards', and 'Grants' including 'European Commission' and 'CALC - Computer-Assisted Language Comparison: Reconciling Computational'.

zenodo Search Upload Communities Log in Sign up

May 24, 2018 Software Open Access

cldf/cldf: CLDF 1.0.1

Robert Forkel; Johann-Mattis List; Michael Cysouw; Christoph Rzymiski; Simon J Greenhill

Incorporated feedback from reviews.

Preview

- sources.bib 1.4 kB
- test_data.py 4.1 kB
- faq.md 3.0 kB
- modules
 - Dictionary
 - Dictionary-metadata.json 2.8 kB
 - README.md 1.6 kB
 - Generic
 - Generic-metadata.json 254 Bytes
 - ParallelText
 - ParallelText-metadata.json 2.6 kB
 - README.md 3.5 kB
 - StructureDataset
 - README.md 1.2 kB
 - StructureDataset-metadata.json 2.7 kB
 - Wordlist
 - README.md 8.7 kB
 - Wordlist-metadata.json 2.6 kB
- scripts
 - make_defaults.py 6.0 kB

Files (429.3 kB)

Name	Size	
cldf/cldf-v1.0.1.zip	429.3 kB	Preview Download
md5:07be143d21208d73edb2751cf4f149c6		

Citations 1

Show only: ☐ Literature (1) ☐ Dataset (0) ☐ Software (0) ☐ Unknown (0) Search

356 views 48 downloads See more details...

Blogged by 1

Available in GitHub

Indexed in OpenAIRE

Publication date: May 24, 2018

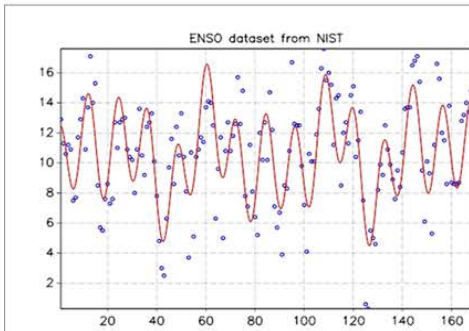
DOI: 10.5281/zenodo.1252097

Keyword(s): historical linguistics computer-assisted language comparison standards computational historical linguistics

Grants: European Commission
• CALC - Computer-Assisted Language Comparison: Reconciling Computational

Identification: Digital Object Identifier (DOI) - Example

1. Take a dataset



2. Describe it

Title
Authors
Year
Description
And others...

3. Assign a DOI



4. Reuse and reference!

ATLAS Collaboration, "Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC: $H \rightarrow \gamma\gamma$,"
<http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44>



Unique



Persistent

5. Enjoy the benefits



Slide by



<https://support.datacite.org/docs/doi-basics>

Other Persistent Identifiers (PIDs)



Researcher identifier: **ORCID** (Open Researcher & Contributor ID)

- Central registry of unique identifiers for individual researchers to address author name ambiguity
- Transparent linking mechanism between ORCID and other author ID schemes
- <https://orcid.org/>

More on PIDs

the PID Forum

What are Persistent Identifiers?

Knowledge Hub Getting Started with PIDs pid-basics



maddenfc FREYA

3 Mar '19

A persistent identifier (PID) is a long-lasting reference to a resource. That resource might be a publication, dataset or person. Equally it could be a scientific sample, funding body, set of geographical coordinates, unpublished report or piece of software.

Whatever it is, the primary purpose of the PID is to provide the information required to reliably identify, verify and locate it. A PID may be connected to a set of metadata describing an item rather than to the item itself.

There are different PID types for different kinds of resources. In the current research environment we most commonly see two varieties: those for objects (publications, data, software) and those for people (researchers, authors, contributors).

Useful Links

- What are Persistent Identifiers and why are they Important? by Jonathan Clark 21
- Persistent identifiers: current features and future properties by Juha Hakala 16
- ANDS Guide to persistent identifiers: awareness level 12
- ANDS Guide to persistent identifiers: expert level 2
- ANDS Webinar series on persistent identifiers 3

<https://www.pidforum.org/t/what-are-persistent-identifiers/295>

plos.org create account sign in

PLOS BIOLOGY
FIFTEENTH ANNIVERSARY

BROWSE PUBLISH ABOUT

SEARCH advanced search

OPEN ACCESS

PERSPECTIVE

Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data

Julie A. McMurry, Nick Juty, Niklas Blomberg, Tony Burdett, Tom Conlin, Nathalie Conte, Mélanie Courtot, John Deck, Michel Dumontier, Donal K. Fellows, Alejandra Gonzalez-Beltran, Philipp Gormanns, Jeffrey Grethe, [...], Helen Parkinson [view all]

Published: June 29, 2017 • <https://doi.org/10.1371/journal.pbio.2001414> • >> See the preprint

Article	Authors	Metrics	Comments	Media Coverage
Abstract				
Introduction				
Conclusion				
Supporting information				
Acknowledgments				
References				

Abstract

In many disciplines, data are highly decentralized across thousands of online databases (repositories, registries, and knowledgebases). Wringing value from such databases depends on the discipline of data science and on the humble bricks and mortar that make integration possible; identifiers are a core component of this integration infrastructure. Drawing on our experience and on work by other groups, we outline 10 lessons we have learned about the identifier qualities and best practices that facilitate large-scale data integration. Specifically, we propose actions that identifier practitioners (database providers) should take in the design, provision and reuse of identifiers. We also outline the important considerations for these

115 Save 31 Citation

24,396 View 99 Share

Download PDF Print

Check for updates

Included in the Following Collections

Open Data
Meta-Research: Policy and Interventions

<https://doi.org/10.1371/journal.pbio.2001414>

Data Citation of Evolving Data



Recommendations of the Working Group on Data Citation (WGDC)

Andreas Rauber, Ari Asmi, Dieter van Uytvanck and Stefan Pröll

Revision of October 20th 2015

I. MAKING DATA CITABLE

These WGDC recommendations enable researchers and data centers to identify and cite data used in experiments and studies. Instead of providing static data exports or textual descriptions of data subsets, we support a dynamic, query centric view of data sets. The proposed solution enables precise identification of the very subset and version of data used, supporting reproducibility of processes, sharing and reuse of data.

Goals of this WG are to create identification mechanisms that:

- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner

A. *Preparing the Data and the Query Store*

Prepare existing data sources and provide the required infrastructure, which is needed for implementing the query based approach.

- **R1 – Data Versioning:** Apply versioning to ensure earlier states of data sets can be retrieved.
- **R2 – Timestamping:** Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.
- **R3 – Query Store Facilities:** Provide means for storing queries and the associated metadata in order to re-execute them in the future.

B. *Persistently Identify Specific Data Sets*

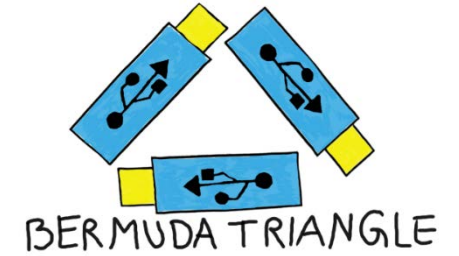
When a data set should be persisted, the following steps need to be applied:

- **R4 – Query Uniqueness:** Re-write the query to a

<http://dx.doi.org/10.15497/RDA00016>

Data Preservation/Archiving

Data Archiving



- After project completion, data which is no longer required should be stored in an archivable file format.
- at least in an archive folder at a different location, but better at an archive server (audit-proof and irreversible storage of data)
- Good scientific practice: **ALL** data shall be securely stored for **10 years** in a durable form
- Data selections for long-term archiving (for eternity)
 - basis of a published research paper
 - irretrievable measurements & observations
 - expensive to sample/collect
 - milestones

Rex (2019)

Aspects of Archiving

Maintenance of data
content

- bitstream preservation

Maintenance of
functionality

- migration
- emulation

Maintenance of
usability

- documentation
- metadata

Rex (2019)

File Formats I

Suitable for long-term archiving:

- unencrypted, uncompressed, unpatented/non-proprietary, standardized, ideally open
- If data are stored in a proprietary format (format of original software), they should be always additionally stored in an open/archivable format.
- (long-term) archiving takes time (documentation, meta data) and money (infrastructure). Plan at an early stage!

Rex (2019)

File Formats II

File Type	Recommendation	Do not use
Tables	CSV, TSV, SPSS portable	Excel
Text	TXT, ODT, HTML, RTF, PDF/A only if layout is important	Word, Powerpoint
Multimedia	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	QuickTime, H264
Pictures	TIFF, JPEG2000, PNG	GIF, JPG
Structured Data	XML, RDF, JSON	RDBMS

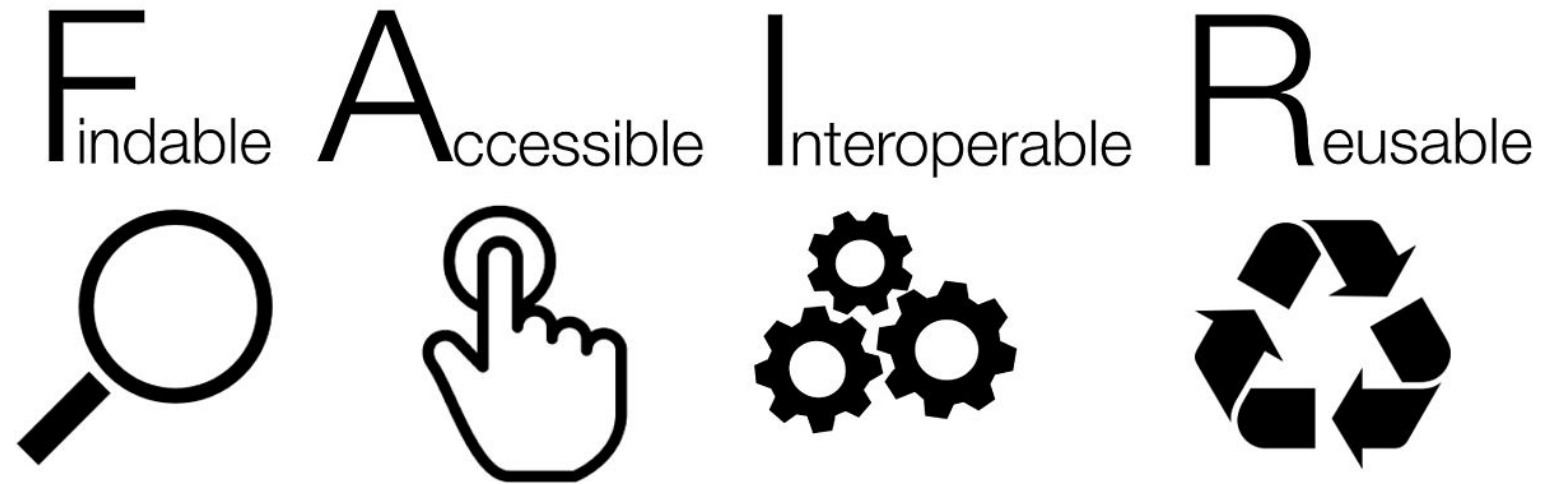
Helbig (2017)

More information:

<https://www.forschungsdaten.info/themen/bewahren-und-nachnutzen/formate-erhalten/>

FAIR Principles & Implementation

FAIR Data



- **Not only for publication of data, but also internally!**
(e.g. for yourself, colleagues, collaborators)

Box 2 | The FAIR Guiding Principles

To be 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 it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

Slide by Eric Schultes, GO FAIR DCSS IN Workshop Leiden 18/12/2019; https://drive.google.com/file/d/1G_bcUhsAyVG3SFZMoBnoJ59M-2zlqR9K/view



Technical infrastructure (generic operations)
Data / social agreements (domain-specific content)

Box 2 | The FAIR Guiding Principles

To be 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 it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

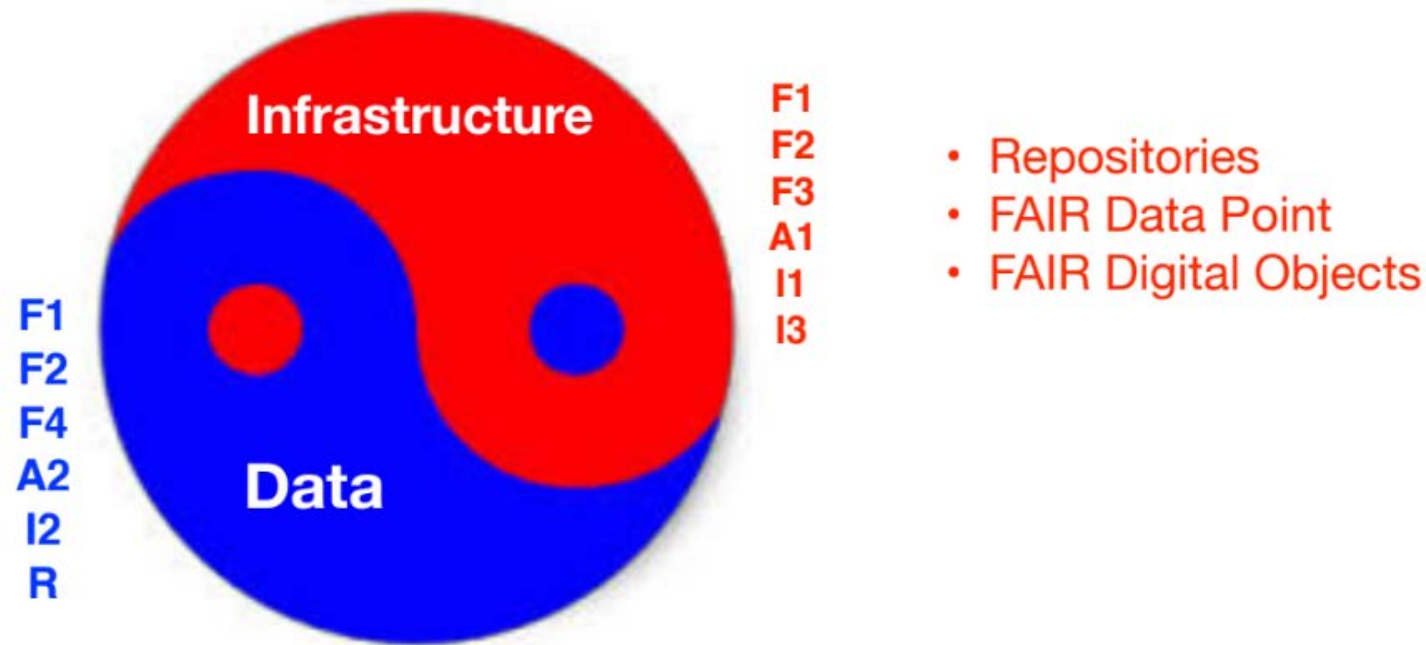
- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

Slide by Eric Schultes, GO FAIR DCSS IN Workshop Leiden 18/12/2019; https://drive.google.com/file/d/1G_bcUhsAyVG3SFZMoBnoJ59M-2zIqR9K/view

There are no fundamental technology barriers to FAIR

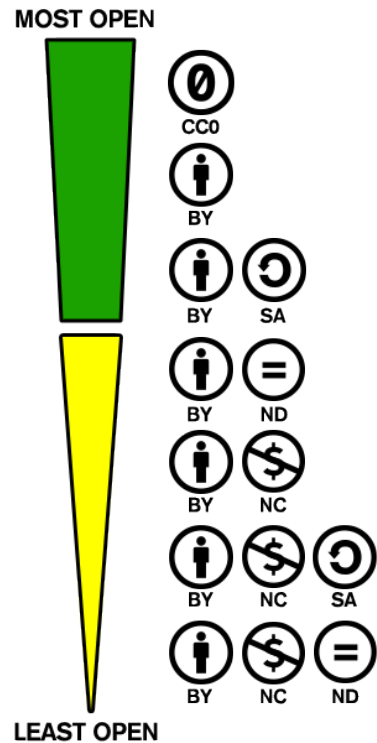


The hard problems are social

- **Community Agreements**
- **Choices**

Slide by Eric Schultes, GO FAIR DCSS IN Workshop Leiden 18/12/2019; https://drive.google.com/file/d/1G_bcUhsAyVG3SFZMoBnoJ59M-2zIqR9K/view

Licences e.g. Creative Commons



creativecommons.org,
CC BY 4.0

Only version 4.0 of:

- CC0
- CC-BY
- CC-SA

conform to the Open Definition (i.e. Open Access)

<http://opendefinition.org/od/2.1/en/>

 Search all of FAIRsharing

Search

☒ Standards ☒ Databases ☒ Policies ☒ Collections/Recommendations

Advanced Search



Fine grained control over your search.

Search Wizard



Let us guide you to your results.



1367 Standards

Terminology Artifact	763
Model/Format	398
Reporting Guideline	163
Identifier Schema	14
FAIR metrics	29

View all



1342 Databases

Natural Sciences	1237
Engineering Science	231
Humanities	48
Social Sciences	52

View all

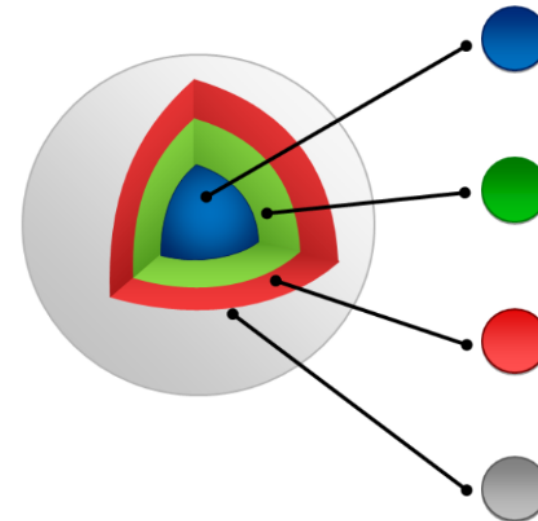
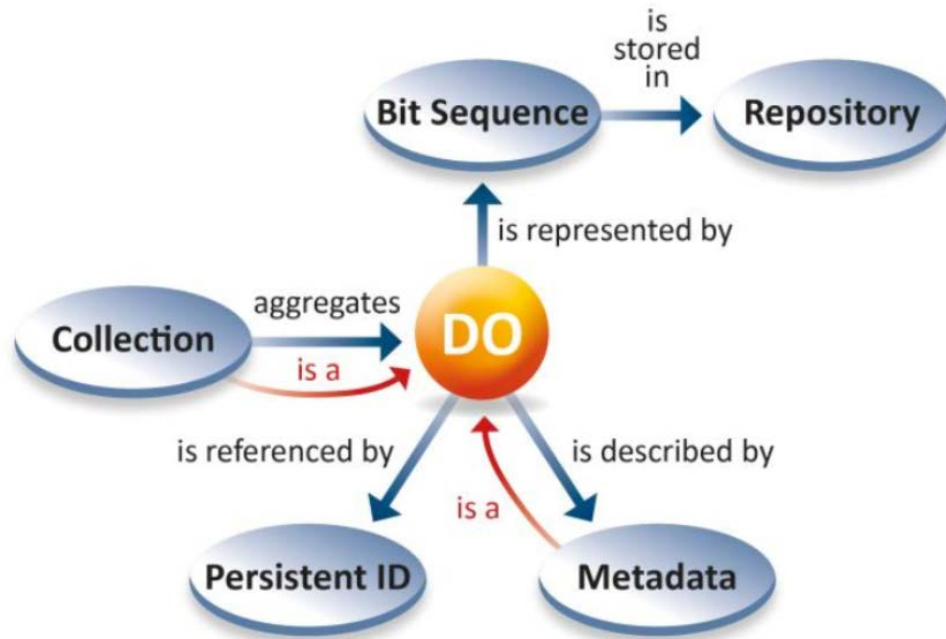


130 Policies

Funder	23
Journal	85
Society	9

View all

From *Digital Objects* to *FAIR Data Objects*



DIGITAL OBJECT

Data, code and other research resources

At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS

Persistent and unique identifiers (PIDs)

Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

STANDARDS & CODE

Open, documented formats

Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

METADATA

Contextual documentation

In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.

Reference/recommended read:

<https://github.com/GEDE-RDA-Europe/GEDE/blob/master/FAIR%20Digital%20Objects/do-fdo-papers/science-view-fdo-final.pdf>

[https://github.com/GEDE-RDA-Europe/GEDE/blob/master/Digital-Objects/Papers\(finished\)/Basic-DO/do-flavours-final.pdf](https://github.com/GEDE-RDA-Europe/GEDE/blob/master/Digital-Objects/Papers(finished)/Basic-DO/do-flavours-final.pdf)

Nationale Forschungsdateninfrastruktur (NFDI)

National Research Data Infrastructure (NFDI)

The aim of the NFDI is to **systematically manage** scientific and **research data**, provide long-term **data storage, backup** and **accessibility**, and **network** the **data** both **nationally** and **internationally**.

The NFDI will bring multiple stakeholders together in a **coordinated network of consortia** tasked with providing science-driven **data services** to **research communities**.

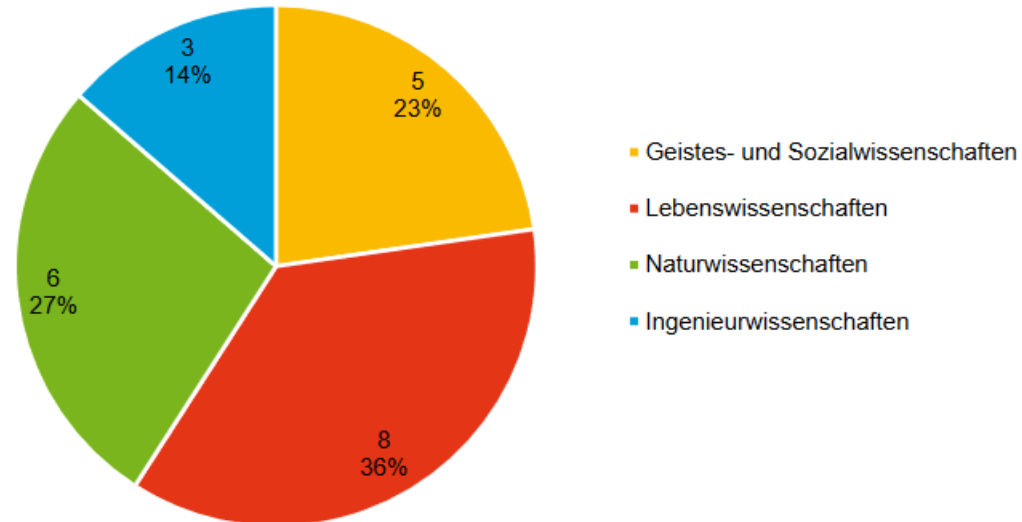
National Research Data Infrastructure (NFDI)

NFDI programme aims:

- Establishment of **data handling standards, procedures** and **guidelines** in close collaboration with the community of interest
- Development of **cross-disciplinary metadata standards**
- Development of **reliable** and **interoperable data management measures** and **services** tailored to the needs of the community of interest
- Increased **reusability** of **existing data**, also beyond subject boundaries
- Improved **networking** and **collaboration with partners outside the German academic** research system with expertise in research data management
- Involvement in developing and **establishing generic, cross-consortia services** and **standards** in research data management together with other consortia

National Research Data Infrastructure (NFDI)

- Up to 30 consortia
- 85 Mio. Euro per year
- 2-5 Mio. Euro per consortia for 5 years
- 22 proposals submitted in 2019
- 26. Juni 2020 decision published for the first round



https://www.dfg.de/download/pdf/foerderung/programme/nfdi/191212_nfdi_statistik_antragseingang.pdf

Übersicht über NFDI-Konsortien

Konsortium	Antragrunde	Absichtserklärung	Webseite	Thüringer Vertreter*
Astro@NFDI	2019	Link	Link	Matthias Hoefl
BERD@NFDI	2019	Link	Link	-
DAPHNE NFDI	2019	Link	-	-
DataPLANT	2019	Link	-	-
FAIRmat	2019	Link	Link	Silvana Botti
ForumX	2019	Link	Link	-
GHGA	2019	Link	-	-
KonsortSWD	2019	Link	Link	-
MaSDI	2019	Link	Link	-
NFDIAAgri	2019	Link	Link	-
NFDIBioDiversity	2019	Link	Link	Birgitte König-Ries, Aletta Bonn
NFDI4cet	2019	Link	Link	-
NFDI4Chem	2019	Link	Link	Christoph Steinbeck
NFDI4Culture	2019	Link	Link	-
NFDI4Earth	2019	Link	Link	Miguel Mahecha, Marica Raichstein
NFDI4Health	2019	Link	Link	-
NFDI4Ing	2019	Link	Link	-
NFDI4Medicine	2019	Link	-	Andre Scherag, Michael Kiehnopf
NFDI4MobilitTech	2019	Link	-	-
NFDI4MSE	2019	Link	Link	-
RAIN+PaN	2019	Link	Link	Stephan Fritzsche, Andre Sternbeck, Thomas Stöhr
Test+	2019	Link	-	-
2linkNFDI	2020	Link	Link	-
Bridge4NFDI	2020	Link	Link	-

https://www.researchdata.uni-jena.de/Forschung+und+Entwicklung/NFDI/NFDI_Konsortien-p-315.html

National Research Data Infrastructure (NFDI)



DFG Deutsche Forschungsgemeinschaft

DFG 2020 FÜR DAS WISSEN ENTSCHEIDEN

DFG MAGAZIN

Startseite > Förderung > Förderung auf einen Blick > Nationale Forschungsdateninfrastruktur

Nationale Forschungsdateninfrastruktur

Die nationale Forschungsdateninfrastruktur (NFDI) soll die Datenbestände von Wissenschaft und Forschung systematisch erschließen, nachhaltig sichern und zugänglich machen sowie (inter-)national vernetzen. Sie wird in einem aus der Wissenschaft getriebenen Prozess als vernetzte Struktur eigeninitiativ agierender Konsortien aufgebaut werden.

Formulare und Merkblätter >
FAQs (PDF | 287 KB) [↗](#)
Absichtserklärungen >
Stellungnahme des NFDI-Expertengremiums (PDF | 136 KB) [↗](#)

Aktuelles

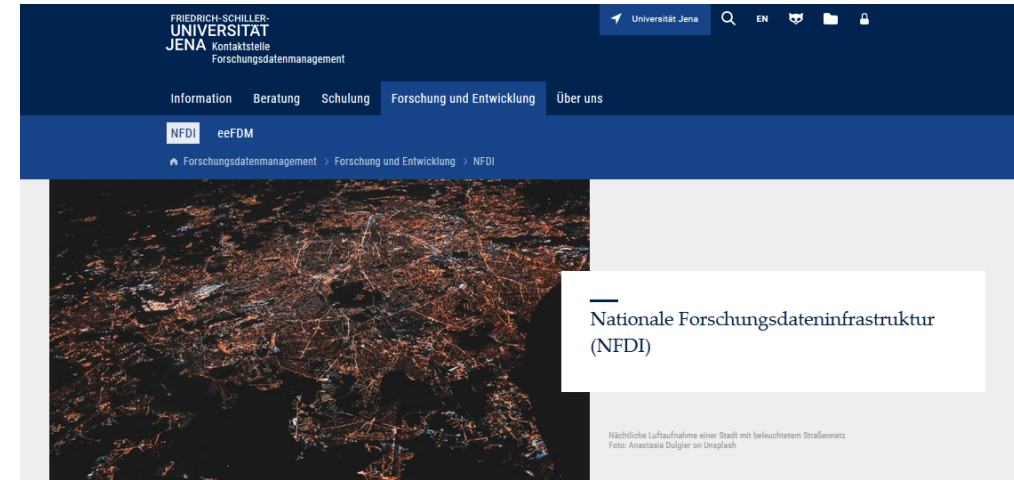
Im Rahmen der Ausschreibung sind zum 15. Oktober 2019 insgesamt 22 Anträge auf Förderung von NFDI-Konsortien eingegangen. An den Anträgen waren insgesamt 142 verschiedene Einrichtungen antrag- und/oder mitantragsstellend beteiligt. 2020 und 2021 sind jeweils weitere Ausschreibungsrunden geplant. Statistische Übersichten zum Antragseingang 2019 finden Sie hier:

- Statistische Übersichten zum Antragseingang [↗](#)

Antragstellung

▼ Kompaktdarstellung

<https://www.dfg.de/foerderung/programme/nfdi/>



FRIEDRICH-SCHILLER-UNIVERSITÄT JENA Kontaktstelle Forschungsdatenmanagement

Information Beratung Schulung **Forschung und Entwicklung** Über uns

NFDI eeFDM

Forschungsdatenmanagement > Forschung und Entwicklung > NFDI

Nationale Forschungsdateninfrastruktur (NFDI)

Nächtliche Luftaufnahme einer Stadt mit beleuchtetem Straßennetz
Foto: Anastasia Dulger on Unsplash

DFG Förderprogramm zum Aufbau einer Nationalen Forschungsdateninfrastruktur

Der Rat für Informations Infrastrukturen (RII) hat 2016 in seinem Positionspapier "Leistung aus Vielfalt" [pdf 1MB] die Etablierung einer Nationalen Forschungsdateninfrastruktur (NFDI) empfohlen und seitdem weitere Hinweise zu diesem Thema veröffentlicht (siehe weitführende Informationen). Am 6. Juni 2019 wurde von der Deutschen Forschungsgemeinschaft (DFG) das Förderprogramm zum Aufbau der NFDI vorgestellt. Im Rahmen dieses von Bund und Ländern finanzierten Programms sollen bis zu 30 Konsortien über einen Zeitraum von zunächst maximal 5 Jahren gefördert werden. Ziel NFDI ist es, Forschungsdatenbestände systematisch zu erschließen, langfristig zu sichern und unter Einhaltung der FAIR Prinzipien international zugänglich zu machen.

Informationen und Dokumente der DFG

- Statistische Übersichten zum Antragseingang zur NFDI (2019) [pdf 645KB]
- Ausschreibungstext zur NFDI
- Stellungnahme des NFDI-Expertengremiums
- Förderkriterien für NFDI-Konsortien
- Merkblatt zu Antragsstellung
- EN

<https://www.researchdata.uni-jena.de/Forschung+und+Entwicklung/NFDI.html>

European Open Science Cloud



Some Links ...

DARIAH-DE

<https://de.dariah.eu/daten>

CLARIN for Researchers

<https://www.clarin.eu/content/clarin-for-researchers>

UK Data Service

<https://www.ukdataservice.ac.uk/manage-data.aspx>

MANTRA Data Management Plans

<https://mantra.edina.ac.uk/datamanagementplans/>

Some Links ...

João Luiz Rebelo Moreira, Luiz Bonino, Luís Ferreira Pires, Marten van Sinderen, Patricia Henning, 2019, Towards Findable, Accessible, Interoperable and Reusable (FAIR) Data Repositories: Improving a Data Repository to Behave as a FAIR Data Point.

<https://doi.org/10.18617/liinc.v15i2.4817>

Wilkinson, M. D. et al. A design framework and exemplar metrics for FAIRness. Sci. Data 5:180118 doi: 10.1038/sdata.2018.118 (2018).

Turning FAIR into reality, Final report and action plan from the European Commission expert group on FAIR data (2018), DOI: [10.2777/1524](https://doi.org/10.2777/1524)



Thanks for joining!

Questions?

References I

- Biernacka, Katarzyna, Dominika Dolzycka, Kerstin Helbig & Petra Buchholz (2018). Train-the-Trainer Workshop zum Thema Forschungsdatenmanagement. DOI: 10.5281/zenodo.1215376.
- DataCite DOI Basics. <https://support.datacite.org/docs/doi-basics#section-dois-for-research-data>. [05.04.2019].
- DataONE Education Module: Data Citation. DataONE. http://www.dataone.org/sites/all/documents/L09_DataCitation.pptx. Retrieved April 05, 2019.
- DataONE Education Module: How to Write Good Quality Metadata. DataONE. Retrieved Nov 12, 2012. http://www.dataone.org/sites/all/documents/L08_WriteQualityMetadata.pptx
- DataONE Education Module: Metadata. DataONE. Retrieved Nov 12, 2016. http://www.dataone.org/sites/all/documents/L07_Metadata.pptx .

References II

- DFG (2015). Leitlinien zum Umgang mit Forschungsdaten.
https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf [09.09.2019].
- Helbig, Kerstin (2017). Einführung in das Forschungsdatenmanagement RDA-DE-Trainings-Workshop A1.
- Mantra. Data management plans. <https://mantra.edina.ac.uk/datamanagementplans/>
- Rex, Jessica (2019). *Forschungsdaten effizient handhaben - eine Einführung ins Forschungsdatenmanagement*. DOI: 10.5281/zenodo.2579580 .
- Rümpel, Stefanie (2011). Der Lebenszyklus von Forschungsdaten. Büttner, Stephan, Hans-Christoph Hobohm & Lars Müller, Hg. *Handbuch Forschungsdatenmanagement*. Bad Honnef: Bock + Herchen. 25-34.
<https://opus4.kobv.de/opus4-fhpotsdam/frontdoor/index/index/docId/207>

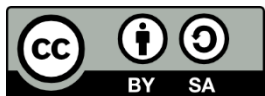
References III

- Spindler, Gerald & Tobias Hillegast (2011). Rechtliche Probleme der elektronischen Langzeitarchivierung von Forschungsdaten. Büttner, Stephan, Hans-Christoph Hobohm & Lars Müller, Hg. *Handbuch Forschungsdatenmanagement*. Bad Honnef: Bock + Herchen. 63-69.
<https://opus4.kobv.de/opus4-fhpotsdam/frontdoor/index/index/docId/207>
- Tarbet, Amanda. Metadata concept map. <https://libguides.bc.edu/c.php?g=44295&p=280677>.
- Vierkant, Paul (2015). Freier Zugang zu Forschungsdaten – Setting the Scene. <http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:1369629:2/component/escidoc:1369630/1369629.pdf>.
- Wilkinson, Mark D., Dumontier, Michel, Aalbersberg, I. Jsbrand Jan, Appleton, Gabrielle, Axton, Myles & Baak, Arie u.a. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific data* 3, Artikelnr. 160018. <https://www.nature.com/articles/sdata201618.pdf> .

Please, quote as follows:

Gerlach, Roman, Hesse, Bettina, König-Ries, Birgitta, Schröter, Annett, Rzymiski, Christoph, Schwartze, Volker, Steiner, Petra u. Olena Tykhostup (2020). *What's on in Research Data Management?*. Presentation by Roman Gerlach on 29/01/2020 at „Reproducible Research and Data Management workshop” at the Max Planck Institute for the Science of Human History.

License:



(CC BY-SA 4.0)