



MAX
PLANCK
DIGITAL
LIBRARY

Onboarding into Research Data Management

Training Course for Max Planck Early Career Researcher

Göttingen, Max Planck Institute for Multidisciplinary Sciences, 8th February 2024

<https://hdl.handle.net/21.11116/0000-000E-1943-B>

Michael Franke, franke@mpdl.mpg.de

David Walter, d.walter@mpdl.mpg.de

Dr. Yves Vincent Grossmann grossmann@mpdl.mpg.de



Introduction

Speakers

Michael Franke

- <https://orcid.org/0000-0002-2661-8242>
- Division Manager MPDL Collections
- 2011-2019: Alliance WG “Research Data”
- 2015-2018: Science Europe WG “Research Data”
- 2019-2023: Alliance WGs “Digital Tools” and “Federated IT Infrastructures”
- 2014-2020: RDM teaching position at HföD Bavaria
- franke@mpdl.mpg.de

David Walter

- <https://orcid.org/0000-0001-6807-5007>
- since 2024: Research Data Management Officer at the MPDL
- since 2021: Data management of S/Y Eugen Seibold, editor at PANGAEA
- 2014-2020: Greenhouse gas measurements and data management at ATTO
- 2013: PhD in Physics (DOAS technique onboard CARIBIC aircraft)
- d.walter@mpdl.mpg.de

Programme for Today

09:00-09:15	Welcome and introduction
09:15-09:30	Topic block 1: Digital research data (D)
09:30-10:10	Topic block 2: Normative Frameworks (M)
10:10-10:30	Topic block 3: Data management plans (M)
10:30-10:50	Coffee break
10:50-11:15	Topic block 4: Order and Structure (D)
11:15-11:45	Topic Block 5: Saving
11:45-12:00	Topic block 6: Access security

12:00-13:00	Lunch (self-pay basis) and photo
13:00-14:00	Topic block 7: MPG-internal services (D)
14:00-14:20	Topic block 8: Metadata (M)
14:20-15:15	Topic block 9: Publishing (D)
15:15-15:45	Coffee break
15:45-16:05	Theme Block 10: Electronic Laboratory Notebook (M)
16:05-16:25	Theme Block 11: New Horizons (M)
16:25-16:30	Closing Remarks

Organisational Issues

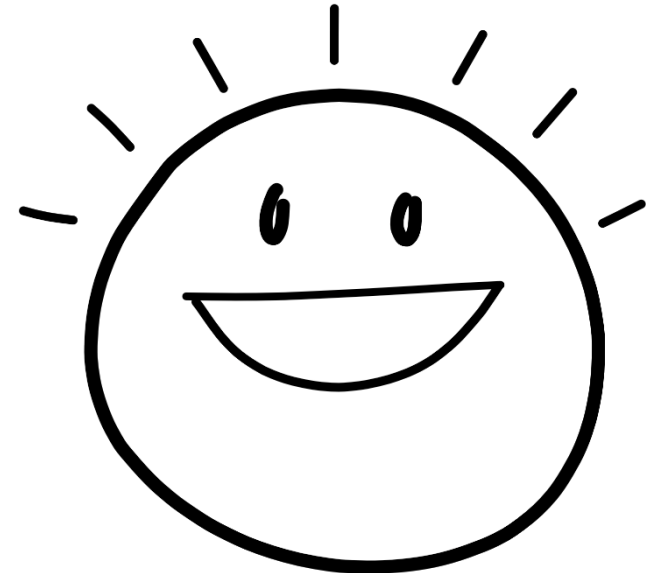
- Questions:
 - Ask immediately
 - Ask after the thematic block
 - Ask in the break
 - Ask afterwards: rdm@mpdl.mpg.de
- Slides are available on MPG.PuRe



<https://doi.org/10.5281/zenodo.3674561>

Objectives for Today

1. Develop basic understanding of Research Data Management
2. See that research data management is not rocket science (except in rocket science)
3. Insight into the various RDM topics
4. Exchange and clarify questions
5. Suggestions and communication of new topics



<https://doi.org/10.5281/zenodo.3674561>

Questions to You via Slido

Join via slido.com and enter 1174667

1. How would you rate your level of experience in working with research data?
2. What do you expect from this course?

Digital Research Data

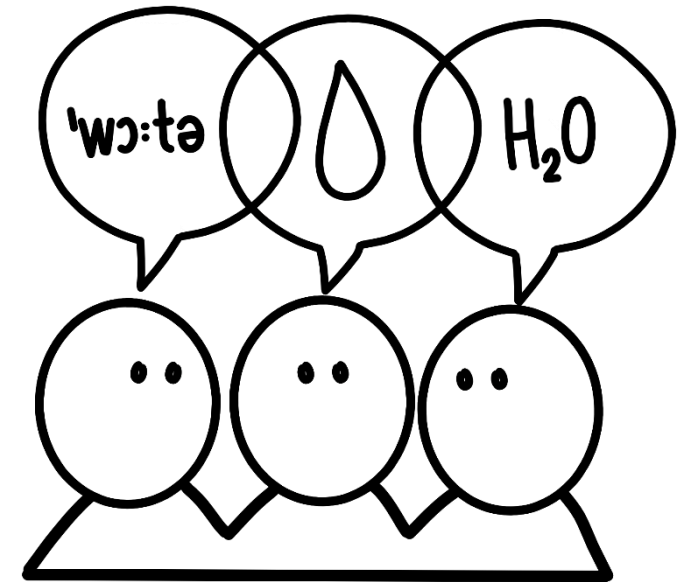
- Research Data Management
- Definition of research data
- Research data lifecycles
- Examples

Worum geht es?



What is Research Data Management?

- Explicit handling of research data before, during and after the project



<https://doi.org/10.5281/zenodo.3674561>

Why Research Data Management?

You **should** think about it, because:

- Re-use data, at least oneself, but also by others
- Scientific reputation increasing
- Acceptance of data as a separate publication is likely to increase in the future
- Because “fits somehow” has never won a Nobel Prize
- ...

Why Thinking about RDM?

- Time and resource savings through early RDM
 - Potential problems are identified earlier
 - Save time by organising so that less time is spent searching
- Easier project management for participating scientists
 - Less communication effort
- Reduction of project complexity through structured procedure
 - Easier processing for the oneself

Why Research Data Management?

You **have to** think about it, because:

- MPG Good Scientific Practice
- DFG Code of Conduct
- Funding Agencies, i.e. by DFG and Horizon Europe, ERC, Volkswagen Foundation
- Publishers
- ...

Example RDM Tasks

1. Planning the handling of research data at the beginning of a research project and, if necessary, presenting the planned measures in funding applications
2. Defining a folder structure and file naming conventions
3. Documentation of research data and labelling with metadata
4. Backup and long-term archiving of research data
5. IT security and access rights for research data
6. Publication of research data
7. Finding and reusing existing research data
8. Consideration of data protection and copyright law when handling research data

Definition Attempts Data

“Research data include measurement data, laboratory values, audiovisual information, texts, survey data or observational data, methodological test procedures and questionnaires.”

DFG Checklist, 21st December 2021, p. 1,

https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_en.pdf

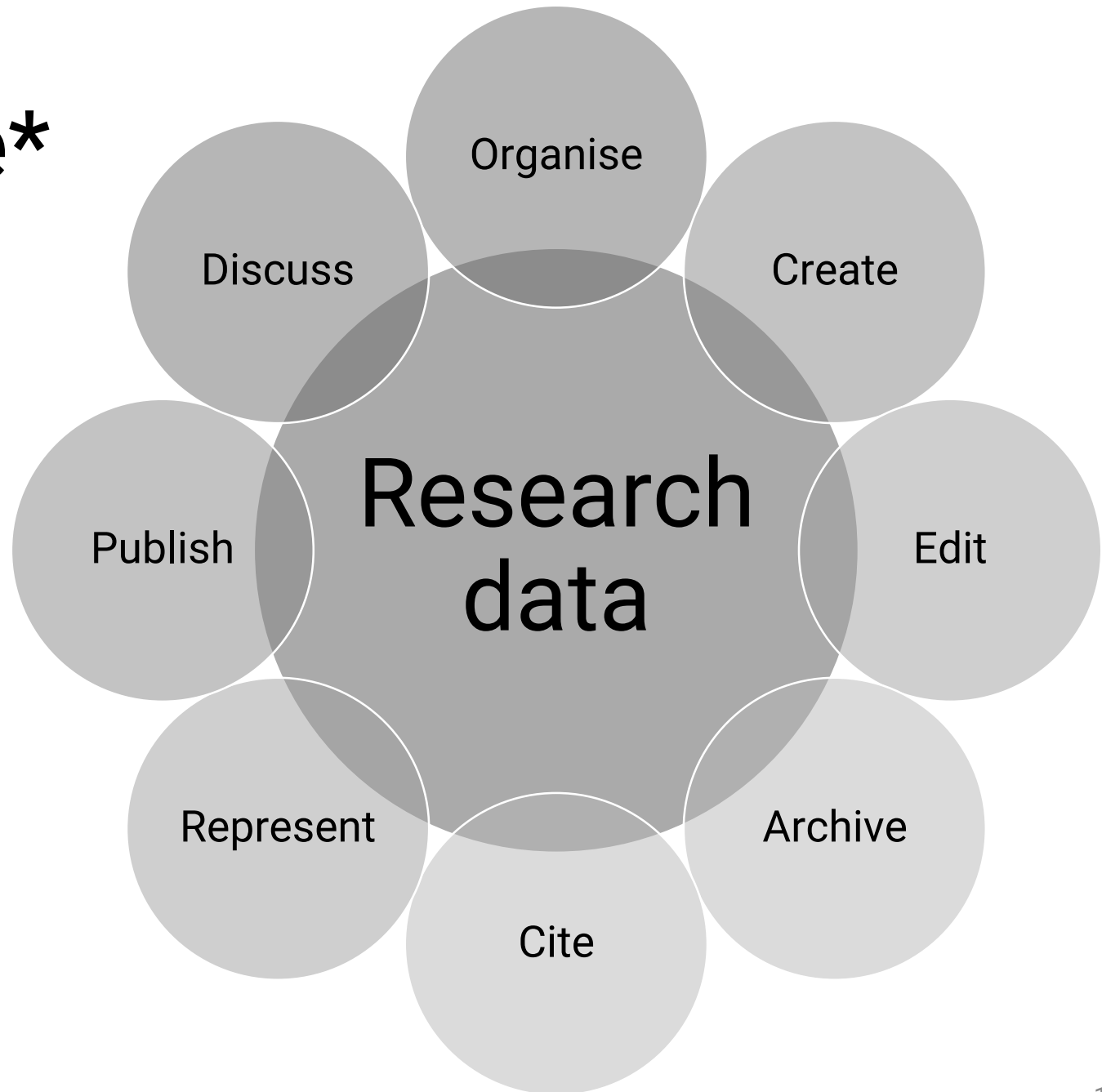
Research data are “data generated in the course of scientific projects, e.g. through digitisation, source research, experiments, measurements, surveys or interviews.”

Allianz AG “Forschungsdaten”, Forschungsdatenmanagement – Eine Handreichung, 2018, p. 4, https://gfzpublic.gfz-potsdam.de/rest/items/item_3055893_5/component/file_3055894/content

“Research data comprise all data generated in the scientific work process and processed in digital form.”

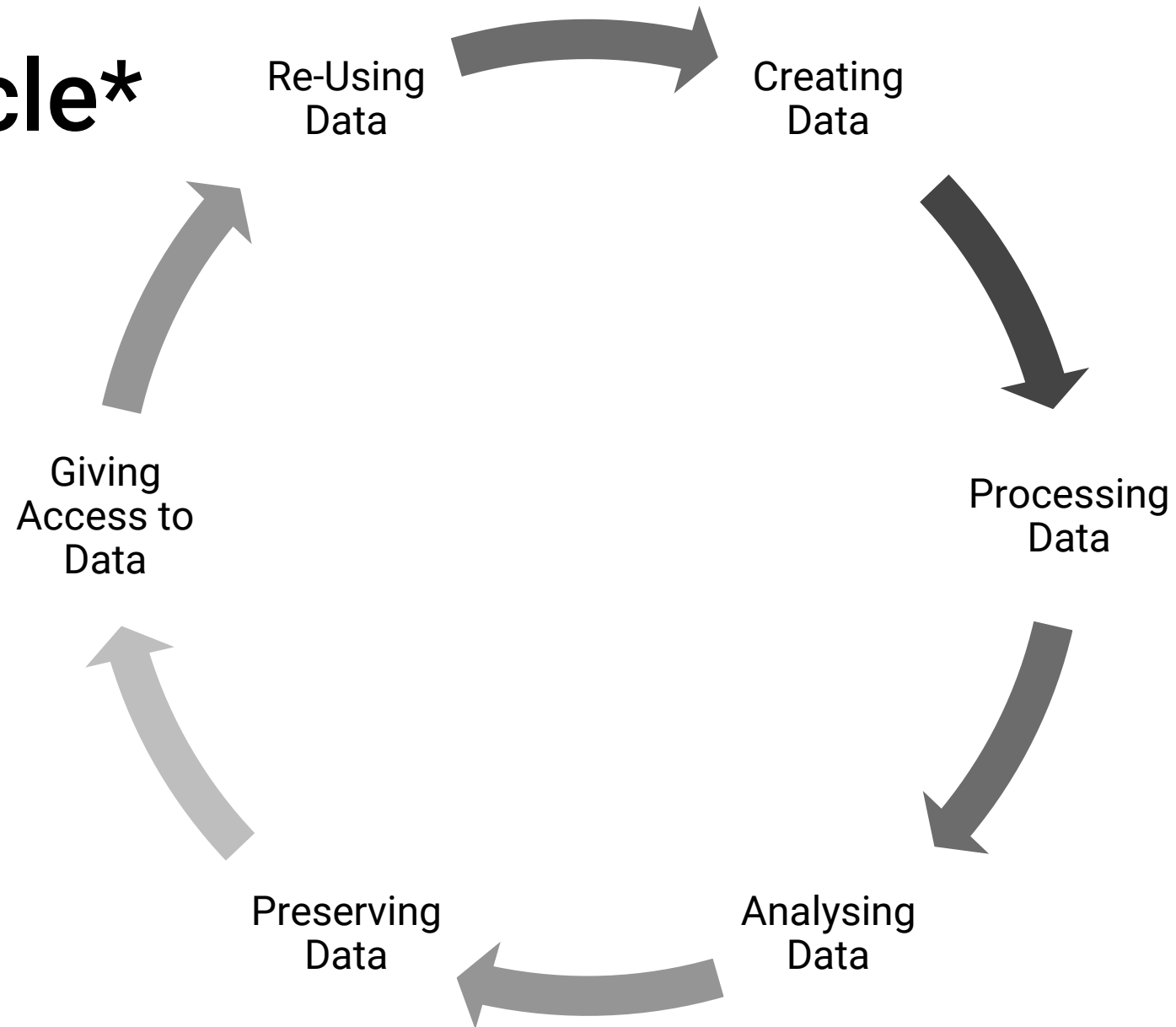
Guideline on handling research data in the Leibniz Association, 2018, p. 1, https://www.leibniz-gemeinschaft.de/fileadmin/user_upload/Bilder_und_Downloads/Forschung/Open_Science/Leitlinie_Forschungsdaten_2018.pdf

The Big Picture*



* which only very schematically presents the handling of research data...

Research Data Cycle*



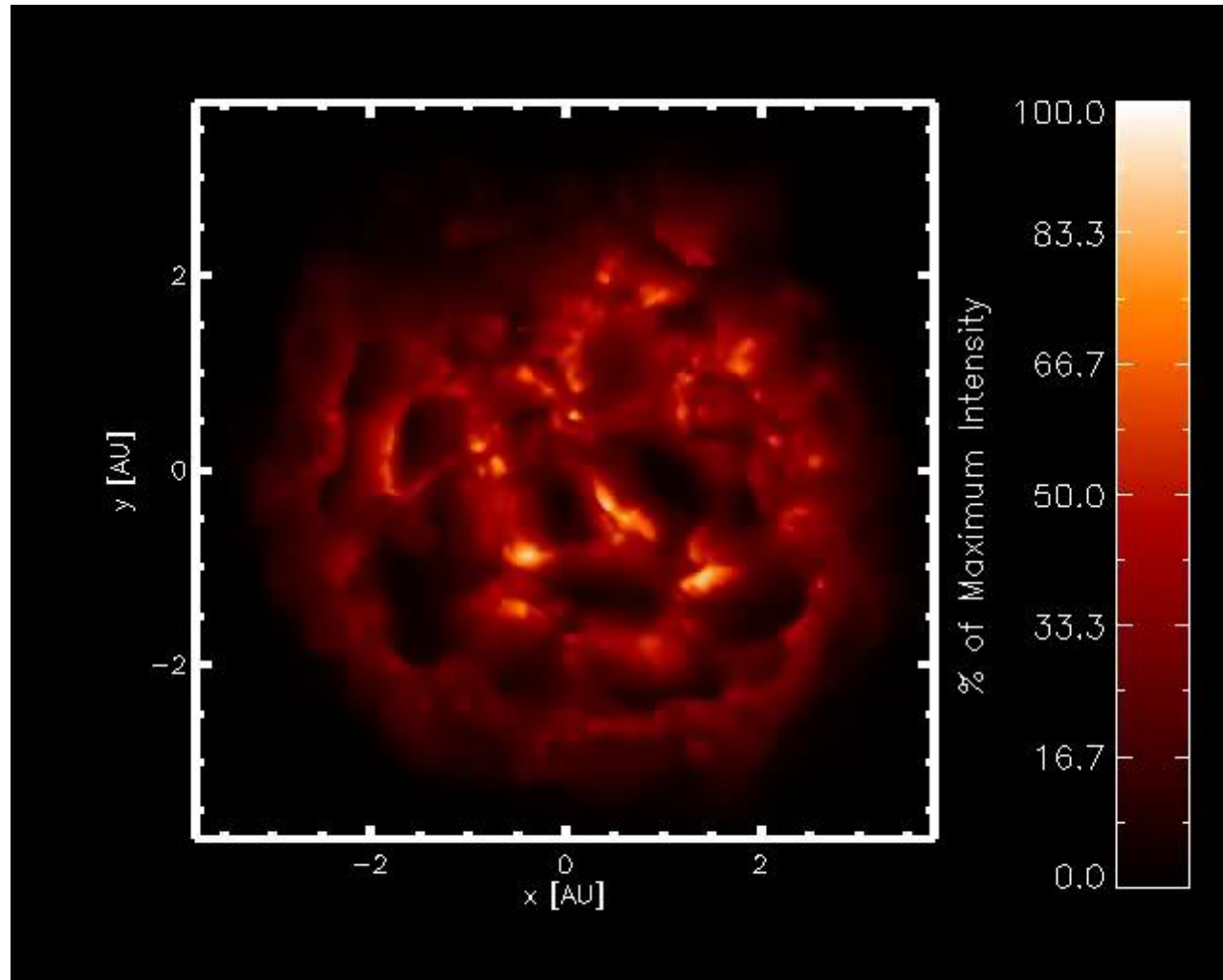
Short Audio Quiz

What is hidden behind this audio file of a research dataset?



Resolution: Cecilia Durojaye, 2020, "Recodings - Nigerian Talking drum", file D1s, Edmond, <https://doi.org/10.17617/3.IUMMLP> , CC BY 4.0.

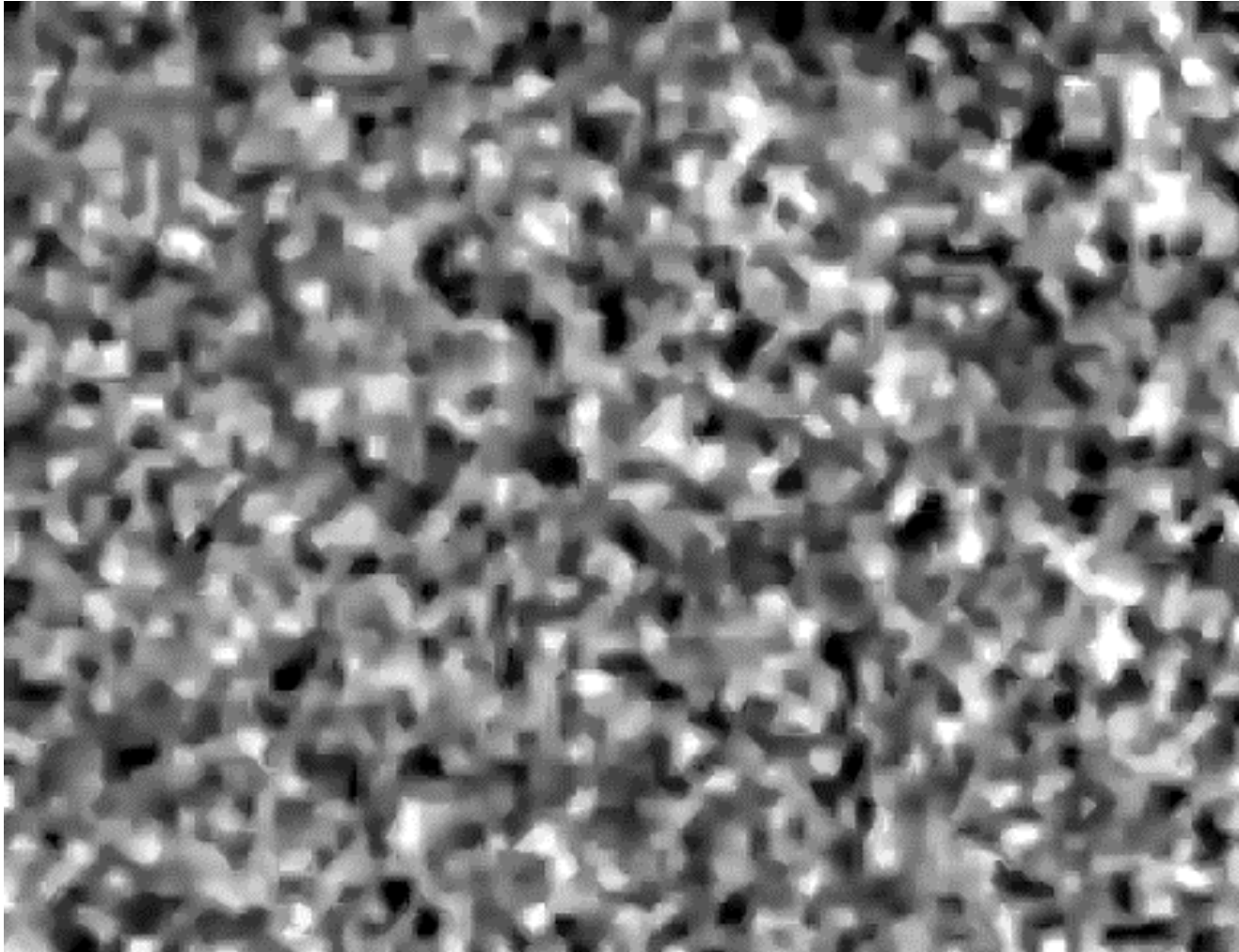
Short Video Quiz



What is hidden behind this video file of a research dataset?

Chiavassa, R. Kudritzki, B. Davies, B. Freytag, & S. E. de Mink. (2022, March 16): Photo-center displacements for RSGs as seen by Gaia (Version v1), Max Planck Institute for Astrophysics, CC BY 4.0, Zenodo, <https://doi.org/10.5281/zenodo.6363011>.
Paper: <https://www.aanda.org/articles/aa/pdf/2022/05/aa43568-22.pdf>

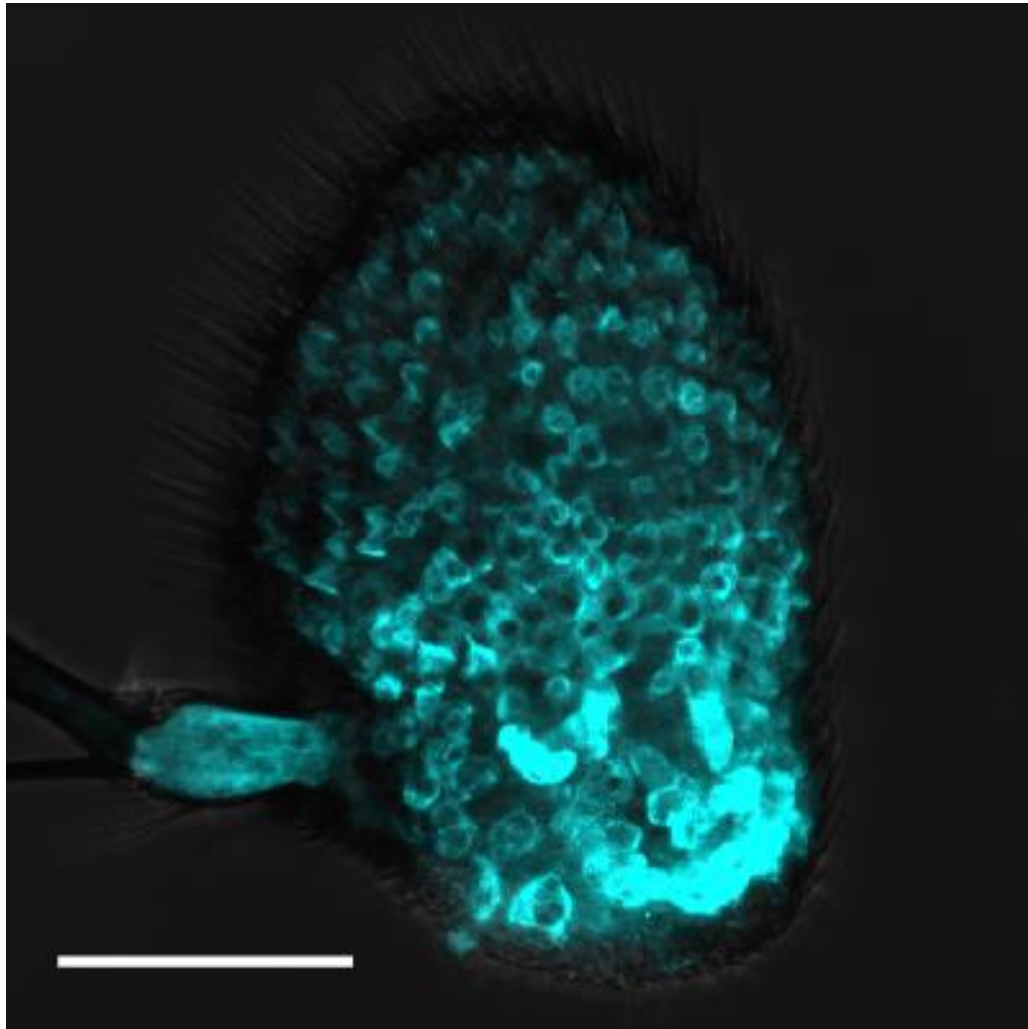
Short Video Quiz



What is hidden behind this video file of a research dataset?

Kaumudi Prabhakara (Max Planck Institute for Dynamics and Self-Organization), Spiralwaves, 2015, CC BY 4.0,
<https://www.youtube.com/watch?v=DmRZn073Uus>.

Short Image Quiz



What is hidden behind this image file of a research dataset?

Prelic, Sinisa, 2021, "Functional interaction between *Drosophila* olfactory sensory neurons and their support cells", Max Planck Institute for Chemical Ecology, <https://doi.org/10.17617/3.7m>, Edmond, V1, CC BY 4.0, 20191021 image 4 ASE5-Gal4 GFP.tif.

Normative Frames

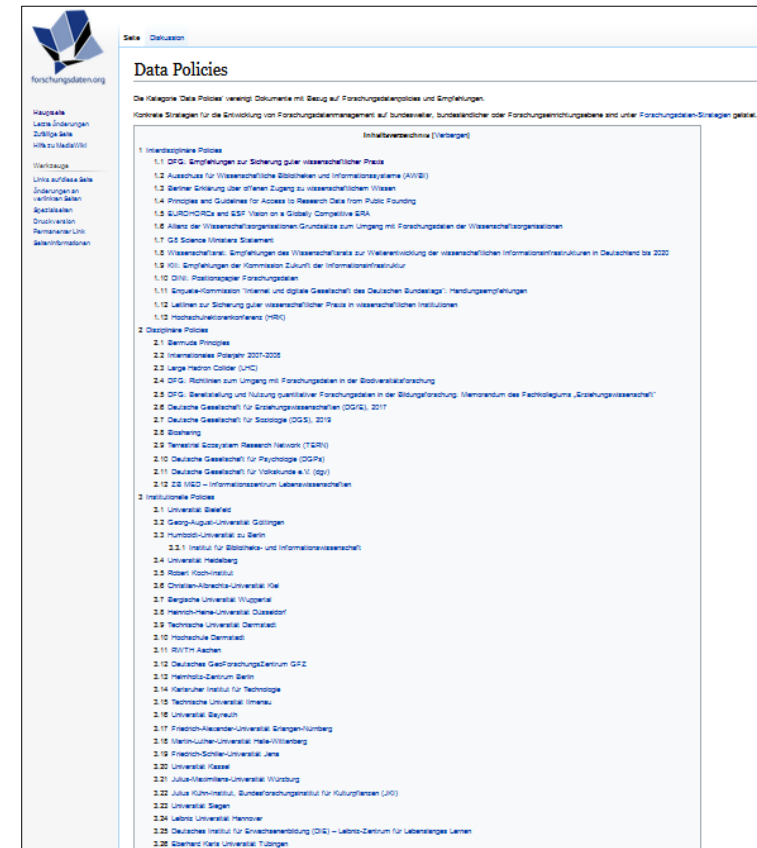
- What is an RD policy?
- New MPG-GWP rules
- Possible institutional policies
- Subject-specific policies
- Perspectives of the funders
- Perspectives of the publishers

What is a Research Data Policy?

A research data policy describes the guidelines for handling data.

There are different types of policies, for example:

- Journal and publisher policies
- Institutional policies
- Project-specific policies
- Subject-specific policies
- Policies of research funders

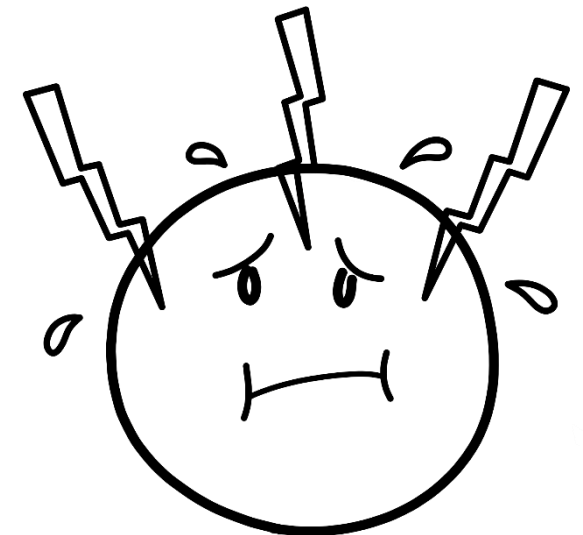


The screenshot shows a webpage titled "Data Policies" from the research data portal. The page is in German and provides a list of various research data policies. The content is organized into three main sections: 1. Interdisziplinäre Politik (Interdisciplinary Policy), 2. Disziplinäre Politik (Disciplinary Policy), and 3. Institutionelle Politik (Institutional Policy). Each section contains a numbered list of specific policies and guidelines, such as "1.1 DFG: Empfehlungen zur Sicherung guter wissenschaftlicher Praxis" and "3.1 Universität Bielefeld". The page also includes a navigation menu on the left and a search bar at the top.

https://www.forschungsdaten.org/index.php/Data_Policies, CC BY 4.0

Normative Requirements

- Standard for “Good Scientific Practice” by institutions
- Research data policies of institutions
- Requirements of funding institutions
- Request from publishers
- ...



<https://doi.org/10.5281/zenodo.3674561>

MPG “Good Scientific Practice”

“2.4 Securing and storing primary data – Documentation and archiving

[.] The Institute Management is expected to provide the usual storage media for the field concerned and to guarantee that information stored both digitally and in analogue format is secured and remains accessible. The framework conditions must be such that protection from unauthorized access, loss, destruction, theft, and manipulation can be guaranteed. [.]

Research Group Leaders and individual researchers are obligated to make use of the protection options provided by Institute Management and retain and store both research data and research results. It does not matter in this context whether the research results are published or not.“

(<https://www.mpg.de/197494/rulesScientificPractice.pdf>, p. 41)

MPG “Good Scientific Practice”

“2.4 Securing and storing primary data – Documentation and archiving

[..] Cooperative research also requires continuous, long-term availability of research data. For this, corresponding research data management specific to the discipline must also be ensured. All Research Group Leaders involved bear joint responsibility for this.“

(<https://www.mpg.de/197494/rulesScientificPractice.pdf>, p. 41)

“2.7 Accessibility of research data

For reasons of traceability, follow-on research and potential later use, scientists store the research data and central materials on which the publication is based whenever possible – following the FAIR principles (see I. 2.5 Publications – Authorships) – e.g. in accessible, commonly recognized archives and repositories.“

(<https://www.mpg.de/197494/rulesScientificPractice.pdf>, p. 45)

FAIR Data Principles

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

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

Accessible

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

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

Data Culture in the Own Research Field

- Comparatively difficult to catch → more tacit knowledge
- Common understanding about “data”
- Community specific file formats
 - E.g. [NeXus](#) for neutron, X-ray and muon research (HDF5 format with specified metadata information)
- Established venues for data publication
 - e.g. subject-specific repository

DFG's Code of Conduct "Safeguarding Good Research Practice"

Guideline 13: Providing public access to research results

"[...] If it has been decided to make results available in the public domain, **researchers describe them clearly and in full**. Where possible and reasonable, this includes making the research data, materials and information on which the results are based, as well as the methods and software used, available and fully explaining the work processes. [...]"
(<https://doi.org/10.5281/zenodo.3923601>, p. 17)

Guideline 17: Archiving

"When scientific and academic findings are made publicly available, the research data (generally raw data) on which they are based are generally archived in **an accessible and identifiable manner for a period of ten years at the institution** where the data were produced or in cross-location repositories." (<https://doi.org/10.5281/zenodo.3923601>, p. 20)

Deutsche Forschungsgemeinschaft (DFG)

“For this reason, the DFG expects research projects to include a description of how research data is handled. The description should be based on the checklist for handling research data (to the questionnaire).”

- “The recommendation is that **contact** should be established as early as possible during the project planning phase with a research data centre or **repository** where the research data can be deposited”.
- “The description of how research data will be handled is included in the review and also forms part of the **reporting obligation** after completion of the project.”

www.dfg.de/proposal_process/research_data

Versionsdatum 21.12.2021

Umgang mit Forschungsdaten


Checkliste für Antragstellende zur Planung und zur Beschreibung des Umgangs mit Forschungsdaten in Forschungsvorhaben

Diese Checkliste unterstützt Sie, wesentliche Aspekte des Umgangs mit Forschungsdaten strukturiert zu beschreiben sowie die für die Umsetzung benötigten Ressourcen und Kompetenzen sichtbar zu machen. Bitte nehmen Sie zu den unten genannten Themenfeldern im Antrag unter Punkt 2.4 Stellung.

Zu Forschungsdaten zählen u. a. Messdaten, Laborwerte, audiovisuelle Informationen, Texte, Surveydaten oder Beobachtungsdaten, methodische Testverfahren sowie Fragebögen. Korpora, Software und Simulationen können ebenfalls zentrale Ergebnisse wissenschaftlicher Forschung darstellen und werden daher ebenfalls unter den Begriff Forschungsdaten gefasst. Da Forschungsdaten in einigen Fachbereichen auf der Analyse von Objekten basieren (z. B. Gewebe-, Material-, Gesteins-, Wasser- und Bodenproben, Prüfkörper, Installationen, Artefakte und Kunstgegenstände), muss der Umgang mit diesen ebenso sorgfältig sein und eine fachlich adäquate Nachnutzungsmöglichkeit, wann immer sinnvoll und möglich, mitgedacht werden. Steht die Nachnutzbarkeit der entstehenden Forschungsdaten in engem Zusammenhang mit Objekten, so bitten wir Sie, auch entsprechende Angaben zu diesen zu ergänzen.

Bitte berücksichtigen Sie die in Ihrer Fachdisziplin existierenden Standards und ggf. bestehende fachspezifische Empfehlungen und Angebote existierender Infrastrukturen (z. B. Datenrepositorien, Archive oder Sammlungen). Einen Überblick über existierende Strukturen bietet das Portal für Forschungsinfrastrukturen Resources (<https://resources.dfg.de>) sowie das Verzeichnis von Forschungsdatenrepositorien re3data (<http://re3data.org>).

Weitere Informationen zum Thema und fachspezifische Empfehlungen finden Sie unter: www.dfg.de/antragstellung/forschungsdaten/



https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/forschungsdaten/forschungsdaten_checkliste_en.pdf

European Commission

ERC Work Programme 2021:

“Finally, as from 2021 it is no longer possible for applicants to opt out of the submission of Research Data Management plans.”

(European Commission Decision C(2021) 930 of 22/02/2021, p. 4)

Marie Skłodowska-Curie Actions:

“[...] data management plan submitted at mid-term and an update towards the end of the project if needed” (European Commission Decision C(2022)7550 of 6 December 2022, p. 88)

RDM as Career Development Activities

3. Career Development Activities

In line with the [European Charter for Researchers](#), the MSCA put special emphasis on skills development and improving the career prospects of researchers.

Therefore you may undertake professional training and development activities during your fellowship, such as:

- Complementary training in transferable skills such as proposal writing, research management, open science, FAIR² data management, intellectual property rights, patent submission, innovation and entrepreneurship, communication, etc.
- Attendance of conferences and workshops to boost your competences and networking capacity
- Outreach and/or citizen engagement activities
- Teaching
- Language courses

Horizon Europe

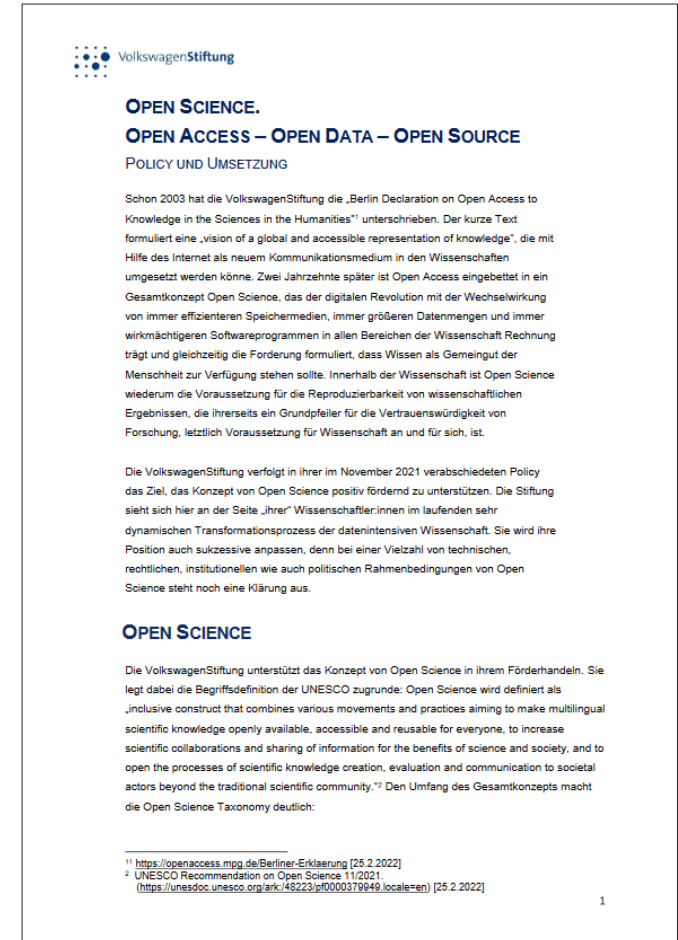
"The beneficiaries must manage the digital research data generated in the action ('data') responsibly, in line with the FAIR principles and by taking all of the following actions:

- **establish a data management plan** ('DMP') (and regularly update it)
- as soon as possible and within the deadlines set out in the DMP, **deposit the data in** a trusted repository [...].
- as **soon as possible** and within the deadlines set out in the DMP, ensure **open** access - via the repository - to the deposited data [plus CC0, CC BY or equivalent]".

Horizon Europe and Euratom: General Model Grant Agreement, Version 1.0, 1st June 2021, p. 109,
https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/agr-contr/general-mga_horizon-euratom_en.pdf

Volkswagen Foundation

- Open Science Policy since 2021
 - “The Volkswagen Foundation encourages researchers to store the data generated within funded projects in public, non-commercial repositories [..]” (p. 4)
 - “In the application process, data generating and data using projects working in disciplines lacking a clear data workflow are asked to upload a digital concept with a data management plan”. (pp. 4-5)
 - “In the application process, the applicants are asked to mark separately the already generated Open Data in the CV for the review and decision process.” (p. 5)
- Funding opportunity: “Data Reuse – Additional Funding for the Preparation and Storage of Research Data (Open Science)”



The image shows a document titled "VolkswagenStiftung OPEN SCIENCE. OPEN ACCESS – OPEN DATA – OPEN SOURCE POLICY UND UMSETZUNG". The text discusses the foundation's commitment to open science, referencing the Berlin Declaration on Open Access to Knowledge in the Sciences in the Humanities and the UNESCO Recommendation on Open Science. It outlines the foundation's goal to support open science and mentions the Volkswagen Foundation's support for the concept of open science in their funding process. The document also includes a footnote with a URL and a page number "1".

VolkswagenStiftung

OPEN SCIENCE.
OPEN ACCESS – OPEN DATA – OPEN SOURCE
POLICY UND UMSETZUNG

Schon 2003 hat die VolkswagenStiftung die „Berlin Declaration on Open Access to Knowledge in the Sciences in the Humanities“ unterschrieben. Der kurze Text formuliert eine „vision of a global and accessible representation of knowledge“, die mit Hilfe des Internet als neuem Kommunikationsmedium in den Wissenschaften umgesetzt werden könne. Zwei Jahrzehnte später ist Open Access eingebettet in ein Gesamtkonzept Open Science, das der digitalen Revolution mit der Wechselwirkung von immer effizienteren Speichermedien, immer größeren Datenmengen und immer wirkmächtigeren Softwareprogrammen in allen Bereichen der Wissenschaft Rechnung trägt und gleichzeitig die Forderung formuliert, dass Wissen als Gemeingut der Menschheit zur Verfügung stehen sollte. Innerhalb der Wissenschaft ist Open Science wiederum die Voraussetzung für die Reproduzierbarkeit von wissenschaftlichen Ergebnissen, die ihrerseits ein Grundpfeiler für die Vertrauenswürdigkeit von Forschung, letztlich Voraussetzung für Wissenschaft an und für sich, ist.

Die VolkswagenStiftung verfolgt in ihrer im November 2021 verabschiedeten Policy das Ziel, das Konzept von Open Science positiv fördernd zu unterstützen. Die Stiftung sieht sich hier an der Seite „ihrer“ Wissenschaftler:innen im laufenden sehr dynamischen Transformationsprozess der datenintensiven Wissenschaft. Sie wird ihre Position auch sukzessive anpassen, denn bei einer Vielzahl von technischen, rechtlichen, institutionellen wie auch politischen Rahmenbedingungen von Open Science steht noch eine Klärung aus.

OPEN SCIENCE

Die VolkswagenStiftung unterstützt das Konzept von Open Science in ihrem Förderhandeln. Sie legt dabei die Begriffsdefinition der UNESCO zugrunde: Open Science wird definiert als „inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community.“¹ Den Umfang des Gesamtkonzepts macht die Open Science Taxonomy deutlich:

¹ <https://openaccess.mpg.de/Berliner-Erklärung> [25.2.2022]
² UNESCO Recommendation on Open Science 11/2021.
https://unesdoc.unesco.org/ark:/48223/pf0000379649_locale=en [25.2.2022]

1

Questions to You via Slido

Join via slido.com and enter 1174667

1. Have you already received an RDM training?
2. Are you currently writing an application with data reference?

20 minutes Coffee Break

Data Management Plans

- What is a DMP
- DMP components
- Tools and templates
- DMP examples

Data Management Plan (DMP)

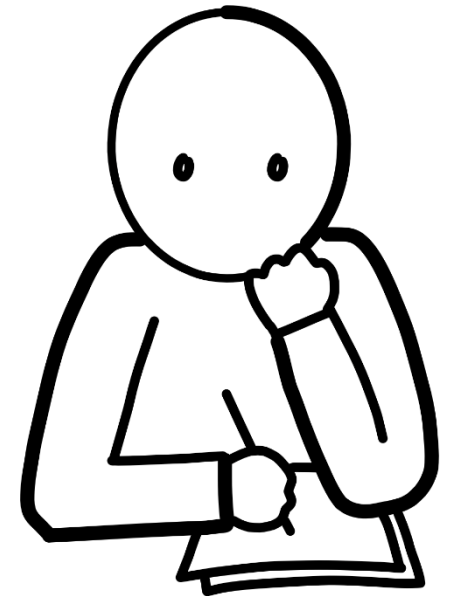
- A DMP structures the handling of research data in a scientific project.
- It describes how to deal with the data used during and after the project.

The DMPs usually cover the following aspects:

1. Description and collection of data
2. Documentation and metadata
3. Storage and backup
4. Legal and ethical requirements
5. Data sharing and long-term storage
6. Responsibilities and resources

Why a DMP?

- Increasingly required by funding organisations (e.g. Horizon Europe)
- Costs for data management (software, hardware, technical expertise) are often eligible for funding
- Making your own handling of research data explicit
- Own project management
- Quality management



Write a DMP

How Do You Write a DMP?

1. Brainstorm yourself
2. Use a template
3. Use digital tools

Tools to Write DMPs

- Argos (<https://argos.openaire.eu>)
- Data Stewardship Wizard (<https://ds-wizard.org>)
- DMPTool (<https://dmptool.org>)
- DMPOnline (<https://dmponline.dcc.ac.uk>)
- RDMO (<https://rdmo.mpdl.mpg.de>)

Use existing DMPs

Templates for DMPs:

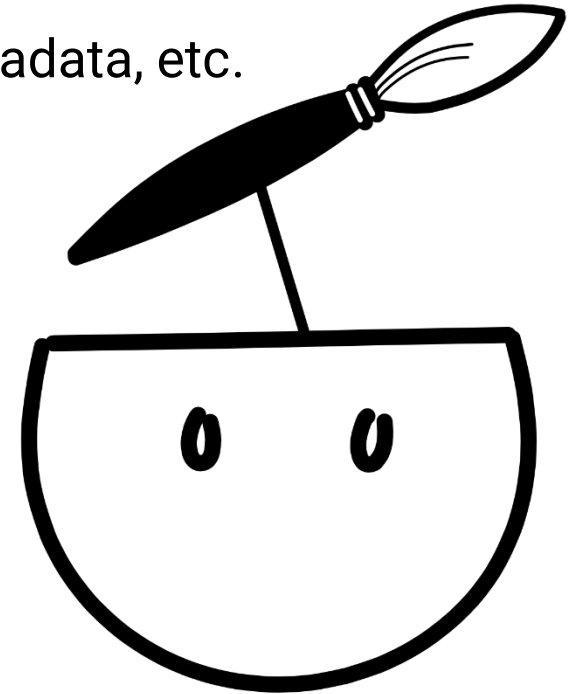
- [Horizon Europe DMP Template by the European Commission, 2021](#)
- [Horizon 2020 DMP Template by the European Commission, 2018](#)
- [ERC Grants DMP Template by the European Research Council, 2021](#)
- [DMP Template by the Swiss National Science Foundation, 2017](#)

Examples of Published DMPs:

- [Digital Curation Centre Example DMPs and Guidance](#)
- [LIBER Europe DMP Catalogue](#)
- [Examples for Horizon 2020 DMPs by the University of Vienna](#)

Get Feedback on DMP

- Ask your own supervisors and colleagues
- Ask science support staff at your own institute
 - e.g. IT regarding storage, library regarding publication and metadata, etc.
- Use centralised DMP service within the MPG
 - e.g. RDM Support by MPDL for cross-checking DMPs
- ...



Questions to You via Slido

Join via slido.com and enter 1174667

1. Have you already written an DMP?
2. If so, for which foundation or funding programm?

Order and Structure

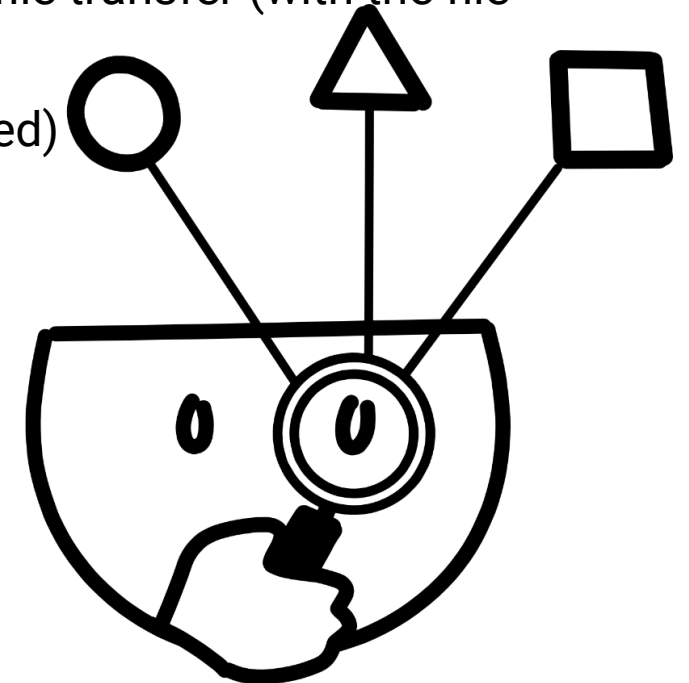
- Directory structures
- File naming
- Version control
- File formats

File Naming – Good Practice

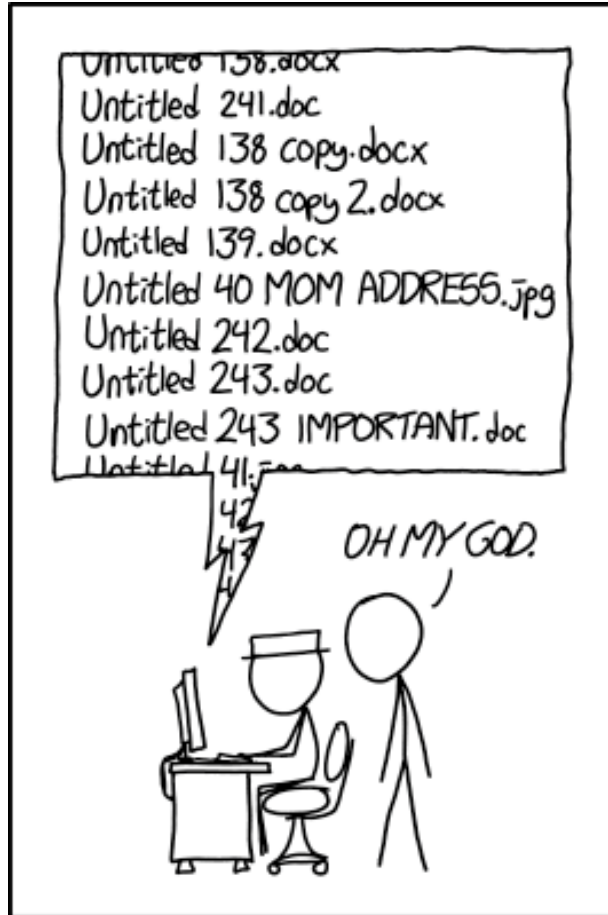
- Include relevant **information** in file names. This may include project, investigator (initials), experiment, location, date, parameter, status of data (raw, processed etc.), file version, or anything else.
- Use YYYY-MM-DD for **dates** (e.g. 2016-08-02) like ISO 8601 and put date at the beginning or end of file names, both to facilitate chronological sorting.
- Use leading **zeros** for other numbers as needed (001, 002, 011, 114 rather than 1, 2, 11, 114) depending on the expected amount of files, and also to allow files to be properly sorted.
- Use **dashes** (file-name.xxx) or **underscores** (file_name.xxx) to separate parts of a filename.
- Generate a **README** file explaining file nomenclature (including the meaning of acronyms or abbreviations), file organization and versioning. Store this file on top of a folder structure for easy access.
- **Good example:** 20140618_exp08_co2data_raw_v01.csv

File Naming – Avoid

- Using **special characters** such as ~ ! / \ @ # \$ % ^ & * () ` ; < > ? , [] { } ' " | as they may have specific meanings in certain operating environments. Also avoid ä ö ü ß å ø ñ or similar.
- Using **dots** other than before the file extension.
- Using **spaces** in file names – they might end up being truncated during file transfer (with the file extension also lost and files potentially becoming unreadable).
- Making file **names** overly **long** (a maximum of 32 characters is suggested) this may involve compromises regarding the recommendation above.
 - In Windows 255 characters is the maximum for a file name (without being in a folder)



File Naming – Examples



PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

xkcd. "Documents.", <https://xkcd.com/1459/>, CC-BY-NC 2.5.

To prefer	To avoid
website-texts-2020-05-v15.docx	Website-Texts-May-finalFinal2_new.docx
Digitized_XY-ZZ_E-2_HT-493887.tiff	Image0001.tiff
Digitized_XY-ZZ_M-5_LS-345-c.tiff	Image0002.tiff
Digitized_XY-ZZ_E-1_M-296778.tiff	Image0003.tiff
Survey_OpenAccess_Clean_ed_v04.csv	OA-Results-clean-4.csv

Examples from <https://www.fu-berlin.de/en/sites/forschungsdatenmanagement/in-der-praxis/durchfuehrung/benennung.html>

File Organization Structure

- Before you start collecting files or data, you should definitely define a **convention** to avoid a backlog of unorganised content.
- In general, for the **naming** of folders, the same rules apply as for files.
- Depending on the structure of your data and documentation, you have to decide on how to arrange your files and folders.
- This may also require judgment and compromises: a very deep folder structure (subsubsubsubsubfolders) can be as inconvenient as one folder with 100 or more individual files.
- In case different project members should have various access restrictions to files, this could be represented by your folder structure.

Example Directory Structures

Project Folder	1. Project Management	<ul style="list-style-type: none">1. Proposals2. Finance3. Reports
	2. Ethics and Governance	<ul style="list-style-type: none">1. Ethical Approvals2. Consent Forms
	3. Experiment 01	<ul style="list-style-type: none">1. Inputs2. Data3. Data Analysis4. Outputs
	4. Dissemination	<ul style="list-style-type: none">1. Presentations2. Publications3. Publicity

After Suse Prejawa (2021). I think I have really good data ... except I cannot find it anywhere: Organising research data in a structured way. Talk presented at Human Research Data in Practice, Virtual Workshop. 2021-04-20, <https://hdl.handle.net/21.11116/0000-0008-662A-7>.

Example Directory Structures

Organized by File type

```
DatasetA.tar.gz
|- Data/
|   |- Processed/
|   |- Raw/
|- Results/
|   |- Figure1.tif
|   |- Figure2.tif
|   |- Models/
```

Organized by Analysis

```
DatasetB.tar.gz
|- Figure1/
|   |- Data/
|   |- Results
|       |- Figure1.tif
|- Figure2/
|   |- Data/
|   |- Results/
|       |- Figure2.tif
```

Example README Template

```
# Title of Dataset

[Access this dataset on XYZ] (Dataset DOI link)

Give a brief summary of dataset contents, contextualized in experimental procedures and results.

## Description of the data and file structure

This is a freeform section for you to describe how the data are structured and how a potential consumer might use them. Be as descriptive as necessary. Keep in mind that users of your data might be new to the field and unfamiliar with common terminology, metrics, etc.

Describe relationships between data files, missing data codes, other abbreviations used. Be as descriptive as possible.

## Sharing/Access information

This is a section for linking to other ways to access the data, and for linking to sources the data is derived from, if any.

Links to other publicly accessible locations of the data:

- [http://...] (http://...)

Data was derived from the following sources:

- []()

## Code/Software

This is an optional, freeform section for describing any code in your submission and the software used to run it.

Describe any scripts, code, or notebooks (e.g., R, Python, Mathematica, MatLab) as well as the software versions (including loaded packages) that you used to run those files. If your repository contains more than one file whose relationship to other scripts is not obvious, provide information about the workflow that you used to run those scripts and notebooks.
```

Brief own work

Question: Take a look at the files and folders in your current project. What problems do you see?

Method: Individual work

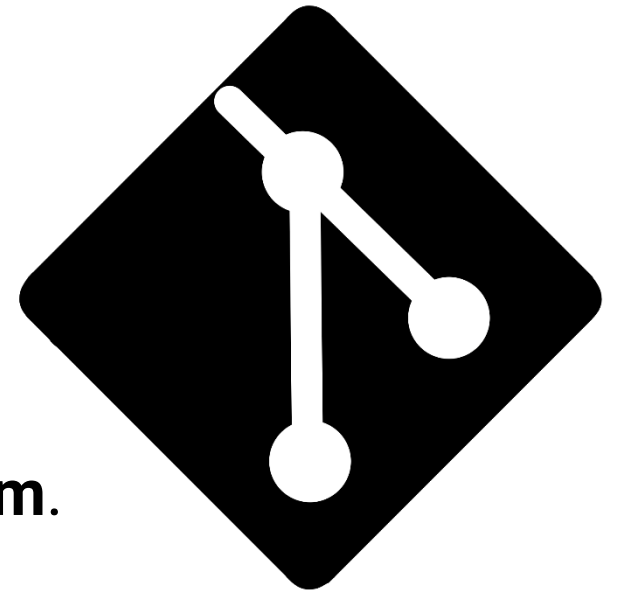
Time: 5 minutes

Securing knowledge: Exchange afterwards within the course

Version Control

Documents may evolve over time, several people may be involved in sequential changes. For data (of any kind, e.g. numeric or images) and text documents alike, file versioning serves two purposes:

- You can **revert** to earlier versions if needed.
- You can **keep track** of changes, including documentation on the underlying rationale and people involved.
- Version control can be done either **manually** by using naming conventions or by using a **version control system**.



Manual Version Control

- When working with a **manual control system**, versions should be numbered consecutively; major changes (v1, v2, v3, ...) can be distinguished from minor ones (v1-1, v1-2, v1-3 or 1a, 1b, 1c). Use leading zeros if you expect more than nine versions.
- However, do not apply such numbering when using version control software, as it will interfere with automatic versioning.
- Qualifiers such as “raw” or “processed” for data, or “draft” or “internal” for documents are useful.
- But terms such as “final2”, “final-revised”, “final-changed_again”, “final_ready” can be confusing. In other words: Avoid “final” naming



<https://doi.org/10.5281/zenodo.3674561>

Version Control Systems

- A prominent example for a **version control system** is Git, as a distributed version-control system for tracking changes. It is widely used in the field of software development. Also a version control for research data with Git is possible as well.
- There may even be a **local** Git instance at your own Max Planck Institute? Ask for it!
- If not, there are a couple of **central** Git systems within the Max Planck Society, which can be used. For example the MPCDF is hosting a [GitLab instance](#). The GWDG is offering a [GitLab system](#) as well.
- A **tutorial** for using Git is given by [Data Carpentry](#) or at [GitHub Education](#).
- An helpful **overview** on tools for version controlling data is given by [The Turing Way](#).



File Formats

Proprietary Data Formats vs. Open Data Formats

- **Proprietary** data formats often are not readable without the corresponding (commercial) software and may become obsolete in the future.
- An Open Data Format is a format with, “a freely available published specification which places no restrictions, monetary or otherwise, upon its use”. (see <https://opendefinition.org/ofd/>)
- **Conversion** from proprietary to open formats is often possible, but may result in some loss of data.



Recommended File Formats I

Type of data	Recommended formats
Text	Plain text (.txt) Rich Text Format (.rtf) Open Document Text (.odt; ISO 26300) PDF (PDF/A-1)
Tabular data (minimal metadata)	Comma-separated values (.csv) Open Document Spreadsheet (.ods; ISO 26300) Tab-delimited (.tab)
Tabular data (extensive metadata)	SPSS portable format (.por) NetCDF HDF5

This list is a guidance and non-binding.

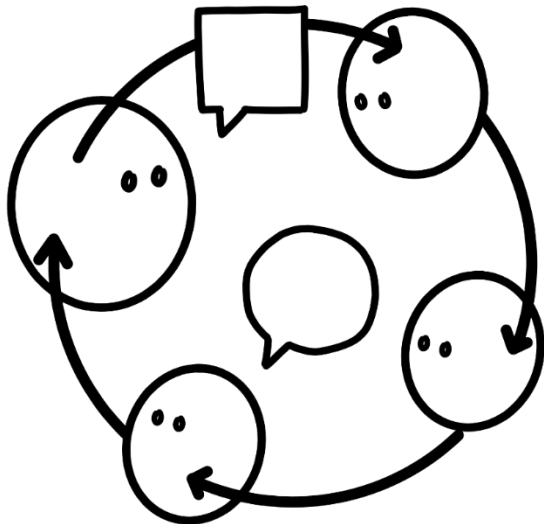
Recommended File Formats II

Type of data	Recommended formats
Images	TIFF6.0 uncompressed (.tif) JPEG2000 (.jp2) JPEG (.jpg) GIF (.gif) Scalable Vector Graphics (.svg) FITS (.fits) PNG (.png)
Video	MPEG-4 (.mp4) OGG video (.ogv, .ogg)
Audio	Free Lossless Audio Codec (.flac) Waveform Audio File Format (.wav)

This list is a guidance and non-binding.

File Standards

- Every standard is better than no standard
- Need to balance: More generic standards vs. community specific standard (i.e. DublinCore vs. DataCite metadata schemas)



<https://doi.org/10.5281/zenodo.3674561>

More Specific File Standards

- Digital Imaging and Communications in Medicine (DICOM, <https://www.dicomstandard.org>), for medical imaging
- FITS (Flexible Image Transport System, https://fits.gsfc.nasa.gov/fits_standard.html), image data file format for astronomical data
- International Image Interoperability Framework (IIIF, <https://iiif.io>), for images i.e. art history
- NeXus Data Format (<https://www.nexusformat.org>), for neutron, x-ray, and muon science

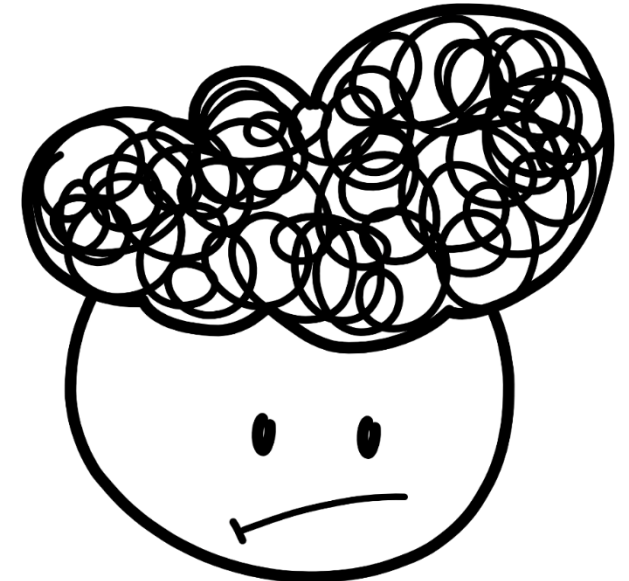
Saving

- Storage and storage locations
- Backup strategies
- Sustainable file formats
- Long-term archiving

Storage and Storage Locations

- Own laptop
- Local institute
- Share systems like ownCloud, Keeper etc.
- ...

- Ask your local IT!



Backups to Ensure your Data

- Backups are an instrument to ensure that your data can be restored in case of damage or losses.
- The data backup strategy is usually already set in your DMP.
- Nevertheless, it is necessary to carry out your planned backup strategy continuously.
- Similarly, the restoring of data, for example, must really be tested for effectiveness. It is your task.
- Keep in mind, that your data is **invaluable**, in comparison to storage space.

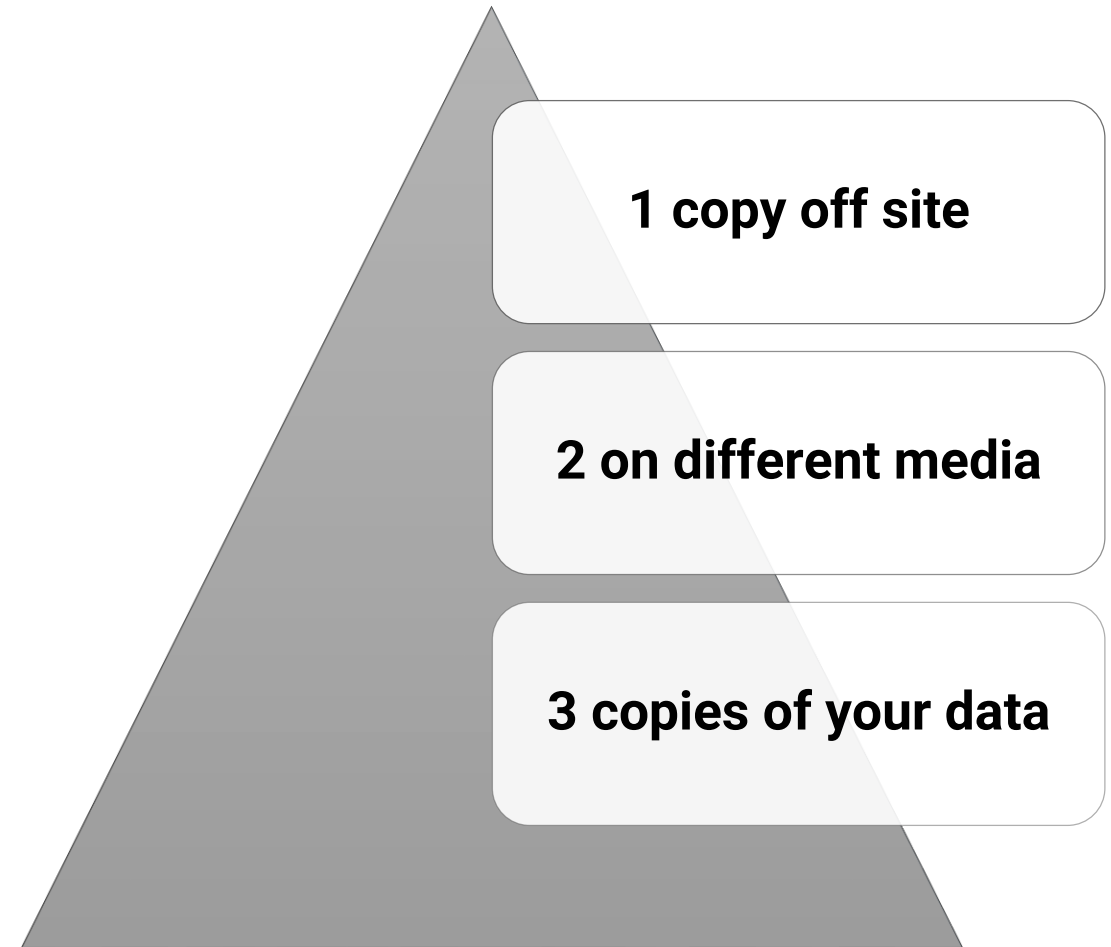
Backup Strategies

3 - 2 - 1 rule

- 3 copies of your data
- On 2 different media
- 1 copy off site

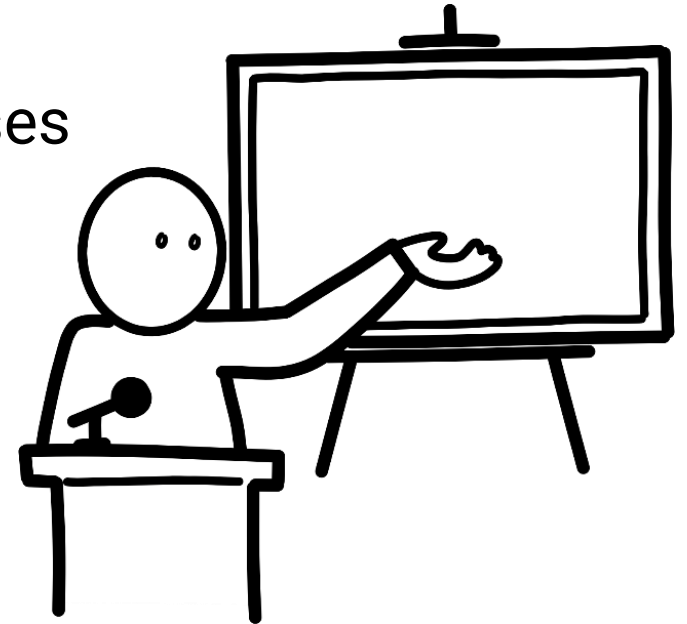
Other concepts:

- [10 steps by CESSDA](#)
- [Checklist by OpenAIRE](#)



Check if it works

- A data backup is only helpful if the recovery of the data is guaranteed
- It is advisable to check the data recovery at the start of the backup and at regular intervals to prevent data loss
- Sometimes files are corrupted and are faulty
- Occasionally the copying process of files itself causes errors



<https://doi.org/10.5281/zenodo.5608845>

Sustainable File Formats

The US Library of Congress defines seven “sustainability factors for digital formats”:

1. **Disclosure:** Do complete specifications and tools for validating technical integrity exist?
2. **Adoption:** A widely used format is less likely to become (rapidly) obsolete.
3. **Transparency:** Are file formats open to direct analysis with basic tools?
4. **Self-documentation:** Are metadata included within the file?
5. **(Limited) external dependencies:** Minimize the degree to which a format depends on specific hardware, software, or operating systems.
6. **Impact of patents:** Patents related to a digital format may inhibit the ability of archival institutions to sustain content in that format.
7. **Technical protection mechanisms:** Encryption poses problems for long-term archival and dissemination, migration to new formats, transfer to new storage media.

<https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml>

Long-Term Archiving

- "Long-term" is an auxiliary term used to describe an unspecified period of time in which technological and socio-cultural changes can occur that can influence the preservation, access to, research in and subsequent use of digital research data.
- Digital long-term archiving therefore comprises a series of measures that need to be planned, controlled and implemented.
- Ten years, as mentioned in the good scientific practice, or even beyond?

Frequent Problems

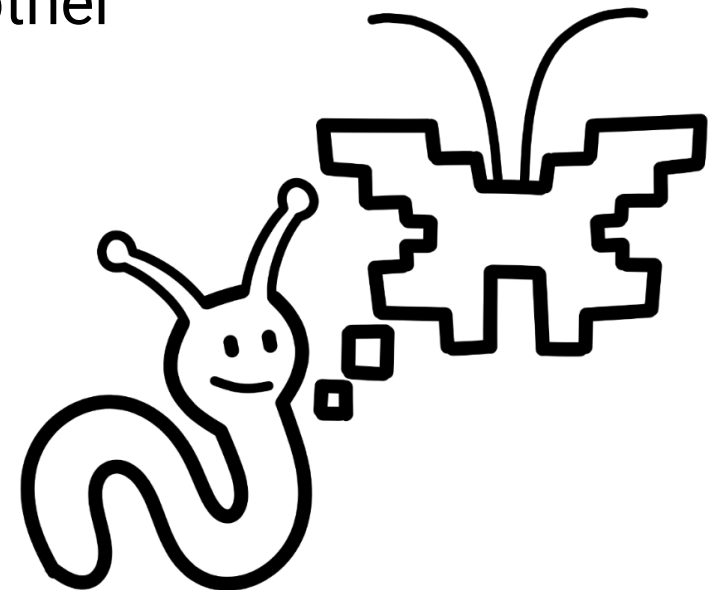
Probleme	Possible solution
Loss of bits	Check the checksums of the data
Delete or overwrite data	Backups and versioning
Defect data storage medium	Backups on external storage
Stolen data carrier	Encryption of your own data and backups on external storage media
Fire, flooding, etc.	Backups on external storage
Loss of compatibility of file formats	Use suitable, non-proprietary file formats (additionally)
Virus and Trojan	Backups on external storage and IT security concept of its own MPI

Secure!

- Encryption
- Checking the usability
- Password protection and access rights

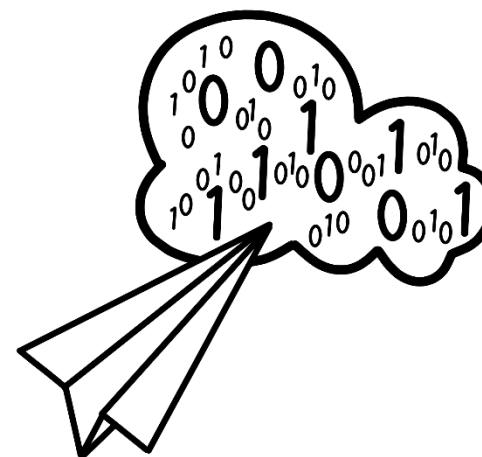
Secure Research Data

- Research data are among the most valuable resources in science
- That is why there is great importance to their security
- The security and access rights aspects of data management should consist of measures to protect against data loss on the one hand and measures to prevent misuse of the data on the other



Encryption

- Physical access to a computer can enable unauthorised access to data, so it may be necessary to encrypt the data
- This measure only makes sense if all affected data is encrypted, especially copies and backups
- All involved parties must be aware of the need for encryption and the storage locations



Password Protection and Access Rights

- Working with sensitive data, secure passwords should be assigned and access be restricted to the circle of persons directly involved.
- The assignment of authorisations determines which persons or groups of persons are allowed to access certain directories and files and with which rights.
- It is possible to assign graduated read and write authorisations as well as execution rights → some users can only view data, while others are granted full access
- It is important to assign permissions carefully not to hinder the workflow

Bundesamt für Sicherheit in der Informationstechnik

BSI-Basisschutz: Sichere Passwörter

Passwörter begleiten uns täglich und trotzdem oder gerade deshalb greifen viele Menschen bei der Wahl ihrer Passwörter auf einfache Zahlenabfolgen oder Namen und Orte in Kombination mit Zahlen oder Sonderzeichen zurück. Diese sind zwar leicht zu merken, können aber ebenso leicht von Cyber-Kriminellen geknackt werden.

Bei einem Cyber-Angriff sind nicht nur persönliche Daten und sensible Informationen in Gefahr, Cyber-Kriminelle können die gehackten Accounts auch für kriminelle Machenschaften und illegale Geschäfte nutzen. Um das zu verhindern, sollte ein Passwort bestimmte Anforderungen erfüllen und immer nur für einen Zugang genutzt werden.

Grundsätzlich können Sie zwei Strategien anwenden, um ein sicheres Passwort zu erstellen:

Weitere Informationen:
<https://www.bsi.bund.de/dok/6596574>

Sicheres Passwort	
Kurzes, dafür komplexes Passwort <ul style="list-style-type: none">• Ist acht bis zwölf Zeichen lang.• Besteht aus vier verschiedenen Zeichenarten.• Groß- und Kleinbuchstaben, Zahlen und Sonderzeichen werden willkürlich aneinandergereiht.	Langes, dafür weniger komplexes Passwort <ul style="list-style-type: none">• Ist mindestens 25 Zeichen lang.• Besteht aus zwei Zeichenarten.• Kann zum Beispiel aus sechs aufeinanderfolgenden Wörtern bestehen, die jeweils durch ein Zeichen voneinander getrennt sind.

Um ihre Accounts und Daten zu schützen, sollten Sie außerdem folgende Tipps beherzigen:

Generell gilt	Zu vermeiden
<ul style="list-style-type: none">✓ Ein individuelles Passwort pro Account!✓ Eine Mehr-Faktor-Authentisierung (ergänzend zum Passwort durch bspw. eine Gesichtserkennung, eine App-Bestätigung, E-Mail oder einer PIN auf einem anderen Gerät) ist empfehlenswert.✓ Alle verfügbaren Zeichen nutzen inklusive Groß- und Kleinbuchstaben, Ziffern und Sonderzeichen (Leerzeichen, !%*...).✓ Das vollständige Passwort sollte nicht im Wörterbuch vorkommen.	<ul style="list-style-type: none">✗ Namen von Familienmitgliedern, Haustieren, Geburtsdaten etc.✗ Einfache oder bekannte Wiederholungs- bzw. Tastaturmuster wie „asdfgh“ oder „1234abcd“✗ Ziffern oder Sonderzeichen an den Anfang oder ans Ende eines ansonsten einfachen Passwortes.✗ Dasselbe Passwort bei mehr als einem Account.

Bundesamt für Sicherheit in der Informationstechnik,
https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Checklisten/sichere_passwoerter_faktenblatt.pdf

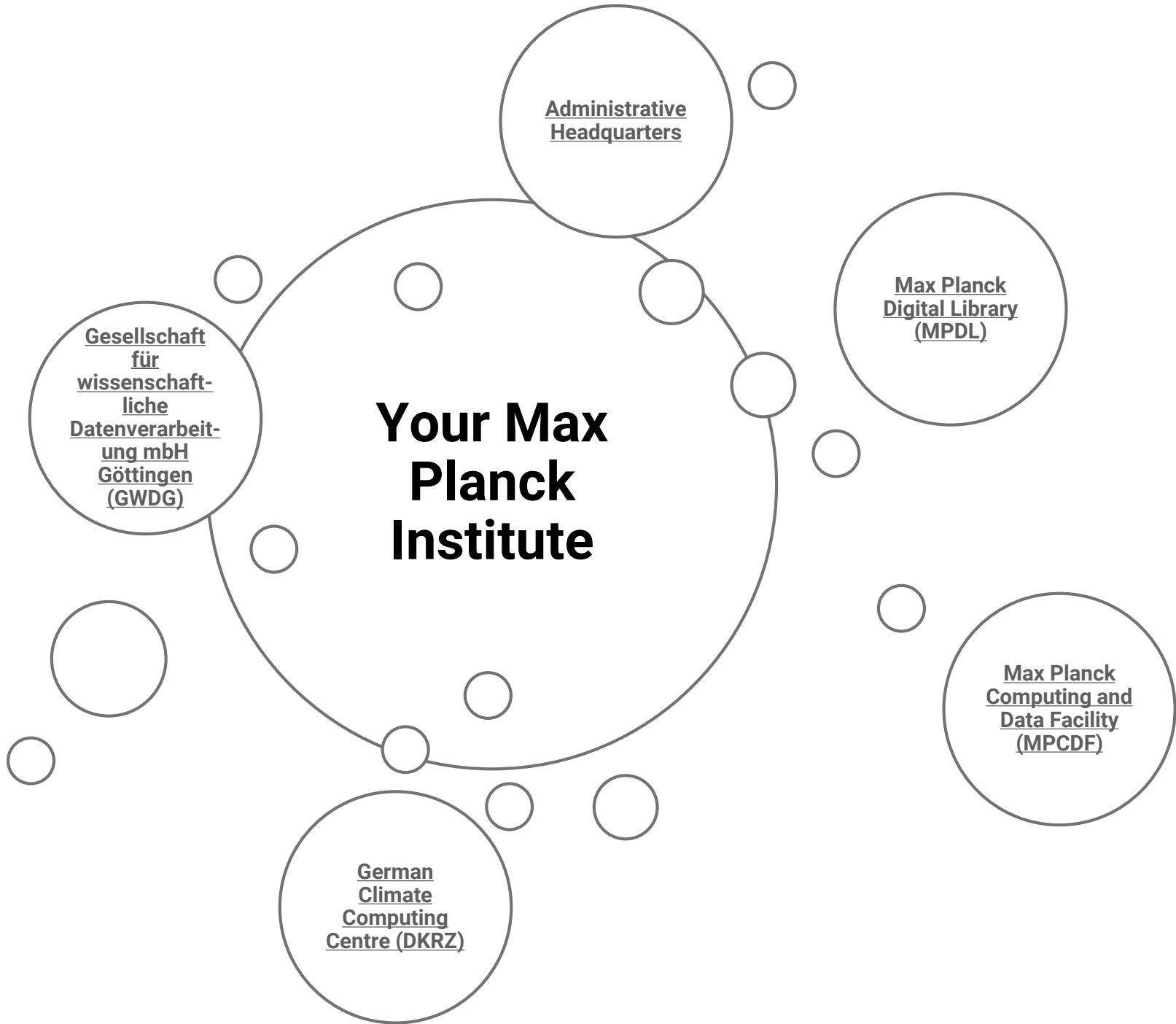
Lunch break until 1 pm

Max Planck Data Services

- MPG Data Services: From local via central to external
- Data services: MPCDF, GWDG, and MPDL

Which RDM Services Are Best To Use?

- It depends!
- Most important thing is to understand which data are generated and used
- Recommendation: A conscious approach to the management of data
- Procedure:
 1. Write a small specification with all the requirements for data and management
 2. Search for different service providers
 3. Choose the best for your research and in accordance with institute's practice
- The fastest first solution does not always have to be the best one



Administrative
Headquarters

Max Planck
Digital Library
(MPDL)

Max Planck
Computing and
Data Facility
(MPCDF)

German
Climate
Computing
Centre (DKRZ)

Gesellschaft
für
wissenschaft-
liche
Datenverarbeit-
ung mbH
Göttingen
(GWDG)

**Your Max
Planck
Institute**

More Max Planck Data Services

1. Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)
2. Max Planck Computing and Data Facility (MPCDF)
3. Max Planck Digital Library (MPDL)

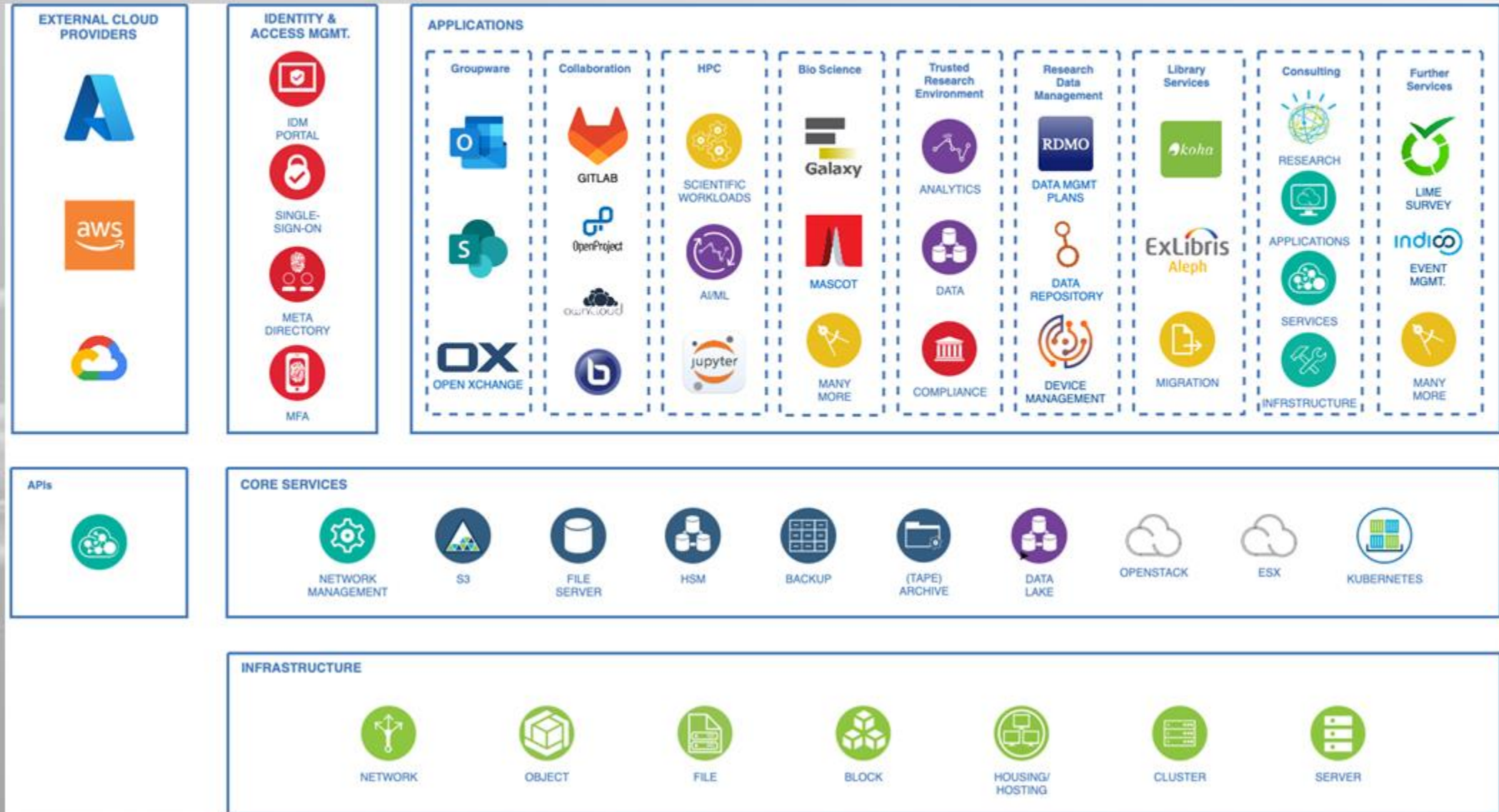
Data Services of the GWGD



The GWWDG

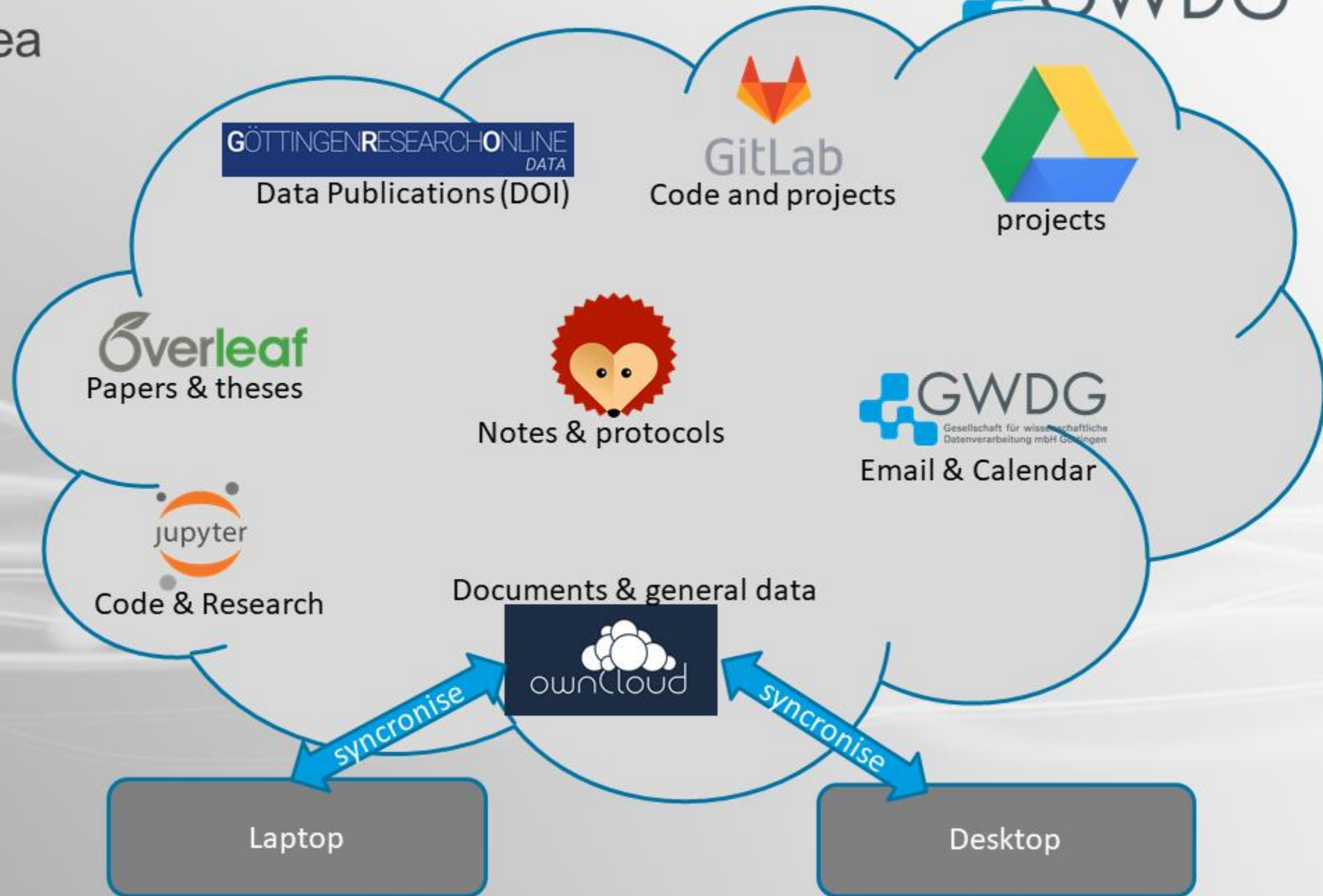
- Founded 1970
- Stakeholder
 - Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V.
 - Georg-August-Universität Göttingen – Stiftung öffentlichen Rechts
- Mission
 - Computing and IT competence centre for the Max Planck Society
 - University Computer Centre for the Georg-August-Universität Göttingen

GWDG Service-Architecture

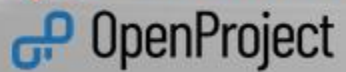


* Only selected applications shown

Daily Work Area



Communication & Collaboration



File Services

- File-based storage of data via standard protocols (NFS, CIFS, CephFS)
- Different features available depending on the platform:
 - Snapshots
 - Sync. Mirroring
 - Storage efficiency process (Dedup, Compression)
 - HSM (transparent storage on tape)
- Different Performance Tiers: Cold-, Warm-, Hot-Storage

The NetApp logo, featuring a blue square icon with a white 'N' shape inside, followed by the text "NetApp" in a bold black sans-serif font.The StorNext logo, with the text "StorNext" in a blue sans-serif font and a registered trademark symbol.The Ceph logo, featuring a red circular icon with a white stylized 'C' shape inside, followed by the text "ceph" in a grey sans-serif font.

ownCloud/Nextcloud

- File synchronisation service with Office collaboration function
- Central instance based on the ownCloud product
- Hosting of Nextcloud instances for individual institutes or cross-institute projects/groups (on request)



Cryptshare

- share files with your web browser
- data will be transferred in encrypted form and
- stored encrypted on the server
- no additional user account necessary
- user-friendly web interface
- Password protected download



S3 Object Storage

- Amazon S3 compatible object storage
- Worldwide access to data via HTTPS
- Rapid growth in use cases (data lakes, Kubernetes, cloud-native architectures)
- Cost-effective and scalable
- Different performance tiers: cold / hot storage
- Support for advanced features: Bucket-Policies, Lifecycle-Policies, Encryption, Online-Compression, Replication, S3-Select, etc.



Backup auf Tape (TSM/ISP)

- Use of TSM / ISP as cost-effective backup software
- Focus on security (ransomware) and cost efficiency
- MPI-specific setups, e.g. :
 - Replication of backups to the MPCDF (geo-redundant second copy)
 - Backups to HDDs for fast restores



**IBM
Spectrum Protect**

Veeam Cloud Connect

- Backup-as-a-Service
- Simple integration into different infrastructures
- Georedundant backups

- Fulfilment of compliance requirements
- Very favourable data storage
- Encryption of the backups

- Tape integration in development

The Veeam logo, featuring the word "VEEAM" in a bold, green, rounded sans-serif font.

Long-term archiving (archive server)

- Secure and cost-effective long-term archiving of large amounts of data (10+ years)
- Storage of two copies in dedicated, geo-redundant tape libraries
- SSD-based file system (150+ TB) for easy access to the archive
- Flexible access via NFS, SMB or SCP / SFTP

The logo for StorNext, featuring the word "StorNext" in a bold, blue, sans-serif font with a registered trademark symbol (®) to the right.

Scientific Compute Cluster - Supplement to the MPCDF offer

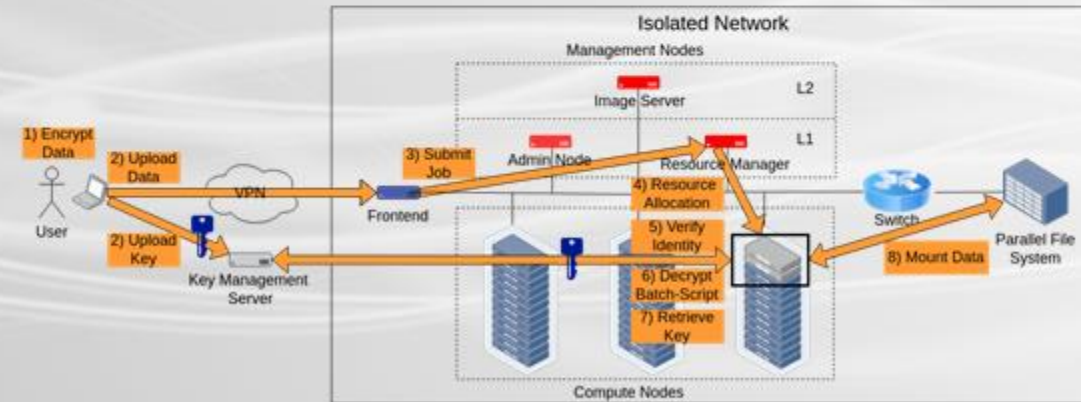
- Close integration with other GWDG services
 - Platform for various BioScience applications
 - Private cloud - enables front ends for scalable HPC applications
 - Data sharing and publication services
- Offers
 - Test environments - stepping stone to MPCDF and national offers
 - Exploration of alternative hardware
 - Future Technology Platform (Graphcore, ARM, RISC-V, ...)
 - Training offer (also locally at MPIs)
 - Maximum data protection (protection class D) and ISO27001 certification

Data Lakes

- Support for scientific projects and institutes with customised solutions

- Specially secured services for processing sensitive data (e.g. health data)

- Highly scalable mass storage
- Connection to HPC service
- Indexing of (sensitive) metadata

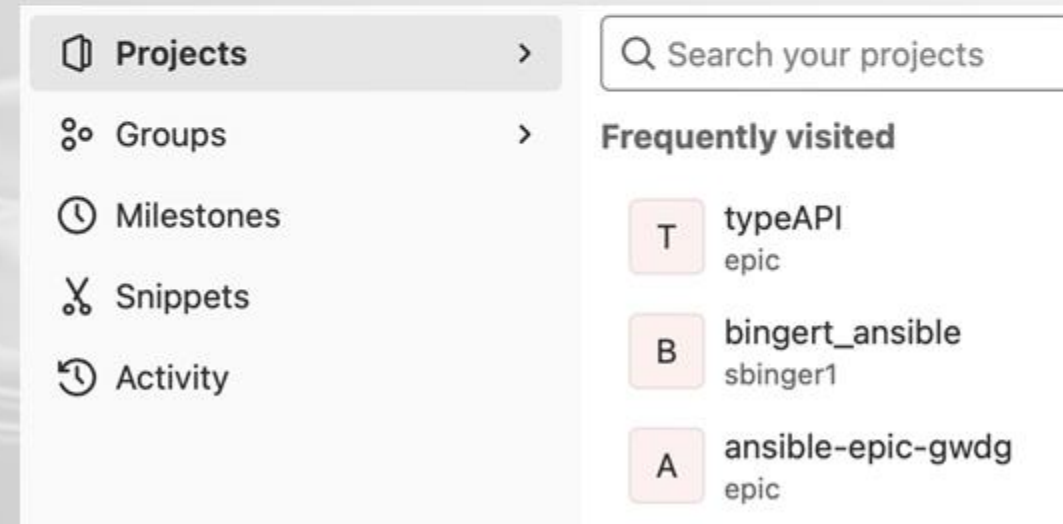


Version Control with GIT

- GWDG Gitlab
 - start a private repo
 - invite colleagues
 - use continuous integration CI/CD



GitLab



<https://jupyter-cloud.gwdg.de>
<https://jupyter-hpc.gwdg.de>

- Python, R, Haskell, Julia ... environment to be used in the browser
- support interactive data science and scientific computing across all programming languages
- create and share documents that contain live code, equations, visualizations and narrative text



Persistent Identifiers

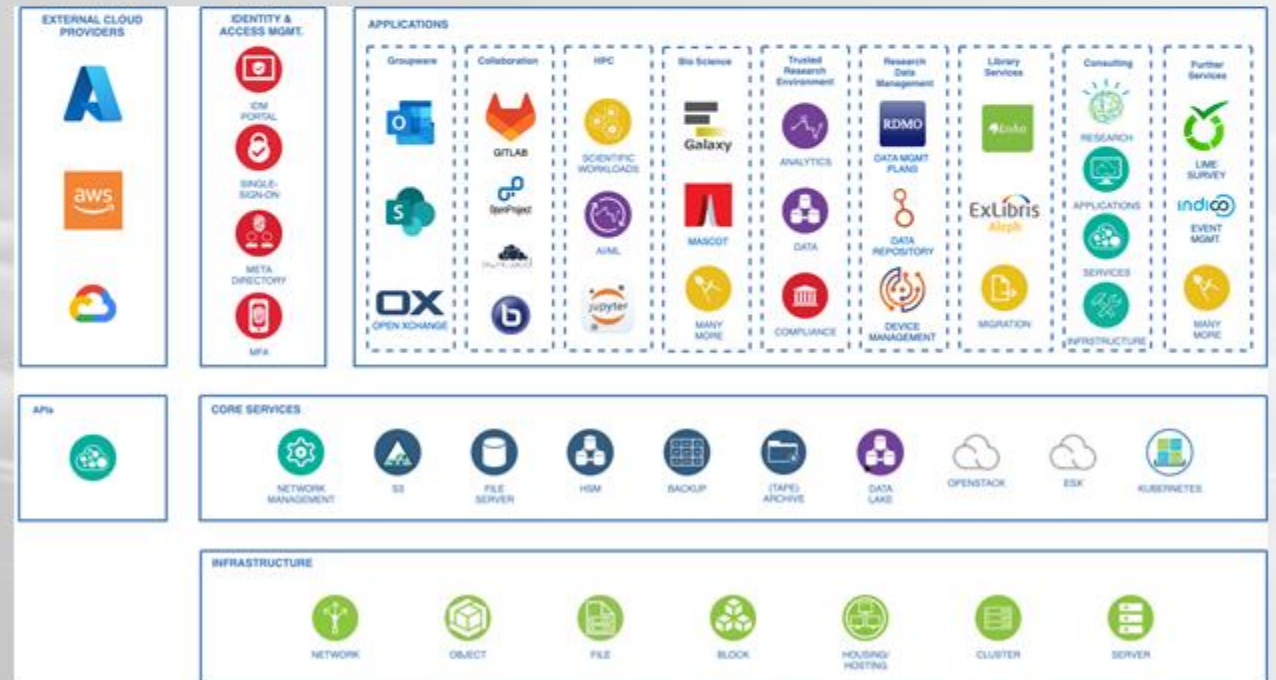
- Elementary component of the FAIR criteria
- Supports the data management processes
- PIDs based on the Handle System
- GWWDG is MPA on DONA (namespace 21.)
- Following the ePIC policies for namespace registration
 - allows also for test namespaces
- Fast provisioning time



How to connect

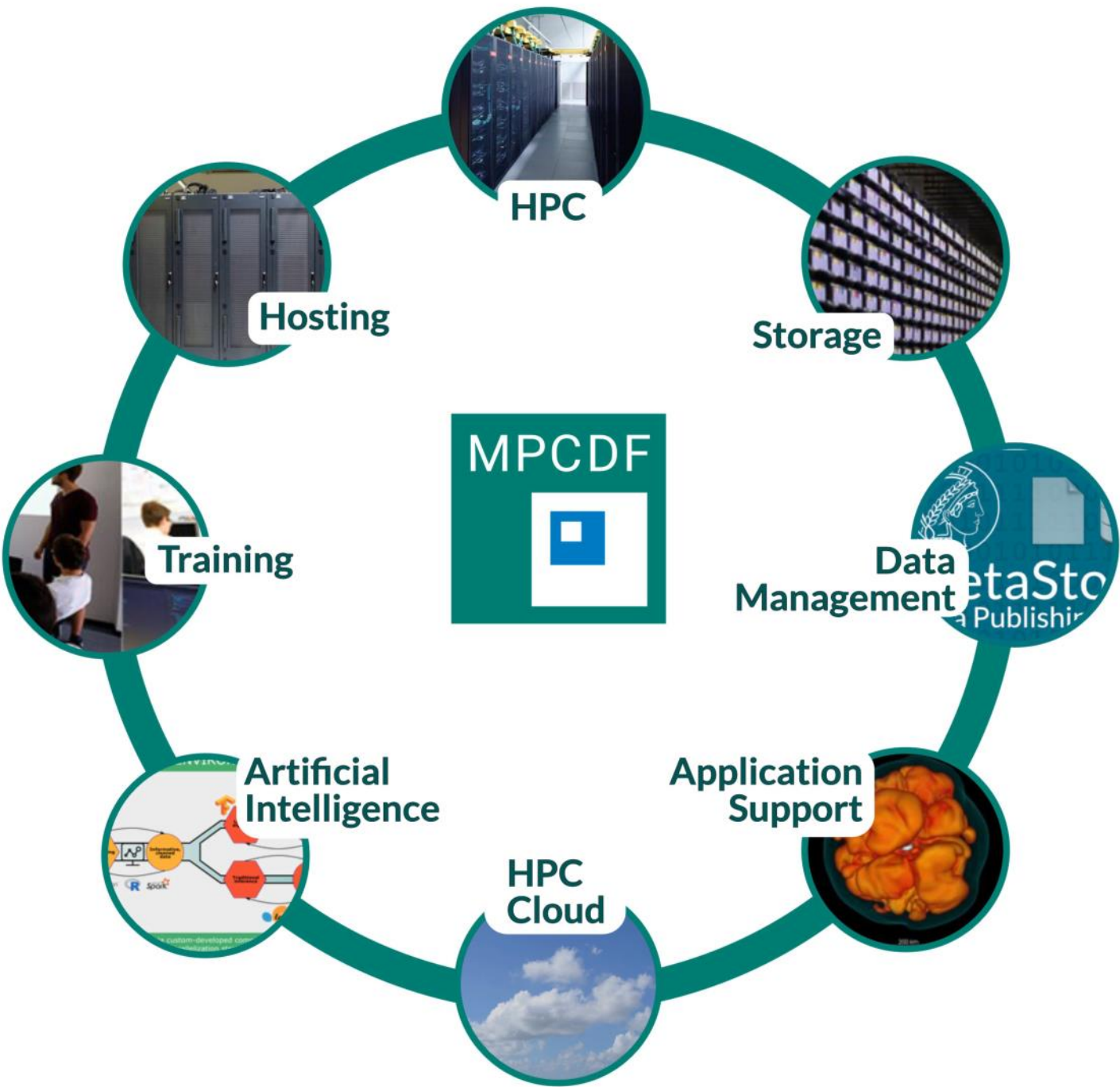
Visit: www.gwdg.de

Requests: support@gwdg.de





MAX PLANCK
COMPUTING & DATA FACILITY



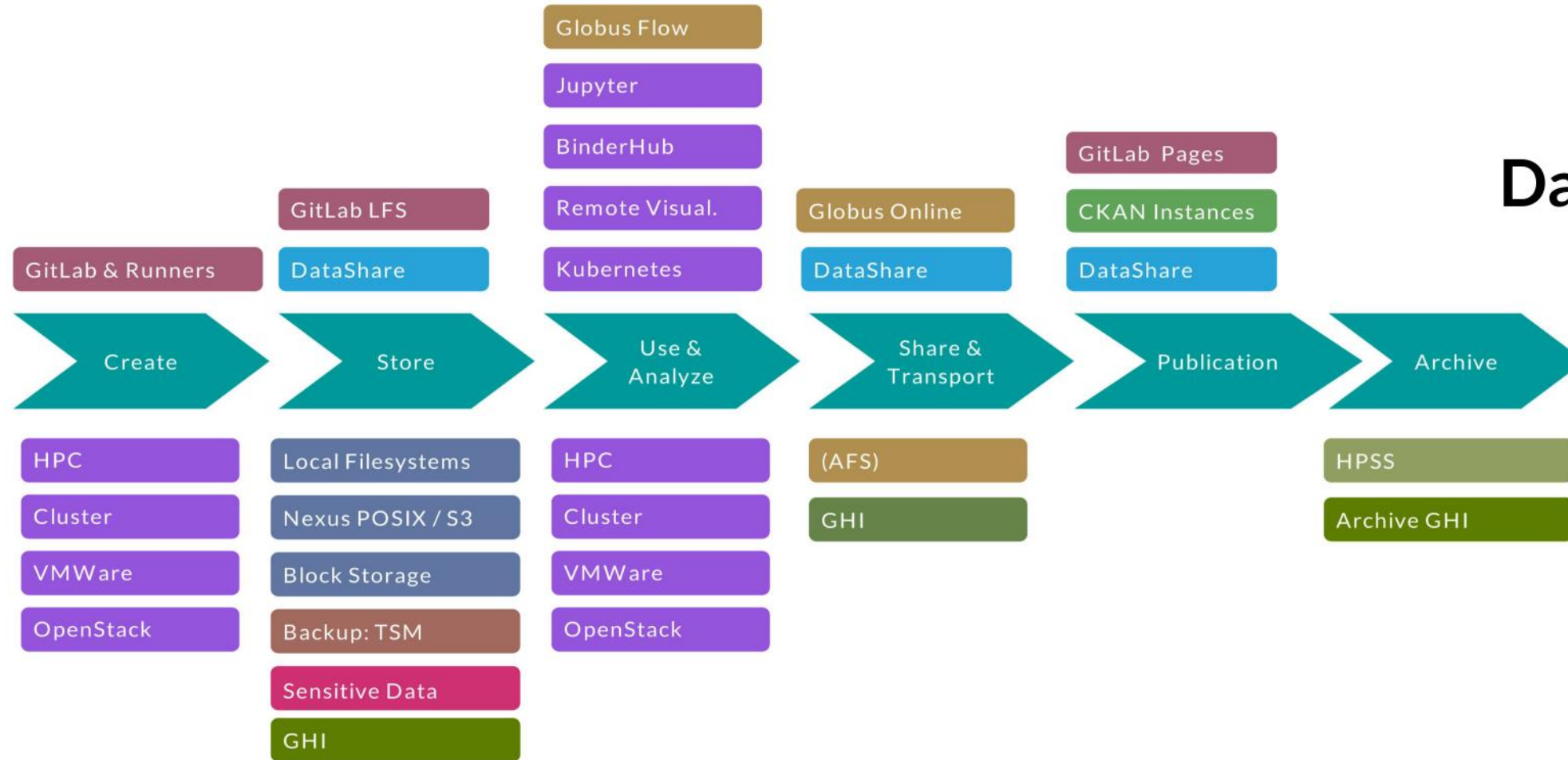
Overview of the MPCDF



AAI

Metadata Management

Along the Data Lifecycle



Network Capabilities

Some MPCDF Data Services

GitLab

The MPCDF GitLab instance is available to all MPCDF users and their external collaborators. Continuous Integration can be done on central hosted shared GitLab Runners, building a bridge to the MPCDF HPC clusters

HPC Cloud

The HPC Cloud enables scientists to combine batch and cloud-based computing within the same pipeline, taking advantage of both massive HPC cluster resources and highly flexible software environments. The system offers standard cloud computing “building blocks”, including virtual machines based on common Linux operating systems, software-defined networks, routers, firewalls, and load balancers, as well as integrated block and S3-compatible object storage services.

CKAN Instances

For data publishing, the MPCDF offers hosting of CKAN-based data repositories. CKAN is a software framework which allows to manage metadata as well as object data. Beside a web-based interface, CKAN offers a REST API for automation of common workflows.

CKAN instances at the MPCDF are meant for Max Planck Institutes, groups or projects and not for individual users.

Globus Online

Globus Online is a third party transfer service which enables fire-and-forget data transfer at TB or multi-TB scale. Globus Online is well established and widely used with many computing centres and research institutes, as well as Universities, having Globus Services installed for their users.

Further Information



Homepage:

www.mpcdf.mpg.de

Email:

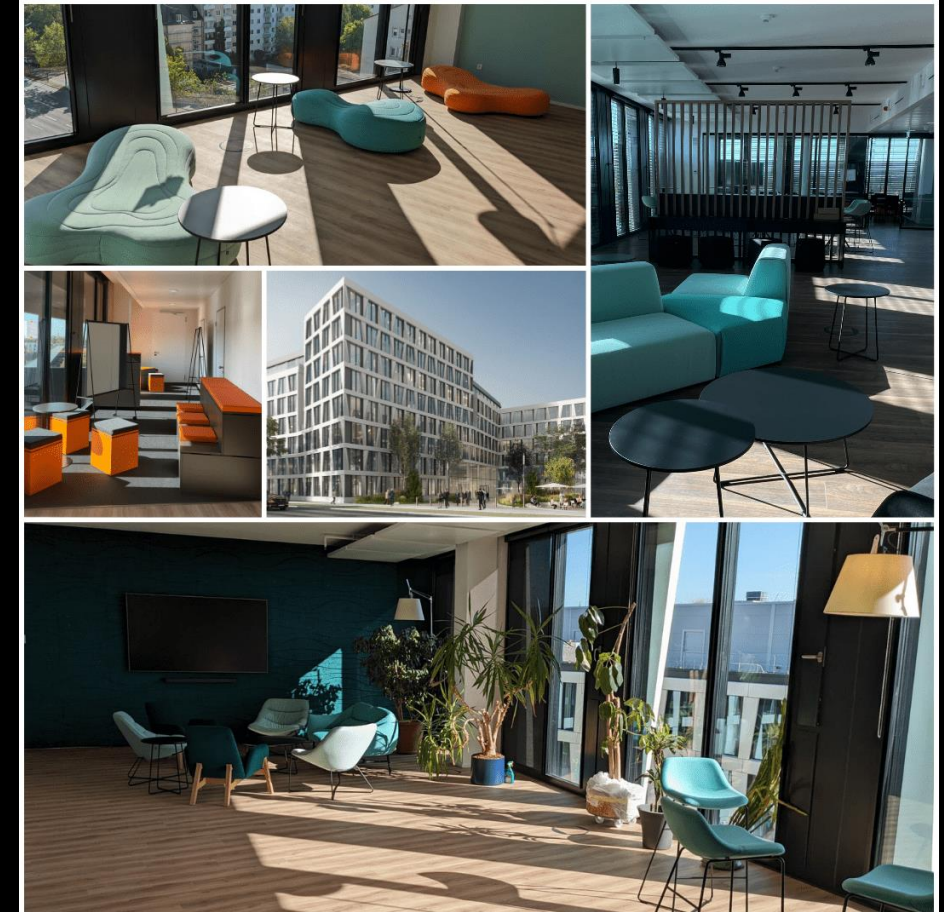
support@mpcdf.mpg.de

Max Planck Digital Library



MAX
PLANCK
DIGITAL
LIBRARY

- Landsberger Straße 346, 80687 Munich
- Information services since 2007, independent service unit since 2012
- MPDL sees itself as
 - one of the four central IT service centres of the Max Planck Society (together with MPCDF, IKT and GWDG)
 - together with the institute libraries, MPDL forms the library system of the Max Planck Society
- Around 80 employees from software development, library, science management and administration
- www.mpdل.mpg.de
- Collections department for the topic of research data



RDM Support



MAX
PLANCK
DIGITAL
LIBRARY

- Consultancy for data management plans
- Support with data transformation
- Handling of metadata
- Identification of suitable data repositories
- Searching for the right repository/archiving software
- and much more
- Contact us via rdm@mpdl.mpg.de

RDM
Support

RDM Information Platform



MAX
PLANCK
DIGITAL
LIBRARY

- Introduction
- Before Research
- During Research
- After Research
- MPDL Services
- News
- Contact
- <https://rdm.mpdl.mpg.de>

The screenshot shows the homepage of the Research Data Management Information Platform. The header features the title "Research Data Management" and the subtitle "Information Platform for Max Planck Researchers - by MPDL". A navigation menu includes links for "Introduction", "Before Research", "During Research", "After Research", "MPDL Services", "News", and "Contact". The main content area has a teal background with a lighthouse illustration. The "Introduction" section explains that digital data are the basis of scientific research and that the website provides data management guidance throughout the research process, introducing Max Planck Digital Library (MPDL) services to support researchers of the Max Planck Society. It also states that the website is a guide for initial orientation and has used a simplified representation for clarity and quick, easy readability.

Edmond

- Open Research Data Repository of the Max Planck Society
- Serves the publication of research data from all disciplines
- Offers scientists the ability to create citable research objects
- Enables secure preservation: provision, describing, documentation, linking, publishing and archiving of all kinds of data
- Supports standardized metadata profiles



MAX
PLANCK
DIGITAL
LIBRARY



<https://edmond.mpdl.mpg.de>

edmond@mpdl.mpg.de





RDMO for MPG

- RDMO stands for Research Data Management Organiser
- You can use it to write Data Management Plans (DMPs)
- You can also organise the data management for a project, e.g. a dissertation
- Collaboration with other users (also external to the MPG)
- Versioning
- Export and import functions

RDMO
for
MPG

<https://rdmo.mpdl.mpg.de>
rdmo@mpdl.mpg.de

RDMO for MPG



MAX
PLANCK
DIGITAL
LIBRARY

Special Catalogues:

- Horizon Europe
- VW-Stiftung – Science Europe
- DFG
- Software Management
- NFDI4Ing
- ...

RDMO for MPG Management Admin

Writing a New DMP

Description	This is a data management plan for demonstration purposes.	
Catalog	Horizon Europe This catalogue is for creating data management plans for projects in Horizon Europe and at the European Research Council.	

Tasks

Tasks are generated automatically from the answers given in the project. On the page of each task you can see which of your answers lead to the activation of the task.

No active tasks found.

Views

Views are created using the answers given in the project and can then be exported in various formats. Initially, all views are empty. Please answer some questions by visiting [Answer Questions](#) (at the top of the sidebar).

View	Description	
Cost overview	Overview of the personnel and non_personnel costs	
Cost Overview	Overview of the personnel and non-personnel costs	
DMPonline template	Template from DMPonline, online https://dmponline.dcc.ac.uk	
DMPTool template	Template from DMPTool, based on "NSF-GEN: Generic", online: https://dmptool.org	
Horizon 2020 FAIR Data Management Plan template	Template for Horizon 2020, from "Guidelines on FAIR Data Management in Horizon 2020", online: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf	

Keeper



MAX
PLANCK
DIGITAL
LIBRARY

- 1 TB storage space
- Long-term archiving of the completed data set, 10 years
- Synchronises local projects
- File sharing with individuals and groups
- Cared Data certificate that confirms good scientific practice
- Digital object identifier (DOI)



<https://keeper.mpdl.mpg.de>
keeper@mpdl.mpg.de

Labfolder



MAX
PLANCK
DIGITAL
LIBRARY

- Electronic laboratory notebook (ELN) for recording, integrating and managing experimental research data
- Two utilization scenarios:
 - In-house service at the respective institute
 - Central MPG Labfolder installation

< LAB

FOLDER >

<https://labfolder.mpdl.mpg.de>
labfolder@mpdl.mpg.de

Software Licensing Service



MAX
PLANCK
DIGITAL
LIBRARY

- Basic software supply
- Open source software solutions
- Preparation of an offer for support and maintenance contracts
- Applications for funding for open source software development projects



SOFTWARE LICENSING
SERVICE

www.soli.mpdl.mpg.de

soli@mpdl.mpg.de

DOI Service



MAX
PLANCK
DIGITAL
LIBRARY

- Assign DOIs to scientific output, e.g. MPG.PuRe, Edmond and Keeper
- Self-service within MPG-IP
- Use for local data repositories possible (MoU needed)



<https://doi.mpdl.mpg.de>
doi@mpdl.mpg.de

Metadata

- Metadata standards
- Identifiers

Metadata Standards

Rich metadata is good, using standards is even better

Example: [Virtual observatory](#)

Find your standard e.g. at [bartoc.org](#), [University of Bath](#)

Identifiers

Digital Object Identifier (DOI)

- For digital objects
- <https://www.doi.org>



Public domain

Open Researcher and Contributor ID (ORCID)

- For persons
- <https://orcid.org>



Public domain, CC0

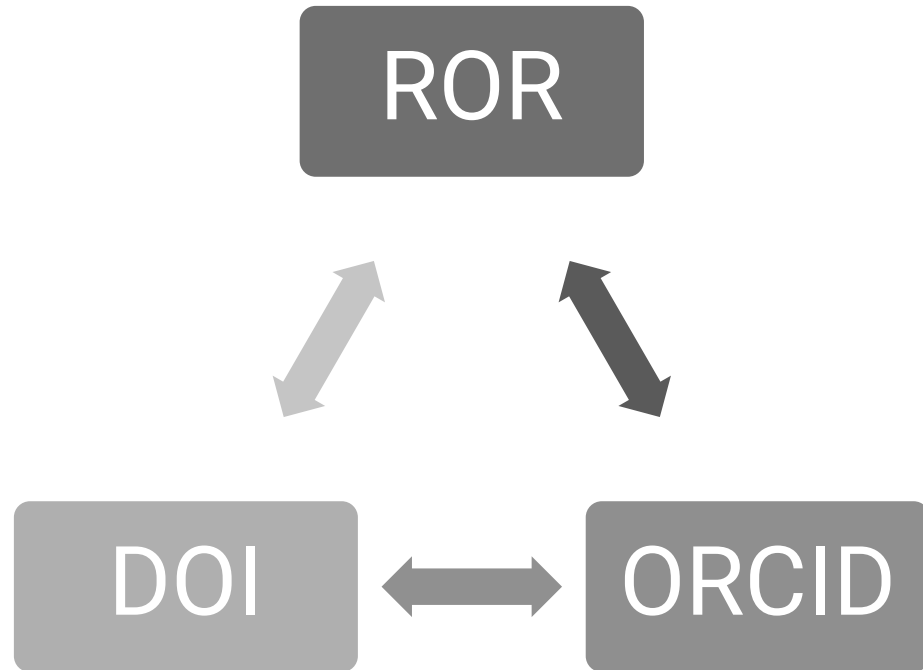
Research Organization Registry (ROR)

- For organisations
- <https://ror.org>



Research Organization Registry, CC BY 4.0,
<https://doi.org/10.5281/zenodo.4701802>

Identifier Network



Max Planck Institute for Evolutionary Biology:

- <https://ror.org/0534re684>
- <https://commons.datacite.org/ror.org/0534re684>

The screenshot shows the DataCite Commons profile for the Max Planck Institute for Evolutionary Biology. The profile includes the following information:

- Organization:** Max Planck Institute for Evolutionary Biology (<https://ror.org/0534re684>)
- Statistics:** 136 Works, 108 Citations, 30,482 Views, 5,374 Downloads
- Other Identifiers:** GRID grid.419520.b, ISNI 0000000122247098, Wikidata Q1493545
- Work Types:** Dataset (88%), CC0-1.0 (75%)
- Publication Year:** A bar chart showing the distribution of works from 2016 to 2023.
- Creators & Contributors:** A list of individuals associated with the organization, including Linnebrink, Miriam (7), Zhang, Boyu (6), Reeves, Richard (3), Pallares, Laisa F (2), Uecker, Hildegard (2), Tautz, Diethard (2), Frenken-Kröte, Carsten (2), Glaino, Stefano (2), Krehenwinkler, Henrik (1), and Hellig, Christian (1).
- Publication Year:** A list of years from 2018 to 2023, with corresponding work counts: 2023 (31), 2022 (10), 2021 (2), 2020 (6), 2019 (12), and 2018 (7).

Questions to You via Slido

Join via slido.com and enter 1174667

1. Do you know your ORCID?
2. Do you know the ROR of your institute?

Publishing

- Data (and software) as a new genre of publication
- Legal aspects of publishing
- Data repositories and journals
- Publish or share data (Open Research Data)
- Reuse data

Three Pillars of Research Results?

Text

- Journal articles
- Books
- Posters
- ...

Data

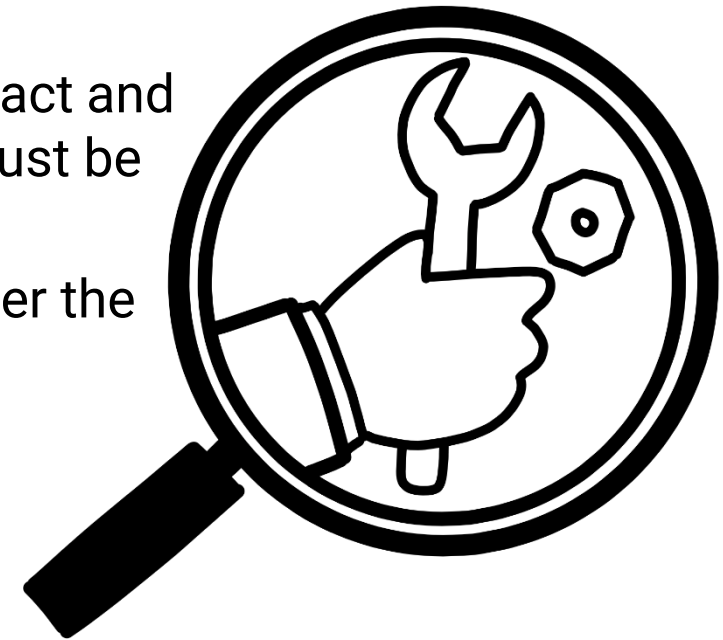
- Raw data
- Lab data
- Analysed data
- ...
-

Code

- Software
- Configuration
- Documentation
- ...

Legal Aspects of Data Publishing

- Who owns the data?
 - The scientist, the supervisor, the Max Planck Institute, the Max Planck Society?
 - There is no universally valid statement on this!
 - Talk to your supervisor. Have a look at your employment contract and local research guidelines. The publication approval process must be clarified.
 - Find out more! It's better to clarify things earlier rather than after the data publication (= too late).



<https://doi.org/10.5281/zenodo.3674561>

Legal Aspects of Data Publishing

- What type of data aggregation should actually be published?
 - That highly depends!
 - Raw data, (pre-)processed data and analysed data are published. It always depends on the context, the specialised culture and the objectives.
 - Discuss this with your supervisor and team colleagues at an early stage.



Licenses Data Publishing

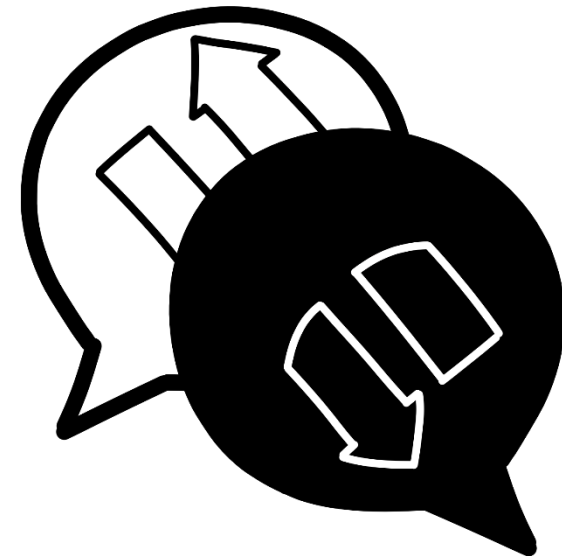
- A licence for a data publication can only be granted by the owner (see above)

Common (Data) License:

- Creative Commons
(<https://creativecommons.org>)
 - CC 0
 - CC BY
- Open Data Commons
(<https://opendatacommons.org>)
 - ODC-By
 - PDDL

Recommendation:

- 1. CC0 or 2. CC BY
- Any license is better than no license!



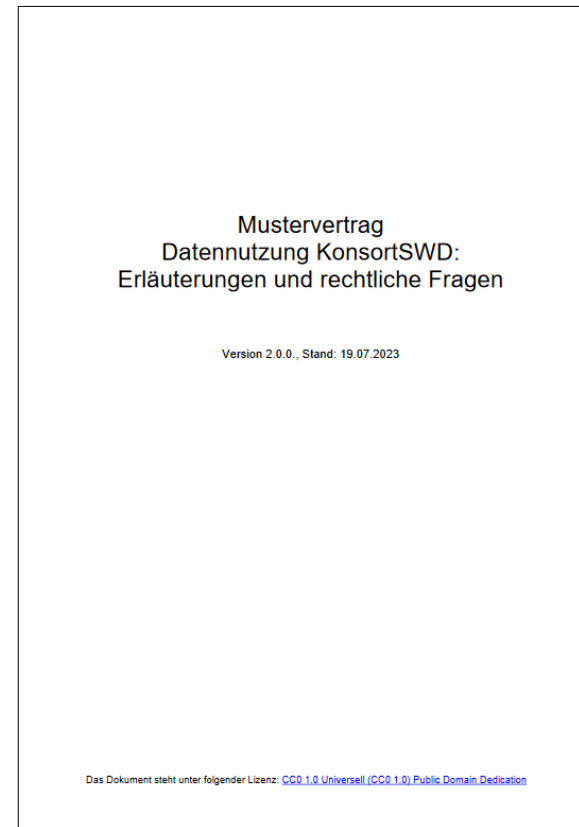
Contracts for Data Collection and Submission

Sample contract for data recording



Schallaböck, J., Kreutzer, T., Hoffstätter, U., & Buck, D. (2023). Mustervertrag Datenaufnahme KonsortSWD, Zenodo, CC BY 4.0, <https://doi.org/10.5281/zenodo.10406480>.

Mustervertrag Datennutzung



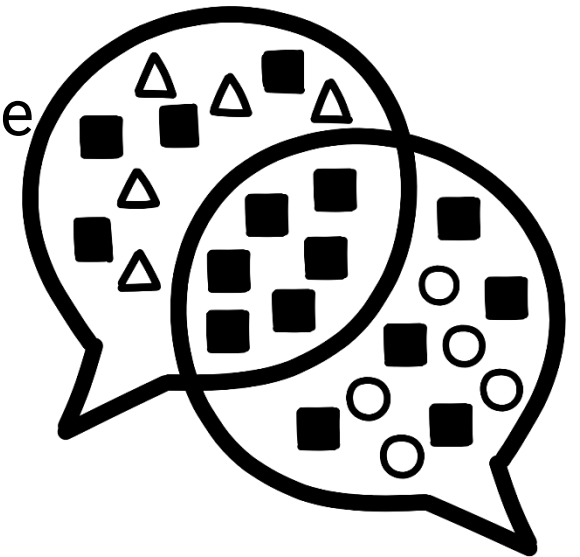
Schallaböck, J., Hoffstätter, U., Buck, D., & Linne, M. (2023). Mustervertrag Datennutzung KonsortSWD, Zenodo, CC BY 4.0, <https://doi.org/10.5281/zenodo.10409864>.

Foreign Trade Law and Dual Use of Data

- The publication of research data is also subject to foreign trade law.
- Not only the physical exchange of goods is a relevant process, but also the pure (electronic or non-physical) exchange of data, information and knowledge/know-how.
- These aspects must be taken into account when publishing data.
- If you have any questions about foreign trade law, please ask your local export control officer at your Max Planck Institute.
- If this is not possible, please contact the team at the Administrative Headquarters at aussenwirtschaft@mpg.de.

Research Data Repositories

- Repositories for research data are specialised services for the (free/restricted) publication of research data.
- They are designed for publishing research data, to ensure the specified publication specifications and processes are automatically adhered to (e.g. FAIR principles).
- Depending on the culture of the discipline and its usage different repositories may be appropriate.



<https://doi.org/10.5281/zenodo.3674561>

Recommendation for the Selection of a Research Data Repository

1. Community Specific

- i.e. [GenBank](#)

2. Institutional

- i.e. [Edmond](#)

3. Generic

- i.e. [Zenodo](#)

Find a (Open) Research Repository

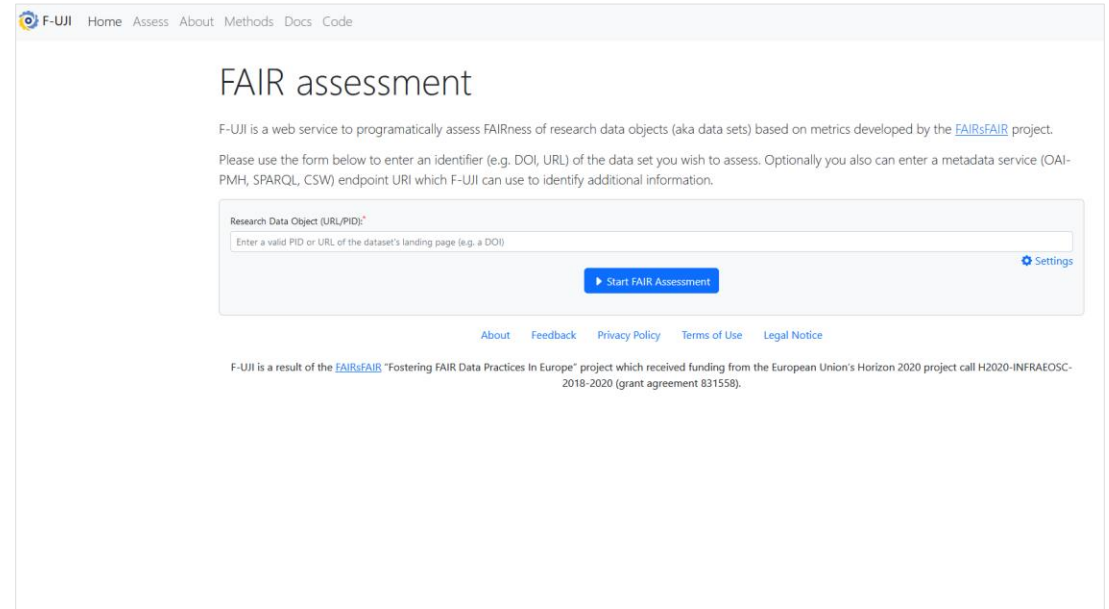
- Registry of Research Data Repositories (= re3data)
- <https://www.re3data.org>
- Best starting point for finding a suitable place for data publication



re3data: Search by Subject, CC BY 4.0, <https://www.re3data.org>.

F-Uji

- Web service to programmatically assess FAIRness of published research data sets
- <https://www.f-uji.net>
- Identifier from research data repository in F-Uji assessment:
 - i.e. <https://doi.org/10.5281/zenodo.6483002>



The screenshot shows the F-Uji web service interface. At the top, there is a navigation bar with links for Home, Assess, About, Methods, Docs, and Code. The main heading is "FAIR assessment". Below this, there is a brief description: "F-Uji is a web service to programmatically assess FAIRness of research data objects (aka data sets) based on metrics developed by the FAIRsFAIR project." This is followed by instructions: "Please use the form below to enter an identifier (e.g. DOI, URL) of the data set you wish to assess. Optionally you also can enter a metadata service (OAI-PMH, SPARQL, CSW) endpoint URI which F-Uji can use to identify additional information." The form itself has a label "Research Data Object (URL/PID):" and a text input field with a placeholder "Enter a valid PID or URL of the dataset's landing page (e.g. a DOI)". To the right of the input field is a "Settings" link. Below the input field is a blue button labeled "Start FAIR Assessment". At the bottom of the form, there are links for "About", "Feedback", "Privacy Policy", "Terms of Use", and "Legal Notice". A footer note states: "F-Uji is a result of the FAIRsFAIR 'Fostering FAIR Data Practices In Europe' project which received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 (grant agreement 831558)."

Anusuriya Devaraju, & Robert Huber. (2020). F-UJI - An Automated FAIR Data Assessment Tool, MIT license, <https://doi.org/10.5281/zenodo.4063720>.

Questions to You via Slido

Join via slido.com and enter 1174667

1. Do you know of a community-specific repository for your discipline?
2. Have you ever published research data in a community-specific repository?

Data Journals

Examples:

- Generic:
 - [Scientific Data](#)
- Engineering
 - [Journal of Chemical & Engineering Data](#)
- Humanities
 - [Journal of Open Humanities Data](#)
 - [Research Data Journal for the Humanities and Social Sciences](#)
- Life Sciences
 - [Gigascience](#)
 - [Biodiversity Data Journal](#)
 - [Open Health Data](#)

For more see https://www.forschungsdaten.org/index.php/Data_Journals

Attig et al. *Journal of Open Psychology Data* DOI: 10.5334/jopd.81 4

WAVE	INTERVIEW-MODE	DIRECT ASSESSMENTS OF CHILD AND HOME LEARNING ENVIRONMENT (HLE)	YEAR	CHILD (TARGET) AGE	PARTICIPATING FAMILIES	CHILDCARE			SUF RELEASE
						EDUCATOR	INSTITUTION	CHILDMINDER	
1	CAPI	Standardized and semi-standardized observational measures (incl. HLE)	Aug-2012- Mar 2013	7 months	3,481				2015
2	CATI, PAPI		Apr-Oct 2013	14 months	2,862	171		73	2015
	CAPI (half sample)	Standardized and semi-standardized observational measures (incl. HLE)	July-Dec 2013	17 months	1,510				2015
3	CAPI, PAPI	Semi-standardized observational measures (incl. HLE)	Apr-Nov 2014	25 months	2,609	449		110	2016
4	CAPI, PAPI	Tabbed-based competence tests	Apr-Nov 2015	3 years	2,478	625	571		2017
5	CAPI, PAPI	Tabbed-based competence tests	Apr-Sep 2016	4 years	2,381	628	521		2018
6	CAPI, PAPI	Tabbed-based competence tests	Mar-Aug 2017	5 years	2,209	683	543		2019
7	CAPI, PAPI	Tabbed-based competence tests	Apr-Sep 2018	6 years	2,116	546	444		2020
8	CAPI, PAPI	Tabbed-based competence tests	Mar-Aug 2019	7 years (mainly grade 1)	2,070				2021
9	CATI Remote/CAPI by phone ¹	Online tests	June-Sep 2020	8 years (mainly grade 2)	1,848				2022

Table 1 Overview of all waves, used modes, target age of the child, participating families and childcare institution.
 Notes: CAPI = Computer-assisted personal interview; CATI = Computer-assisted telephone interview; PAPI = Paper-and-pencil-Questionnaire; SUF = Scientific use file; HLE = Home learning environment; Each year, beginning with wave 3, there was also a small CATI-field for all families who did not take part in the CAPI. ¹Due to the Covid-19 pandemic, in wave 9 nearly all interviews with the parents were administered via telephone from the regular CAPI interviewers from their homes. ²The competence tests for the children were administered online with support of an interviewer by telephone. ³The field of the original CAPI started in March 2020 before it stopped due to the pandemic. The field was re-opened in June 2020 with telephone interviews and the online testing.

Hence, across the waves, for 1,563 children of the SC1 information from early child care personnel are available (at least one questionnaire available from childminders, educators, or head of institution).

2.2 TIME OF DATA COLLECTION
 The first assessment wave started in August 2012, the so far latest assessment included in this paper took place in 2020 (assessments in the SC1 are still continued). Table 1 gives an overview of the duration of the field for each wave.
 For further information see study overview <https://www.neps-data.de/Data-Center/Data-and-Documentation/Start-Cohort-Newborns/Documentation>.

2.3 LOCATION OF DATA COLLECTION
 Data was collected at various sample points in Germany (see also 2.4). In most waves, data collection took place in the households of the families.

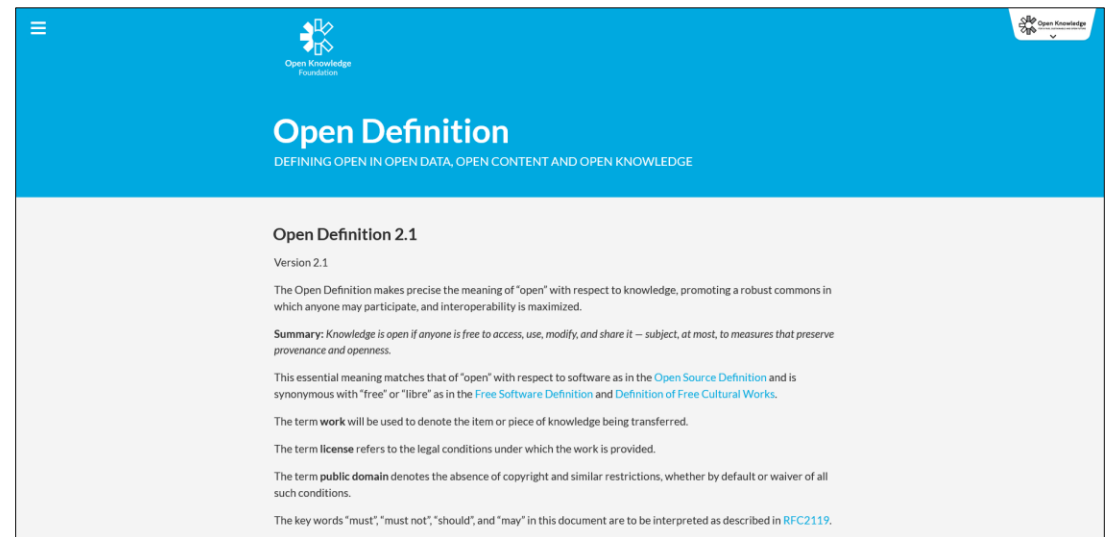
2.4 SAMPLING, SAMPLE, AND DATA COLLECTION
 The targets of SC1 are the children. As already mentioned, the first wave was intended to take place when the infants were 6 to 8 months of age. As context persons, one parent (respondent, in most cases the mother) as well as child care personnel were interviewed.
 The sample was drawn via a register-based sample of addresses available at the level of municipalities. A two-stage disproportional stratified sampling strategy was used to allow for a representative sample of the newborn population in Germany. Overall, 84 German municipalities were considered as primary sampling units (Wirbach et al., 2016). As a secondary sample, addresses out of these municipalities were drawn. Overall, 8,483 addresses from 90 sampling points in 84 municipalities were used. A detailed description of the sampling strategy can be found in Wirbach et al. (2016). Due to the fast developmental progress of infants

Attig, M., Vogelbacher, M. and Weinert, S., 2023. Education from the crib on: The potential of the Newborn Cohort of the German National Educational Panel Study. *Journal of Open Psychology Data*, 11(1), p.13. DOI: <https://doi.org/10.5334/jopd.81>, p. 4. CC BY 4.0.

Open Research Data

EU Directive 2019/1024: *“Open data as a concept is generally understood to denote data in an open format that can be freely used, re-used and shared by anyone for any purpose.”*
(<http://data.europa.eu/eli/dir/2019/1024/oj>)

“Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness.”



The screenshot shows the 'Open Definition' page from the Open Knowledge Foundation. The page has a blue header with the Open Knowledge Foundation logo and the text 'Open Definition' and 'DEFINING OPEN IN OPEN DATA, OPEN CONTENT AND OPEN KNOWLEDGE'. The main content area is white and contains the following text:

Open Definition 2.1
Version 2.1

The Open Definition makes precise the meaning of “open” with respect to knowledge, promoting a robust commons in which anyone may participate, and interoperability is maximized.

Summary: Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness.

This essential meaning matches that of “open” with respect to software as in the [Open Source Definition](#) and is synonymous with “free” or “libre” as in the [Free Software Definition](#) and [Definition of Free Cultural Works](#).

The term **work** will be used to denote the item or piece of knowledge being transferred.

The term **license** refers to the legal conditions under which the work is provided.

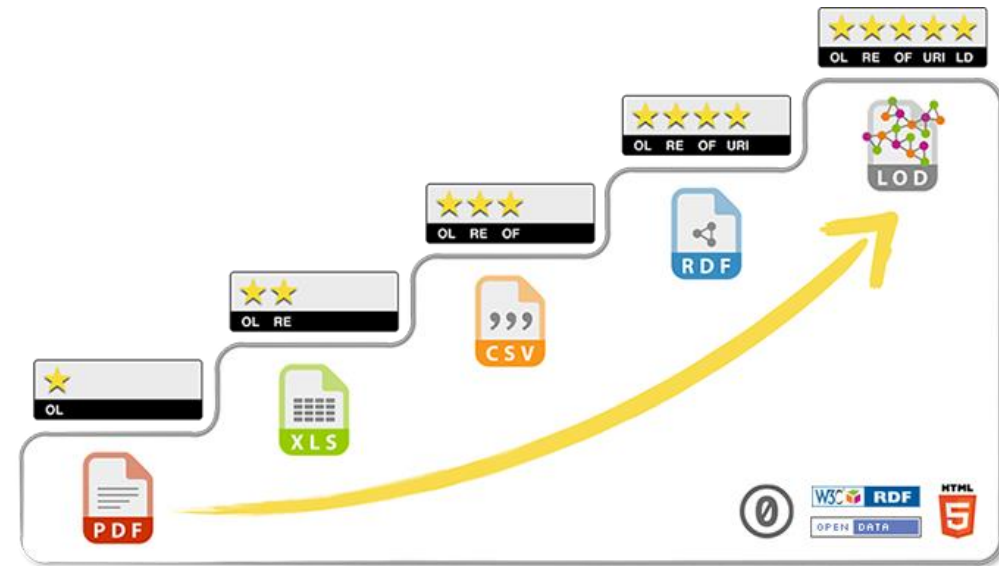
The term **public domain** denotes the absence of copyright and similar restrictions, whether by default or waiver of all such conditions.

The key words “must”, “must not”, “should”, and “may” in this document are to be interpreted as described in [RFC2119](#).

Open Knowledge Foundation, <https://opendefinition.org/od/2.1/en/>, CC BY 4.0.

5 ★ Model for Open Data by Tim Berners-Lee

- ★ Data online
- ★★ Data online available as structured data
- ★★★ Data online available in a non-proprietary open format
- ★★★★ Use identifiers to denote things, so that people can point at your data
- ★★★★★ Link your data to other data to provide context



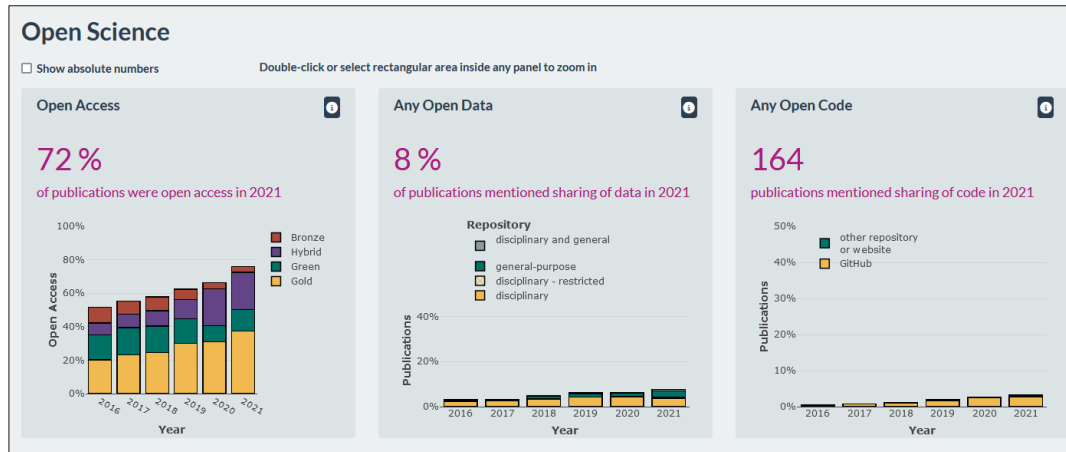
<https://5stardata.info/en/>. CC0.

<https://5stardata.info>

Open Research Data Dashboards

Charité Dashboard on Responsible Research (<https://quest-dashboard.charite.de>)

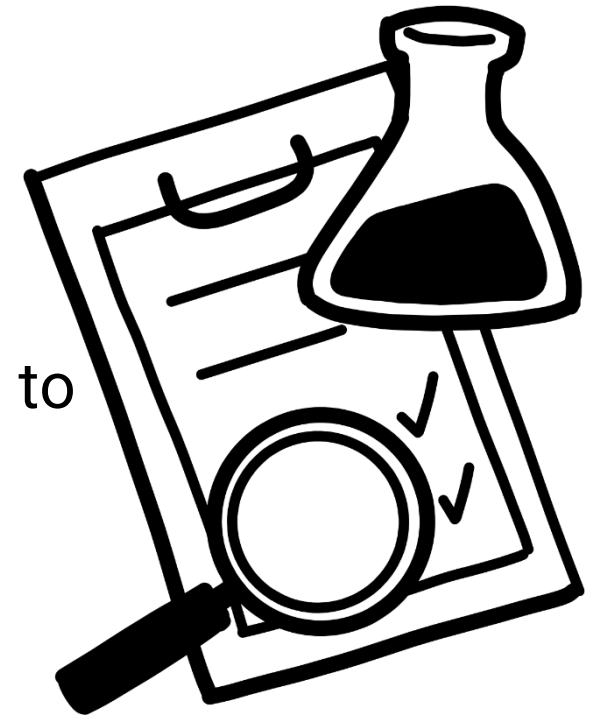
French Open Science Monitor (<https://frenchopensciencemonitor.esr.gouv.fr>)



OPEN LICENCE 2.0/LICENCE OUVERTE 2.0 <https://github.com/etalab/licence-ouverte/blob/master/open-licence.md>

Re-Use of Research Data

- The re-use of research data is a common, legitimate practice. There are rules for it.
- If in doubt, discuss the (subsequent) use of data with your supervisor.
- A licence for data can easily ensure legally compliant subsequent use.
- In the spirit of good scientific practice, you are obliged to cite your sources correctly; also re-used data.



Questions to You via Slido

Join via slido.com and enter 1174667

1. Have you already published research data?
2. Have you re-used research data for your projects?

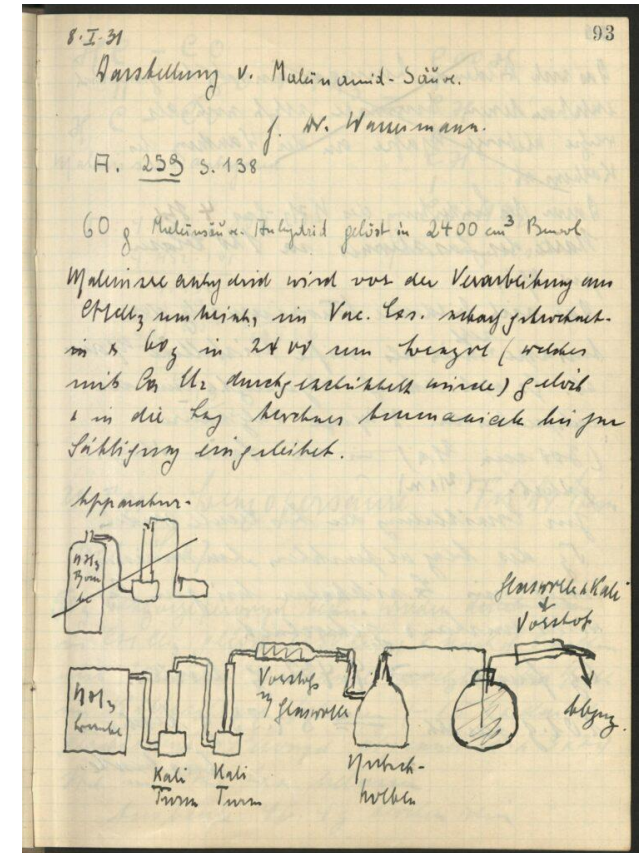
30 minutes Coffee break

Electronic Laboratory Notebooks

- Concept and use
- Selection
- Concrete examples

Aspects of a Laboratory Notebooks

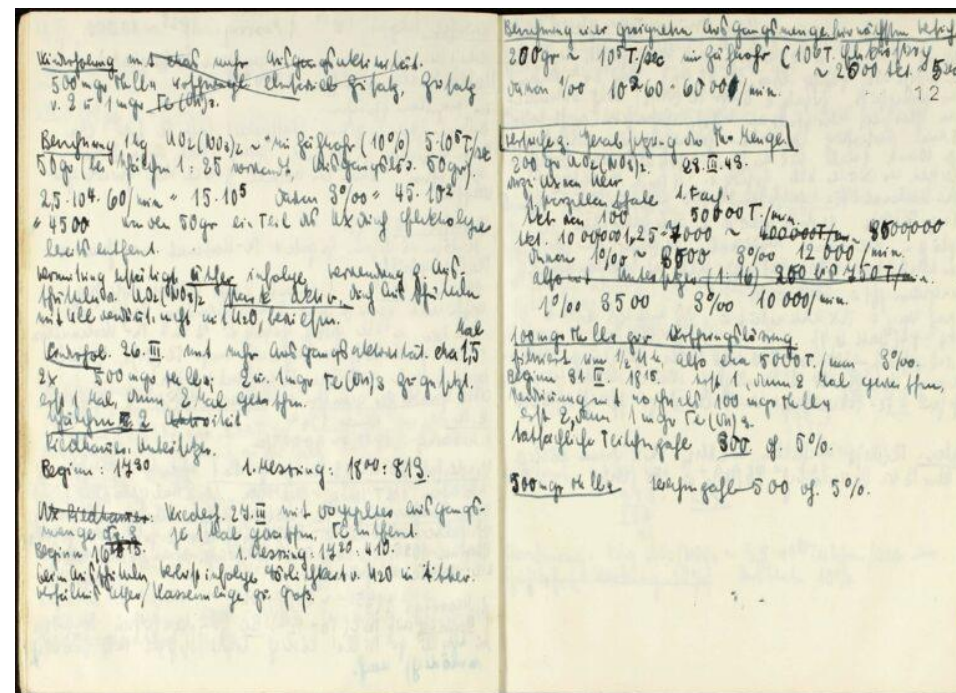
- Permanent records of research ideas, concepts, data and observations
- A laboratory notebook is a **legal** record (i.e. evidence at applying for a patent)
- Protection of your intellectual property rights
- But only, if the laboratory notebook is used **correctly**



Laboratory notebook from Hans Stocker, Kaiser Wilhelm Institute for Medical Research, 1930, folio 97, Copyright: Archive of the Max Planck Society, Berlin.

Usage of a Laboratory Notebooks

- Never remove pages
- Do not write outside the document
- Never use non-permanent pen/ink
- Use sequentially numbered pages
- Cross out blank spaces
- Do not start a new page until the previous one is full
- Never remove entries (errors must still be visible)



Laboratory notebook from Magdalena Wiedemann, Kaiser Wilhelm Institute for Chemistry, folio 13, Copyright: Archive of the Max Planck Society, Berlin.

Mende, Michael (2021): Notebook Origins - why document? Talk presented at Digital Workshop "Living with Electronic Laboratory Notebooks", Munich, 21st September 2021, <http://hdl.handle.net/21.11116/0000-0009-3F16-9>.

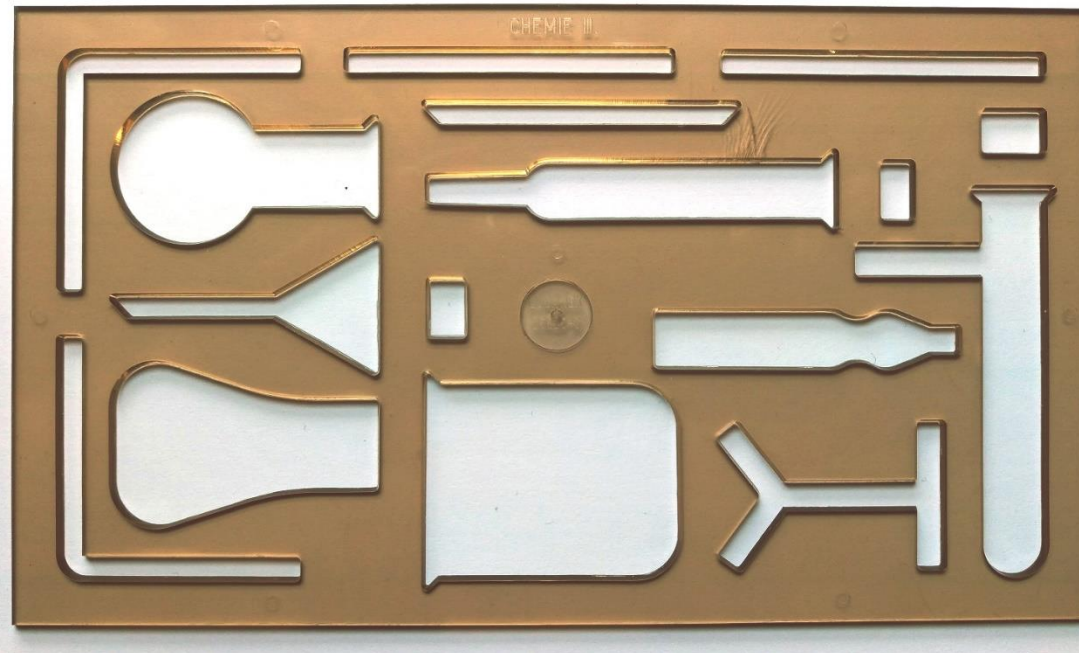
From Analogue to Digital

- Normally no delete-function in an ELN available
 - Tracking down changes etc. to one real user
 - Only one account per user
 - ELN continues to be a legal document
 - Makes sense if the whole team works with the same medium and system
- Analog lab notebook is supposedly more error-tolerant, with an ELN errors are quickly noticed (because things like pulling out pages are not possible)

Bake a Cake with ELNs: Teaching tutorial by Heike Böhm in 2021
<https://hdl.handle.net/21.11116/0000-0009-2D85-F>

Lists of ELN Software Systems

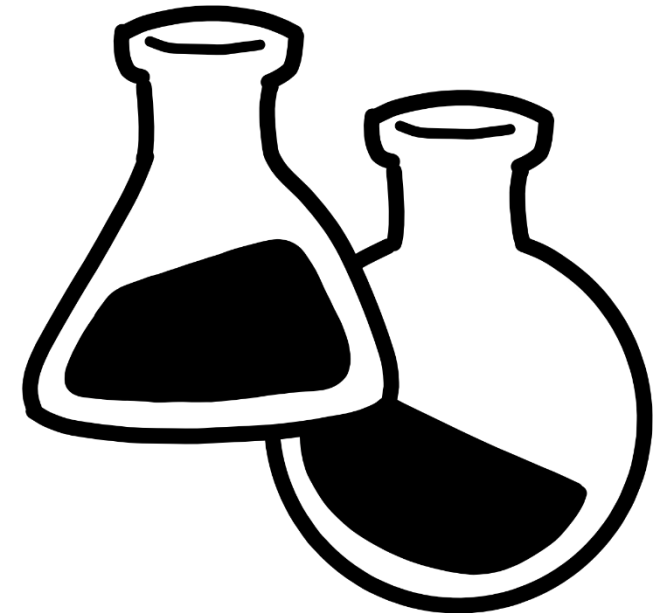
- ELN Finder (<https://eln-finder.ulb.tu-darmstadt.de>)
- Harvard Medical School (<https://datamanagement.hms.harvard.edu/analyze/electronic-lab-notebooks>)
- University of Cambridge (<https://www.data.cam.ac.uk/data-management-guide/electronic-research-notebooks/electronic-research-notebook-products>)



Schablone zum Zeichnen chemischer Gerätschaften, https://commons.wikimedia.org/wiki/File:Schablone_Logarex_25525-S,_Chemie_III.jpg, CC0.

Selection of an ELN: Example MPI-CEC

- Selection process for an MPI-CEC-wide ELN
- Pre-selection of ten ELN systems
- Three ELN systems for testing in the departments
 - eLabJournal, eLabFTW, Labstep

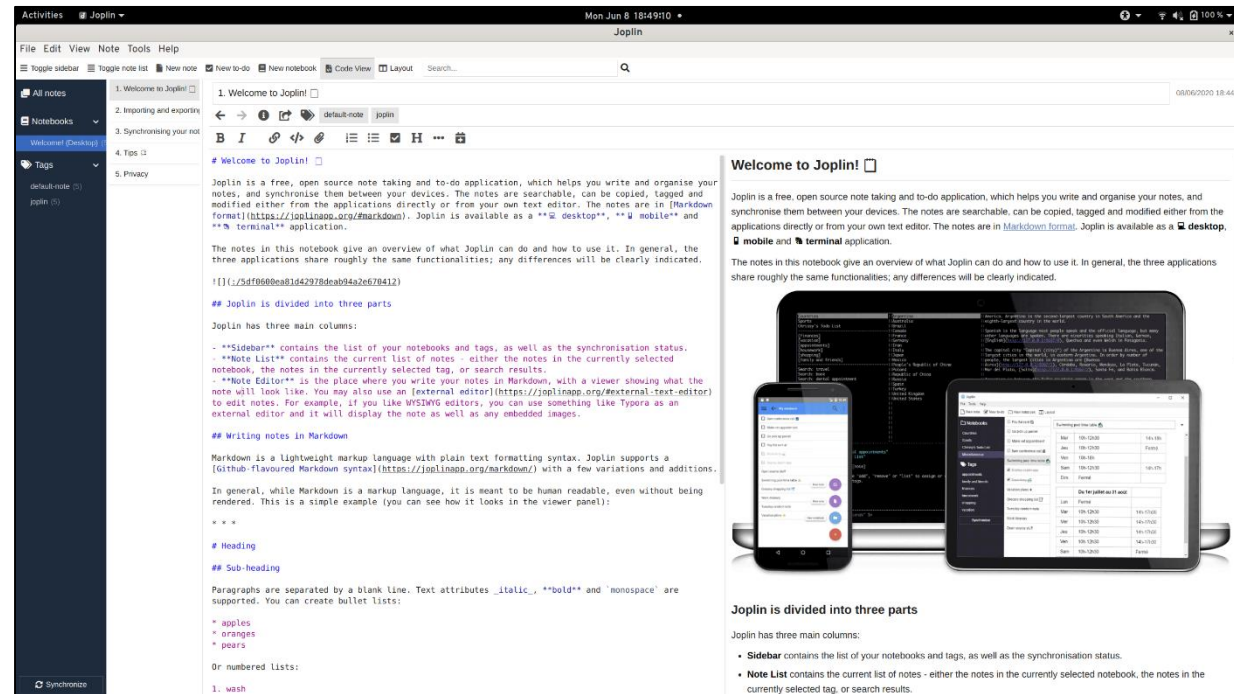


Greiner & Vaid (2021): Challenges in choosing a full-featured ELN/LIMS for a multi-disciplinary research organization. Talk presented at Digital Workshop "Living with Electronic Laboratory Notebooks", Munich, 20th September 2021, <https://hdl.handle.net/21.11116/0000-0009-3EE9-C>.

<https://doi.org/10.5281/zenodo.3674561>

Example: JoplinApp

- Website: <https://joplinapp.org>
- Demo Video: <https://www.youtube.com/watch?v=VAAA6uNPxec>
- Software Type: Open Source
- Offline functions with Nextcloud, WebDAV etc.
- E2EE encryption



Screenshot of Joplin 1.0.218 on GNOME Desktop Environment on Fedora Linux 32, by Tonica, 2020-06-08, <https://en.wikipedia.org/wiki/File:Joplin-1.0.218-on-fedora-32-20200608.png>, Public domain.

Ampuero Ruiz (2021): Can 'electronic library notebooks' (ELN) become the new field diary?. Talk presented at Digital Workshop "Living with Electronic Laboratory Notebooks", Munich, 20th September, <https://hdl.handle.net/21.11116/0000-0009-3F06-B>.

Example: eLabFTW

- Website: <https://www.elabftw.net>
- Demo: <https://demo.elabftw.net>
- Software Type: Open Source
- Some LIMS possibilities

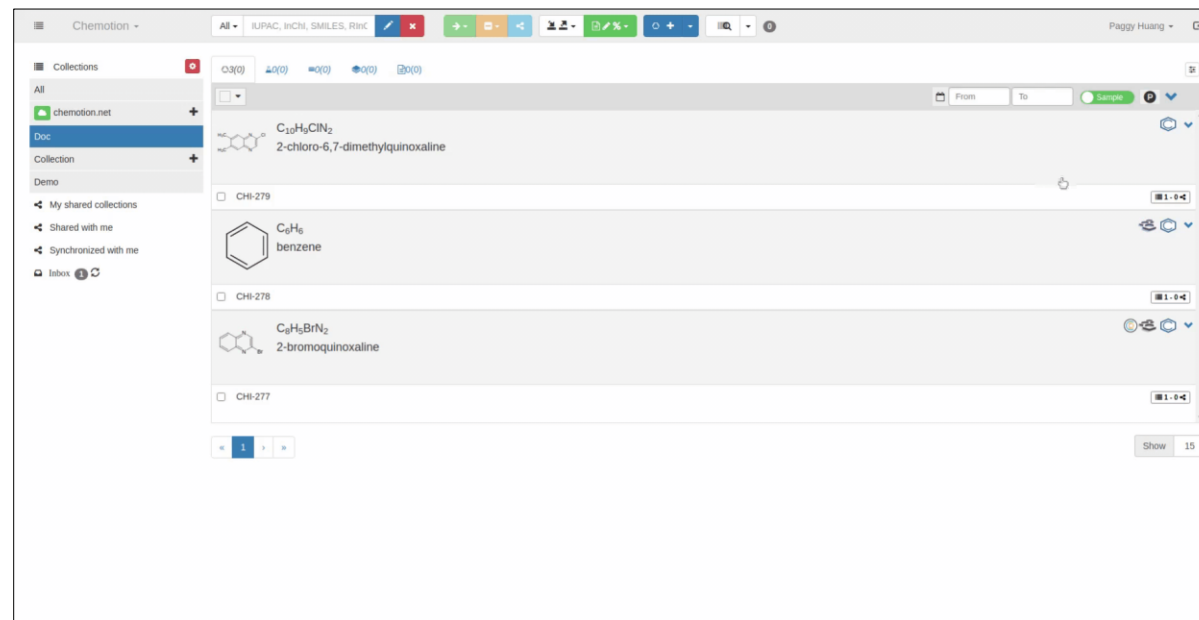


Logo eLabFTW, CC BY-SA 4.0.

See more: Niels Cautaerts. (2021): Using eLabFTW for materials science. FOSDEM21, <https://doi.org/10.5281/zenodo.4707352>, Max-Planck-Institut für Eisenforschung.

Example: Chemotion

- Website: <https://chemotion.net>
- Demo: <https://demo.chemotion.ibcs.kit.edu/home>
- Software Type: Open Source
- Developed at KIT
- Specialised in chemical research
- Lively community, e.g. NFDI4Chem
- Extensive “ecosystem”



Copyright © 2024 Chemotion – KIT, https://chemotion.net/docs/elN/ui/first_steps.

Example: OpenBIS

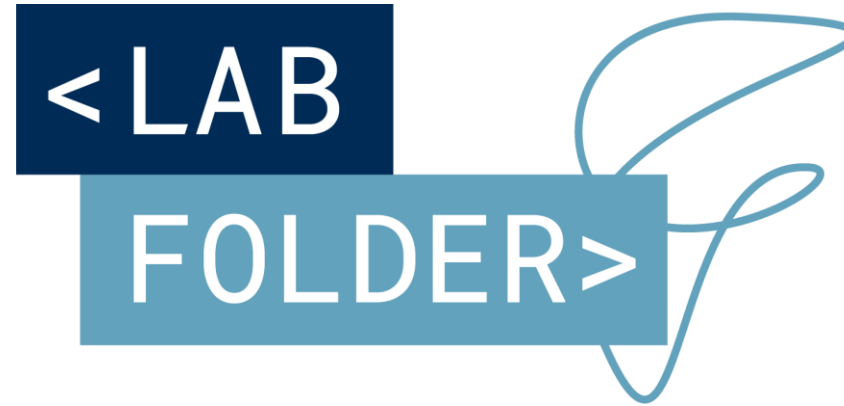
- Website: <https://openbis.ch>
- Demo: <https://openbis-eln-lims.ethz.ch/openbis/webapp/eln-lims/>
- Software-Type: Open Source
- Development: Mainly at ETH Zurich
- ELN + LIMS

FAIRDOM-SEEK

- Web-based resource for sharing heterogeneous scientific research datasets
- FAIRDOM-SEEK and OpenBIS can be combined well → a “seamless” data flow
- Example at the Max Planck Institute for Evolutionary Biology:
 - see slides from ELN workshop 2021 [Ullrich](#) and [Fortmann-Grote](#)

Example: Labfolder

- Website: <https://labfolder.com>
- MPG-wide license through the MPDL
- MPDL instance or local instance possible
- <https://labfolder.mpd.mpg.de>
- Software type: Proprietary
- Labregister as LIMS included



Questions to You via Slido

Join via slido.com and enter 1174667

1. Do you use an ELN in your working group?
2. Which ELN exactly do you use?
3. Is there a LIMS (Laboratory Inventory Management System) in use as well?

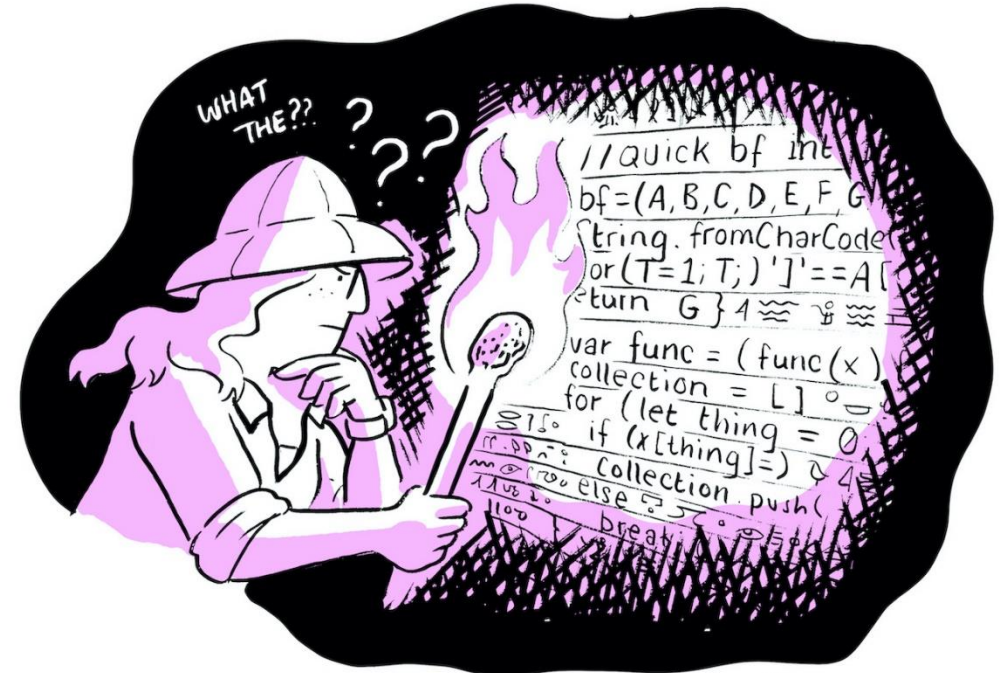
New Horizons

- Research software
- Open Science

Understanding Research Software

- In-house developed research software
- Software applications for research
- Infrastructure software/services

In discussion: Is it really helpful to distinguish between research software and “non-research software” (i.e. MS Word etc.)?



MAKE SURE YOUR CODE IS
NICE AND READABLE

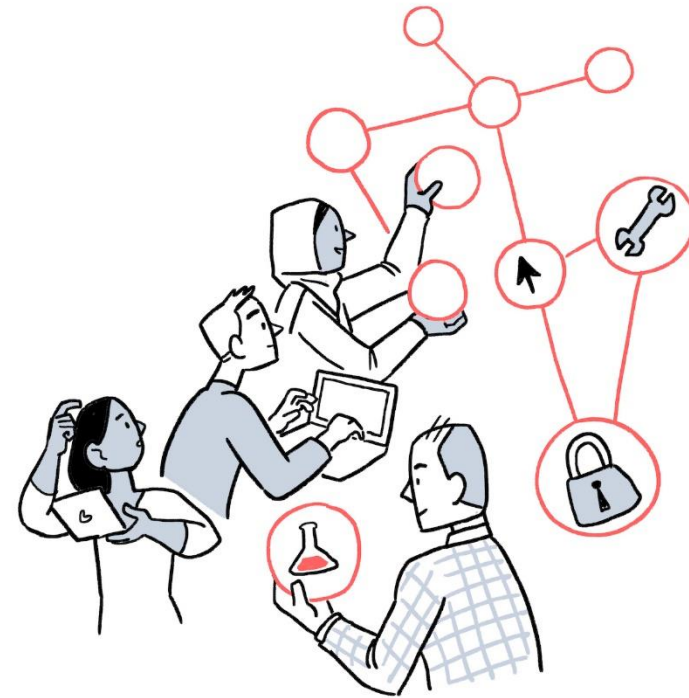
Scriberia 

The Turing Way Community, & Scriberia. (2022). Illustrations from The Turing Way: Shared under CC-BY 4.0 for reuse. Zenodo. <https://doi.org/10.5281/zenodo.6821117>.

Researchers Writing Research Software

Observations:

- Usually not thoroughly trained but autodidactic developers
- Functionality over documentation over sustainability
- After the first publication, nothing happens for a long time, and then maybe data publication and software publication
- Research software is often handed over from one PhD student to the next
- ...



Scriberia

The Turing Way Community, & Scriberia. (2022). Illustrations from The Turing Way: Shared under CC-BY 4.0 for reuse. Zenodo. <https://doi.org/10.5281/zenodo.6821117>.

Questions to You via Slido

Join via slido.com and enter 1174667

1. Are you programming software?
2. Have you attended a code training course?

Reproducibility and Accessibility

- Software is often needed to reproduce research results
- It should be accessible according to good scientific practice
- Internal policies, funders, journals require or recommend the publication of software
 - „Software programmed by researchers themselves is made publicly available along with the source code.“

Guideline 13: Providing public access to research results in the code of conduct of the DFG

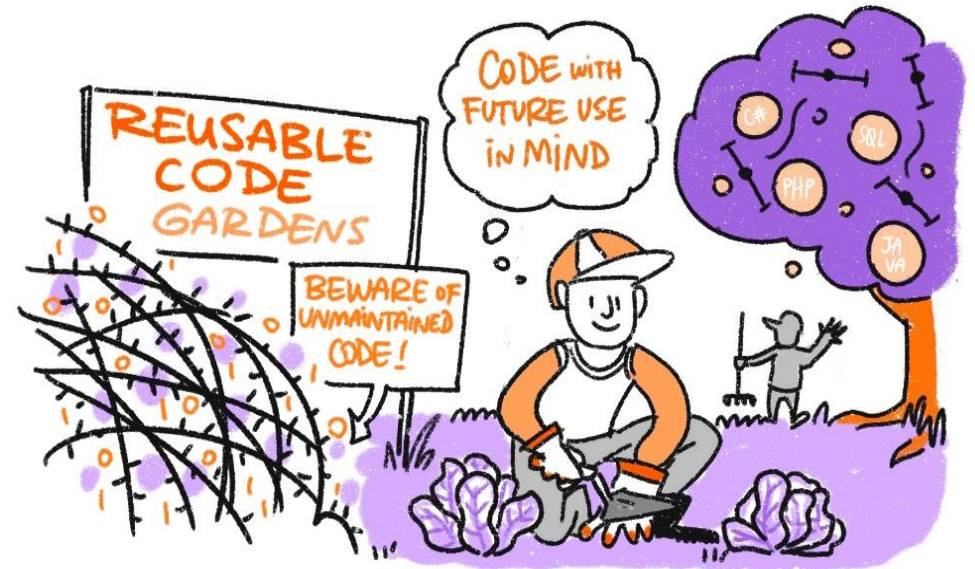


The Turing Way Community, & Scriberia. (2022). Illustrations from The Turing Way: Shared under CC-BY 4.0 for reuse. Zenodo. <https://doi.org/10.5281/zenodo.6821117>.

Re-Usability of Research Software

Conscious handling is increasing the likeliness of re-use:

- Increased probability of publication
- Explicit licensing
- Clear code structure and reflected use of third-party libraries
- A targeted approach to archiving
- ...

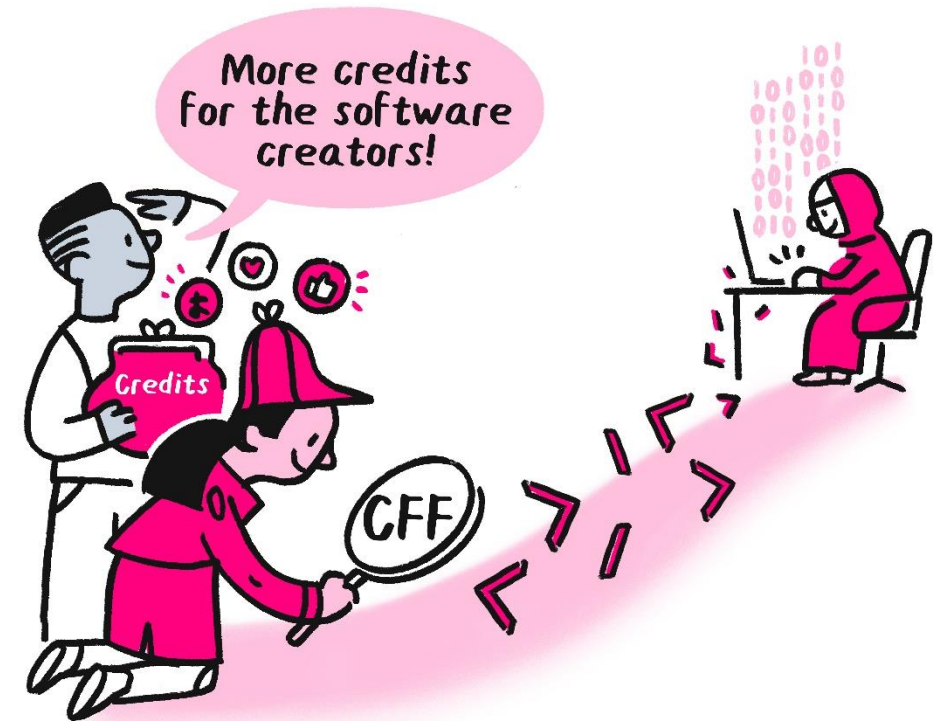


Scriberia 

The Turing Way Community, & Scriberia. (2022). Illustrations from The Turing Way:
Shared under CC-BY 4.0 for reuse. Zenodo. <https://doi.org/10.5281/zenodo.6821117>.

Recognition of Research Software

- Disciplinary credits for software publication
- Credits/funding by funders
- Institution-wide credits for development
- Policies needed
 - Normative framework for software publication
 - Endorsement of software publication



Scriberia

The Turing Way Community, & Scriberia. (2022). Illustrations from The Turing Way: Shared under CC-BY 4.0 for reuse. Zenodo. <https://doi.org/10.5281/zenodo.6821117>.

Examples for Software Policies

Software Policy

MAX-PLANCK-INSTITUT
FÜR METEOROLOGIE

CODE @ MPI-M

Software Licensing and Copyright Policy for Research Software CODE @ MPI-M

1. Preamble

The Max Planck society is determined to promote Open Access to research data, and Open Science (Berliner Erklärung¹).

MPI for Meteorology in Hamburg (MPI-M) has developed, partly on its own, partly in collaboration with partners, various Research Software, for example the ICON Model Code², the Climate Data Operators cdo³, and others.

MPI-M believes that for the benefit of science such Research Software should be released as Open-Source Software.

2. Rules

This policy treats the issues of copyright and licensing. It is applicable and restricted to Research Software Source Code developed at MPI-M (CODE). CODE must either be copyright of MPI-M alone (i.e. new code for, or code developed earlier at MPI-M under the copyright of the institute) or code contributed by third parties to MPI-M code which has been licensed to MPI-M under a permissive license like MIT/X11, (2- or 3-Clause) BSD or Apache 2.0 under the Contributor License Agreement (CLA⁴) or where unlimited copyrights have been transferred to MPI-M by other means under similar conditions.

a. Contributor License Agreement (CLA)

Every developer of CODE at MPI-M, being an employee or in any other connection to MPI-M (freelancer, guest, post-doc, scholarship etc.) (CONTRIBUTOR) must sign a CLA to be allowed to contribute to an MPI-M project in the field of Research Software development. To clarify: MPI-M can agree upon co-operations with other institutions where this policy is not applicable.

b. License

CODE shall be licensed under the BSD-3-Clause License⁵, also see attachment. For any other open source license, you must consult with the MPI-M person responsible for licenses⁶.

¹ <https://openaccess.mpg.de/Berliner-Erklärung>
² <https://mpimet.mpg.de/en/science/models>
³ <https://code.mpimet.mpg.de/projects/cdo/>
⁴ Link t.b.d. on MPI-Met Internet site.
⁵ See <https://opensource.org/licenses/BSD-3-Clause>, SPDX short identifier: BSD-3-Clause
⁶ Currently: The person responsible for strategic IT partnerships, or the head of the MPI-M administration. mailto:licenses@mpimet.mpg.de

Authors: Reinhard Budich, MPI-M; Maximilian Funk, MFS 1 / 7 Version: V3.8.2, Date: 2022-07-04; License: CC BY 4.0

Reinhard Budich und
Maximilian Funk:
Software Licensing and
Copyright Policy for
Research Software CODE
@ MPI-M, v3.8.2 vom 4.
Juli 2022,
[https://hdl.handle.net/21.
11116/0000-000C-80B1-A](https://hdl.handle.net/21.11116/0000-000C-80B1-A)
, CC BY 4.0.

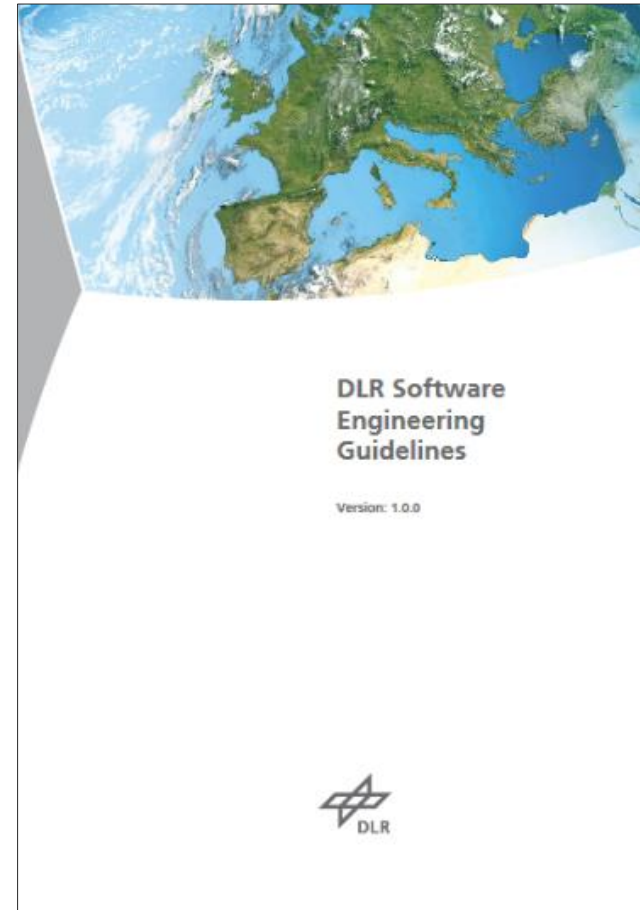
TU Delft Guidelines on Research Software Licensing, Registration and Commercialisation



Bazuine et al. (2021): TU Delft
Guidelines on Research
Software: Licensing,
Registration and
Commercialization, CC BY 4.0,
[https://doi.org/10.5281/zenodo.
4629635](https://doi.org/10.5281/zenodo.4629635).

German Aerospace Center (DLR)

- Support for developed software in terms of good software development and documentation practices
- Focus of the recommendations is on knowledge preservation and the promotion of sustainable software development in research
- Categorization according to application classes 0 to 3, depending on the scope and use of the software



Schlauch, Meinel & Haupt
(2018): DLR Software
Engineering Guidelines
(1.0.0), CC BY 4.0,
<https://doi.org/10.5281/zenodo.1344612>.

Use of Software Repositories

- Differing publication patterns between software and data
- Our observation: software that is mentioned in a publication is often directly linked to GitHub, GitLab, ...

„In 2021, one out of five publications in the arXiv corpus included a URI to GitHub“ (p. 1)



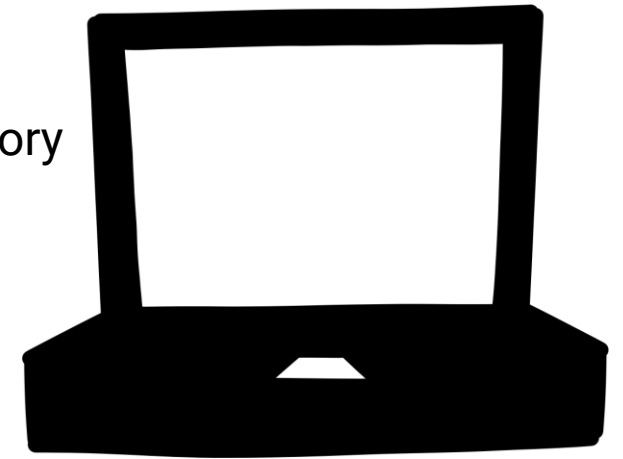
Emily Escamilla, Martin Klein, Talya Cooper, Vicky Rampin, Michele C. Weigle, Michael L. Nelson: The Rise of GitHub in Scholarly Publications, 2022, <https://doi.org/10.48550/arXiv.2208.04895>, CC BY-NC-SA 4.0.

Why Should I Write a Software Management Plan?

- **For myself!**
- Together with IT/Scientific Computing Unit/... to better design a software project
- For funding applications
- For internal planning
- For sustainability and a possible publication/archiving (good scientific practice)
- Quality assurance
- ...

Publish Software

- Checks before publication:
 - Legal considerations (e.g. integration of external library with restrictive license, dual use ...)
 - Commitment of the supervisor(s)
- Software repository, i.e. <https://github.com>, <https://gitlab.mpcdf.mpg.de>, <https://gitlab.gwdg.de>, ...
- Data repository, i.e. [Edmond](#) or [Zenodo](#)
 - manual publication
 - Automated from GitHub as push with GitHub Action towards repository
- Journal, i.e. [Journal of Open Source Software \(JOSS\)](#)
- <https://research-software-directory.org>
- Search repositories via [re3data](#)



<https://doi.org/10.5281/zenodo.3674561>

Archiving Software

1. Decision on Archival Value:

- Yes or no?
- Intention of archiving
- Resources and responsibilities
- Duration (10 years or more?)

2. Essential Elements:

- Description (e.g. documentation or detailed readme)
- (machine-readable) metadata (e.g. .cff file) on authors, affiliation, date, etc.
- License

3. Decision Conservation Status:

- Bitstream-only preservation
- Preservation of functionality
- Continuous development

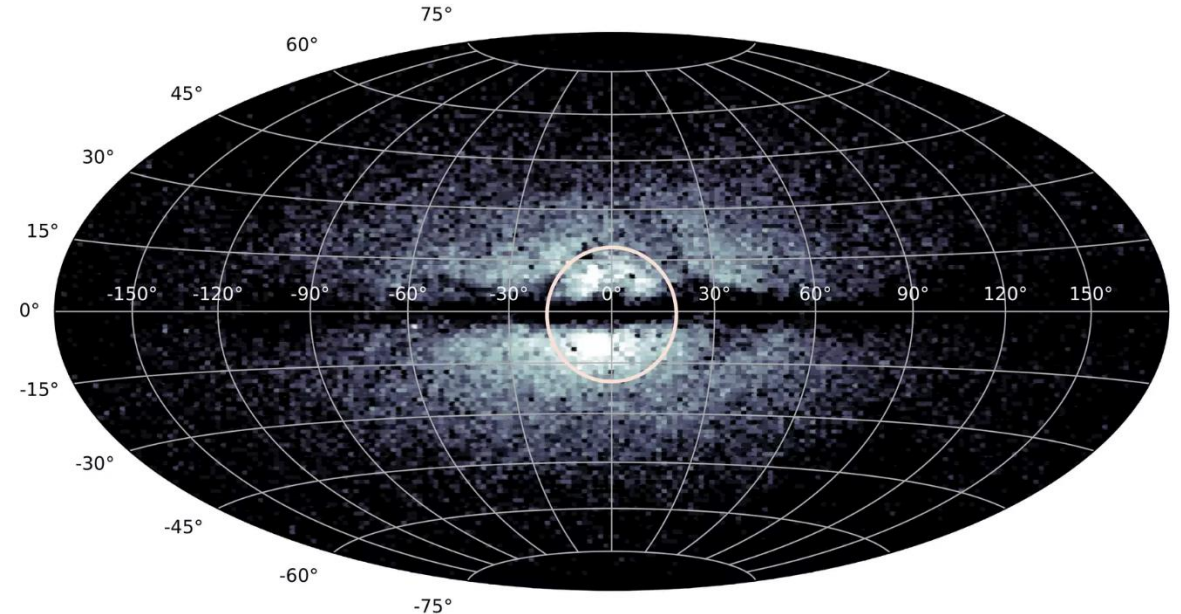
4. Decision on Archiving Location:

- At the own institute in a computer center
- In a subject repository
- Archiving of public software via [Software Heritage](#)

Research Data and AI

- AI use still new in the area of research data
- There are many possible applications for AI in research data...
 - Collect
 - Prepare
 - Search and find
 - Analyse
- ...and vice versa: be aware of what and where your research data is passed on to (external) AI providers such as OpenAI

Discoveries in old data sets using neural networks



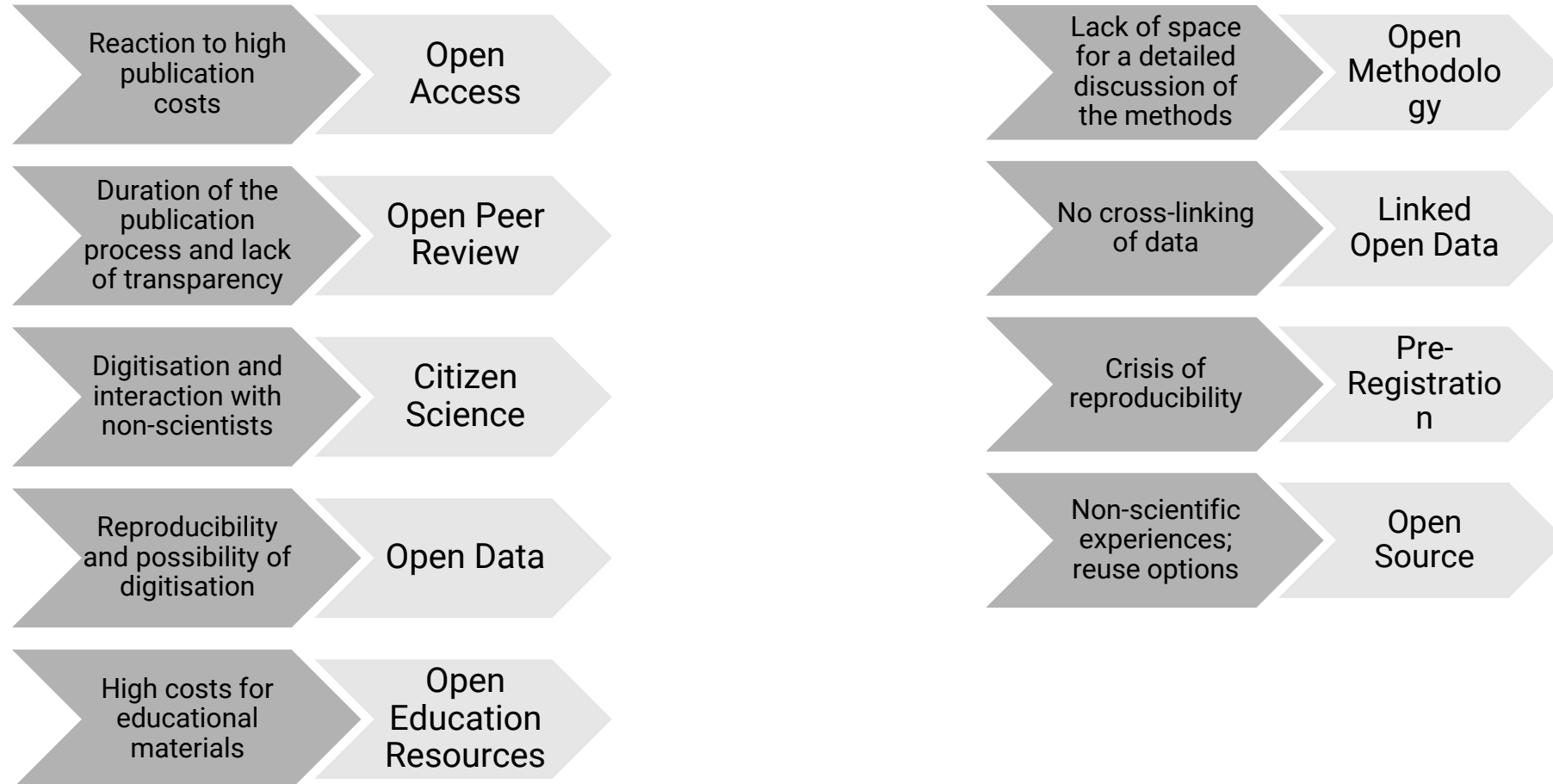
Map of the particularly metal-poor giant stars that could be identified thanks to the Gaia-DR3 dataset, image: H.-W. Rix / MPIA, <https://www.mpia.de/aktuelles/wissenschaft/2022-19-ancient-heart>.

Open Science

- Better access to and participation in scientific research and its results and methods
- Enable fairer access to science worldwide and thus also better contributions to solving global problems
- Improve efficiency, transparency and comprehensibility within science



Hypothesis: Open Science is a Reaction



UNESCO Recommendation on Open Science

- Adopted in November 2021
- <https://www.unesco.org/en/open-science>
- Increases scientific collaborations and sharing of information for the benefits of science and society
- Makes multilingual scientific knowledge openly available, accessible and reusable for everyone
- Opens the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community



UNESCO,
<https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en>, CC BY-SA 3.0.

G6 Statement on Open Science

*„We share a common commitment to the principle of making research data **“as open as possible and as closed as necessary”**.“*

- Co-signed among others by the Max Planck Society in 2021
- Commitment to FAIR and open data



G6 statement on Open Science, 2021,
<https://www.cnrs.fr/sites/default/files/download-file/G6%20statement%20on%20Open%20Science.pdf>

Open Science within Max Planck



Open Science @ Max Planck Society!

- Open Science Group within PhDnet
- Open Science Ambassadors
- Open Science in Practice by MPDL
 - Open Science Days
- ...

- Local initiatives at MPIS
 - CBS Open Science
 - ...

Max Planck Open Science Ambassadors

Open Science Ambassadors

Start

Why Open Science?

Ambassadors

Resources

Contact



OPEN SCIENCE
AMBASSADORS
MAX PLANCK SOCIETY



Why Open Science?

OSAP
2023

2023 meeting



About the
ambassadors



Resources

<https://osambassadors.mpdl.mpg.de>

**If you still haven't had
enough...**

Carpentries



<https://carpentries.org>, CC BY 4.0

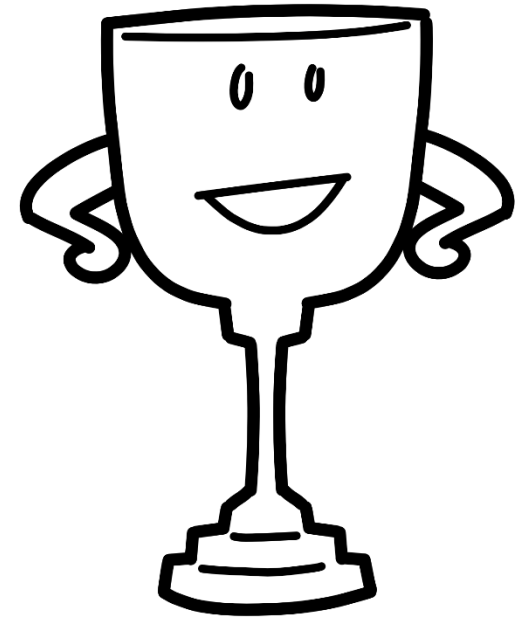
Software Carpentry

- <https://software-carpentry.org>
- Teaching basic knowledge for research software
- Participate: <https://software-carpentry.org/workshops/>
- Request workshop: <https://software-carpentry.org/workshops/request/>

Data Carpentry

- <https://datacarpentry.org>
- Data skills in a scientific context
- Example: "Data Analysis and Visualization with Python for Social Scientists"
<https://datacarpentry.org/python-socialsci/>

Your final questions?



<https://doi.org/10.5281/zenodo.3674561>

Feedback

- Here directly and/or
- via [slido.com](https://www.slido.com) and enter 1174667