# WORKSHOP ON RESEARCH DATA MANAGEMENT
## Introduction

Torsten Rathmann, Servicezentrum Forschungsdatenmanagement Wuppertal
Daniela Kastrup, ULB Düsseldorf
Bastian Weiß, UB Siegen
B. Lindstädt, A. Shutsko & J. Vandendorpe, ZB MED - Information Centre for Life Sciences

ZB MED Information Centre for Life Sciences
BERGISCHE UNIVERSITÄT WUPPERTAL
UNIVERSITÄT SIEGEN
hhu Heinrich Heine Universität Düsseldorf

Photo by ZB MED

# Outline

- ➔ Icebreaker
- ➔ Introduction of our institutions
- ➔ Research data & Research data management
- ➔ Research data lifecycle & ZB MED’s services
- ➔ Requirements of funding organizations
- ➔ Good scientific practice
- ➔ Policies & guidelines on managing research data
- ➔ Q&A
- ➔ Feedback

# Outline

- ➔ **Icebreaker**
- ➔ Introduction of our institutions
- ➔ Research data & Research data management
- ➔ Research data lifecycle & ZB MED’s services
- ➔ Requirements of funding organizations
- ➔ Good scientific practice
- ➔ Policies & guidelines on managing research data
- ➔ Q&A
- ➔ Feedback

# Participants’ background in RDM

- Your current handling of data and data management?

→ **Poll**

# Participants’ expectations

## Discussion

Photo by Volodymyr Hryshchenko on Unsplash

# Outline

- ➔ Icebreaker
- ➔ **Introduction of our institutions**
- ➔ Research data & Research data management
- ➔ Research data lifecycle & ZB MED’s services
- ➔ Requirements of funding organizations
- ➔ Good scientific practice
- ➔ Policies & guidelines on managing research data
- ➔ Q&A
- ➔ Feedback

# Universitätsbibliothek Siegen

eSs e-Science-Service
Forschungsnahe Dienste

- Research Data Management as a cooperative service from UB and ZIMT
  - Consultation (regarding DMPs, grant applications, managing/archiving/publishing data, ...), Tools (RDMO, FoDaSi, ...), self-learning resources
  - https://e-science-service.uni-siegen.de/
  - e-science-service@uni-siegen.de

# University and State Library Düsseldorf

RDM competence center HHU

- Founded this year
- A service of university and state library and ZIM
- Consulting: data management planning,  funding applications
- Tools: eLabFTW, RDMO, DSpace etc.

- https://www.fdm.hhu.de/
- fdm@hhu.de

# Servicezentrum Forschungsdatenmanagement Wuppertal

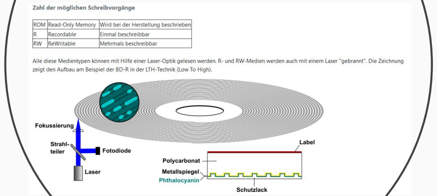
Consultation

archiving/publishing data

data management plan

grant application

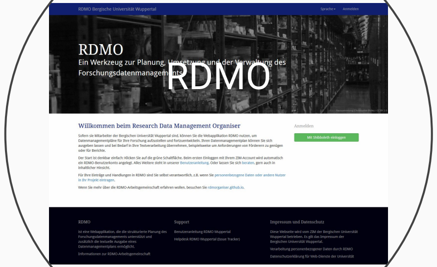
Moodle course



**Zahl der möglichen Schreibvorgänge**

| | | |
|---|---|---|
| ROM | Read-Only Memory | Wird bei der Herstellung beschrieben |
| R | Recordable | Einmal beschreibbar |
| RW | ReWritable | Mehrmals beschreibbar |

Alle diese Medientypen können mit Hilfe einer Laser-Optik gelesen werden. R- und RW-Medien werden auch mit einem Laser "gebrannt". Die Zeichnung zeigt den Aufbau am Beispiel der BD-R in der LTH-Technik (Low To High).



RDMO Bergische Universität Wuppertal

Sprache ▾ Anmelden

RDMO

Ein Werkzeug zur Planung, Umsetzung und der Verwaltung des Forschungsdatenmanagements

RDMO

Willkommen beim Research Data Management Organiser

Sofern sie Mitarbeiter der Bergischen Universität Wuppertal sind, können Sie die Webapplikation RDMO nutzen, um Datenmanagementpläne für Ihre Forschung aufzustellen und fortzuentwickeln. Ihren Datenmanagementplan können Sie sich ausgeben lassen und bei Bedarf in Ihre Textverarbeitung übernehmen, beispielsweise um Anforderungen von Förderern zu genügen oder für Berichte.

Der Start ist denkbar einfach: Klicken Sie auf die grüne Schaltfläche. Beim ersten Einloggen mit Ihrem ZIM-Account wird automatisch ein RDMO-Benutzerkonto angelegt. Alles Weitere steht in unserer Benutzeranleitung. Oder lassen Sie sich beraten, gern auch in inhaltlicher Hinsicht.

Für Ihre Einträge und Handlungen in RDMO sind Sie selbst verantwortlich, z.B. wenn Sie personenbezogene Daten oder andere Nutzer in Ihr Projekt eintragen.

Wenn Sie mehr über die RDMO-Arbeitsgemeinschaft erfahren wollen, besuchen Sie rdmorganiser.github.io.

Anmelden

Mit Shibboleth einloggen

RDMO

ist eine Webapplikation, die die strukturierte Planung des Forschungsdatenmanagements unterstützt und zusätzlich die textuelle Ausgabe eines Datenmanagementplans ermöglicht.

Informationen zur RDMO-Arbeitsgemeinschaft

Support

Benutzeranleitung RDMO Wuppertal

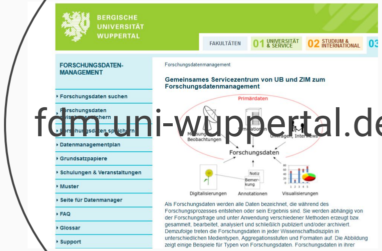
Helpdesk RDMO Wuppertal (Issue Tracker)

Impressum und Datenschutz

Diese Webseite wird vom ZIM der Bergischen Universität Wuppertal betrieben. Es gilt das Impressum der Bergischen Universität Wuppertal.

Verarbeitung personenbezogener Daten durch RDMO

Datenschutzerklärung für Web-Dienste der Universität



BERGISCHE UNIVERSITÄT WUPPERTAL

FAKULTÄTEN | 01 UNIVERSITÄT & SERVICE | 02 STUDIUM & INTERNATIONAL | 03

FORSCHUNGSDATEN-MANAGEMENT

- Forschungsdaten suchen
- Forschungsdaten zwischenspeichern
- Forschungsdaten speichern
- Datenmanagementplan
- Grundsatzpapiere
- Schulungen & Veranstaltungen
- Muster
- Seite für Datenmanager
- FAQ
- Glossar
- Support

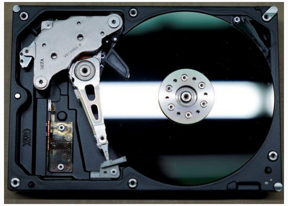
Forschungsdatenmanagement

**Gemeinsames Servicezentrum von UB und ZIM zum Forschungsdatenmanagement**

Als Forschungsdaten werden alle Daten bezeichnet, die während des Forschungsprozesses entstehen oder sein Ergebnis sind. Sie werden abhängig von der Forschungsfrage und unter Anwendung verschiedener Methoden erzeugt bzw. gesammelt, bearbeitet, analysiert und schließlich publiziert und/oder archiviert. Demzufolge treten die Forschungsdaten in jeder Wissenschaftsdisziplin in unterschiedlichen Medientypen, Aggregationsstufen und Formaten auf. Die Abbildung zeigt einige Beispiele für Typen von Forschungsdaten. Forschungsdaten in ihrer

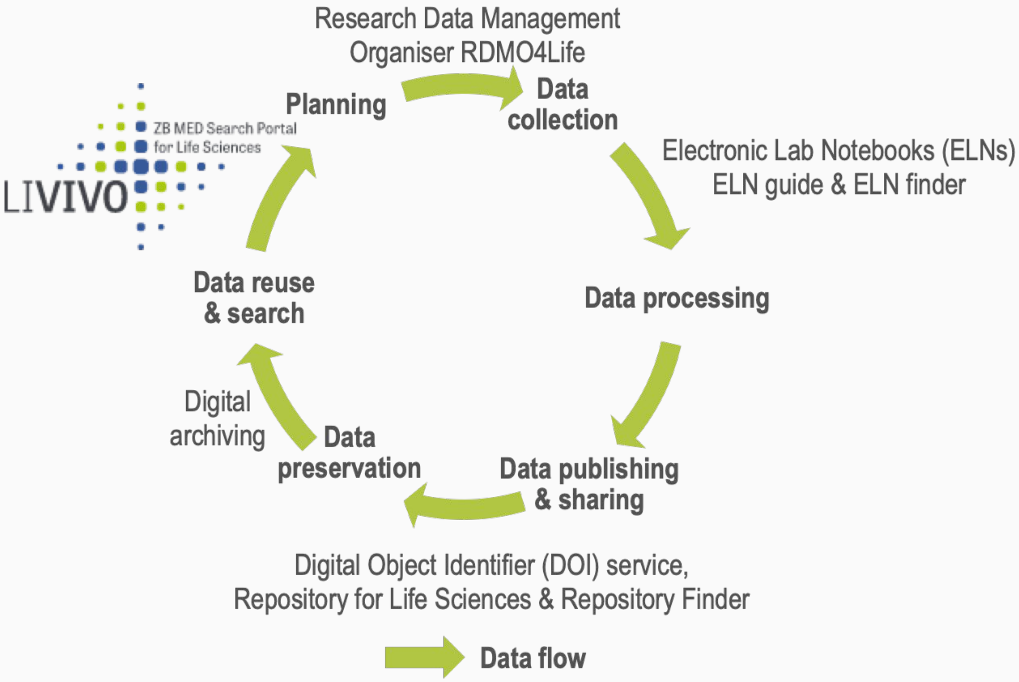
fdm.uni-wuppertal.de

Storage



File server

# ZB MED - Information Centre for Life Sciences



Research Data Management Organiser RDMO4Life

Planning

Data collection

Electronic Lab Notebooks (ELNs)
ELN guide & ELN finder

Data processing

Data publishing & sharing

Digital Object Identifier (DOI) service,
Repository for Life Sciences & Repository Finder

Data preservation

Digital archiving

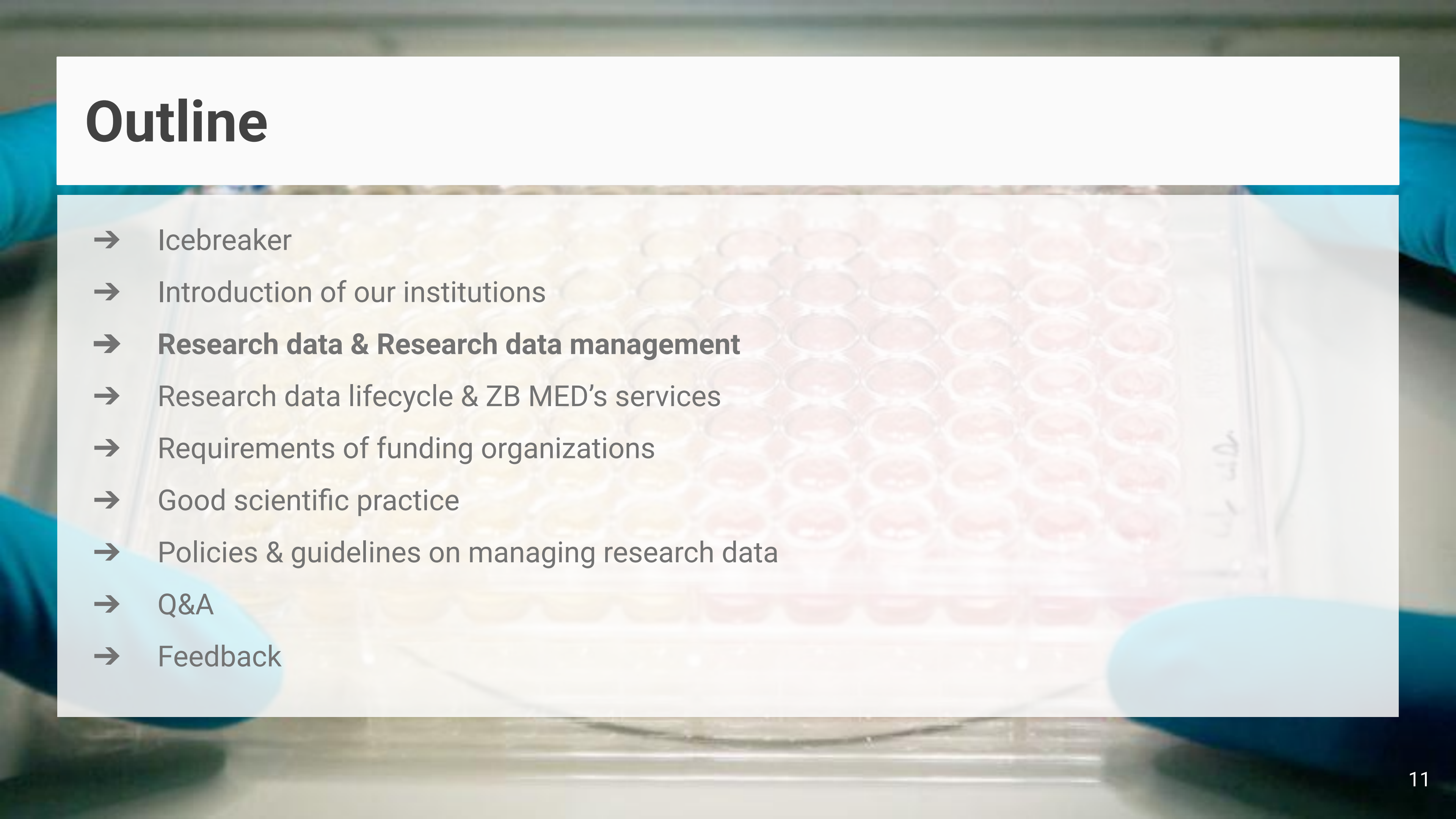
Data reuse & search

ZB MED Search Portal for Life Sciences
LIVIVO

Data flow

Website • FAQs • Tutorials

- **INFORMATION:** fostering Open Access and Open Data.
- **KNOWLEDGE:** conducting applied research to improve ZB MED’s services, and providing research support in the Life Sciences.
- **LIFE:** German National Library of Medicine, Health, Environment, Nutrition and Agriculture (world’s largest library in these fields).



# Outline

- ➔ Icebreaker
- ➔ Introduction of our institutions
- ➔ **Research data & Research data management**
- ➔ Research data lifecycle & ZB MED’s services
- ➔ Requirements of funding organizations
- ➔ Good scientific practice
- ➔ Policies & guidelines on managing research data
- ➔ Q&A
- ➔ Feedback

# Research data

- A uniform definition is missing

- The definition varies depending on:
  - Disciplines
  - Research funders [University of Leicester]

- A definition of research data:
  - “any information that has been collected, observed, generated or created to validate original research findings” [University of Leeds]

# Research data: examples

- documents, spreadsheets
- laboratory notebooks, field notebooks, diaries
- questionnaires, transcripts, codebooks
- audiotapes, videotapes
- photographs, films
- test responses
- slides, artefacts, specimens, samples
- collections of digital outputs
- data files
- database contents (video, audio, text, images)
- models, algorithms, scripts
- contents of an application
- methodologies and workflows
- standard operating procedures and protocols

[University of Leeds]

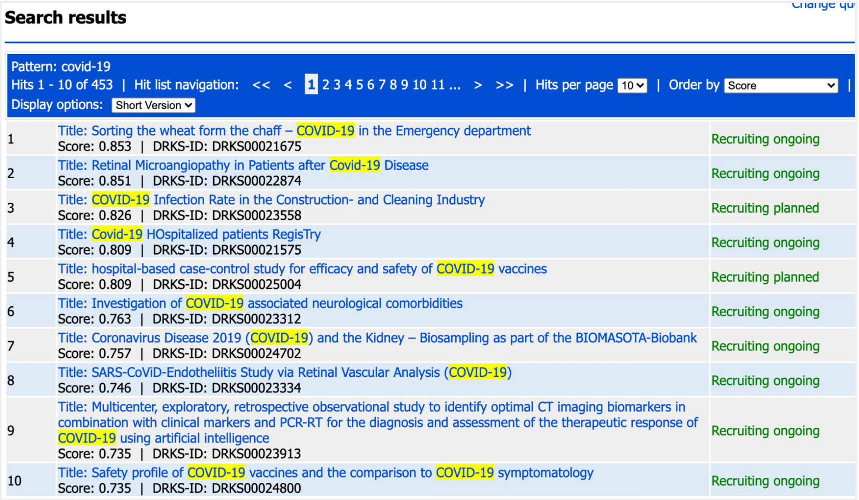
# Examples of research data in medicine

- **Electronic Medical Records** (EMRs) and **Electronic Health Records** (EHRs)
- **Patient/disease registries** (e.g. ENCePP Resources Database)
- **Health surveys** (e.g. The Rhineland Study)
- **Clinical and health data** (e.g. European Health Information Portal)
- **Clinical trials registries** and **databases** (e.g. German Clinical Trials Register (DRKS))
- **Catalogue for population health data**
- **Thesauri**, **ontologies** and **classifications** and **codes** of diseases or substances (e.g. International Statistical Classification of Diseases and Related Health Problems (ICD))

# Example of clinical trials registry

## German Clinical Trials Register (DRKS)

- **Open Access**
- Search, register and **share** information on clinical trials
- **12,000 studies** (+ 2,000/year)
- **Information:** title, short descriptions, inclusion and exclusion criteria, status and outcomes



Change qu

**Search results**

Pattern: covid-19

Hits 1 - 10 of 453 | Hit list navigation: << < 1 2 3 4 5 6 7 8 9 10 11 ... > >> | Hits per page 10 | Order by Score |

Display options: Short Version

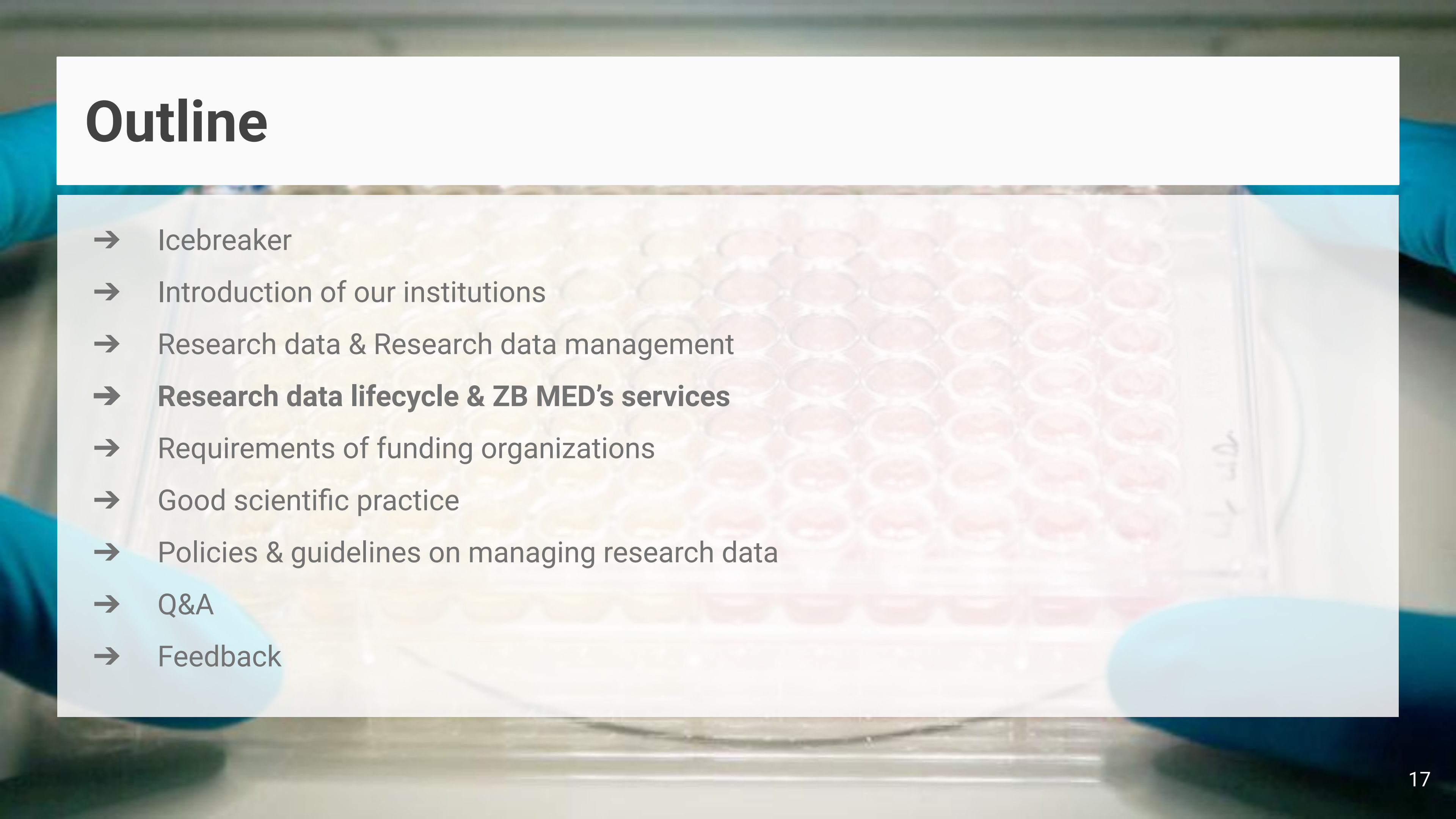
| | | |
|---|---|---|
| 1 | Title: Sorting the wheat form the chaff – COVID-19 in the Emergency department<br>Score: 0.853 \| DRKS-ID: DRKS00021675 | Recruiting ongoing |
| 2 | Title: Retinal Microangiopathy in Patients after Covid-19 Disease<br>Score: 0.851 \| DRKS-ID: DRKS00022874 | Recruiting ongoing |
| 3 | Title: COVID-19 Infection Rate in the Construction- and Cleaning Industry<br>Score: 0.826 \| DRKS-ID: DRKS00023558 | Recruiting planned |
| 4 | Title: Covid-19 HOspitalized patients RegisTry<br>Score: 0.809 \| DRKS-ID: DRKS00021575 | Recruiting ongoing |
| 5 | Title: hospital-based case-control study for efficacy and safety of COVID-19 vaccines<br>Score: 0.809 \| DRKS-ID: DRKS00025004 | Recruiting planned |
| 6 | Title: Investigation of COVID-19 associated neurological comorbidities<br>Score: 0.763 \| DRKS-ID: DRKS00023312 | Recruiting ongoing |
| 7 | Title: Coronavirus Disease 2019 (COVID-19) and the Kidney – Biosampling as part of the BIOMASOTA-Biobank<br>Score: 0.757 \| DRKS-ID: DRKS00024702 | Recruiting ongoing |
| 8 | Title: SARS-CoViD-Endotheliitis Study via Retinal Vascular Analysis (COVID-19)<br>Score: 0.746 \| DRKS-ID: DRKS00023334 | Recruiting ongoing |
| 9 | Title: Multicenter, exploratory, retrospective observational study to identify optimal CT imaging biomarkers in combination with clinical markers and PCR-RT for the diagnosis and assessment of the therapeutic response of COVID-19 using artificial intelligence<br>Score: 0.735 \| DRKS-ID: DRKS00023913 | Recruiting ongoing |
| 10 | Title: Safety profile of COVID-19 vaccines and the comparison to COVID-19 symptomatology<br>Score: 0.735 \| DRKS-ID: DRKS00024800 | Recruiting ongoing |

# Research data management

A definition of research data management:
“The research data management process is a series of steps and methods that aim to make research data usable over the long term”

- Data collection
- Data processing
- Adding metadata
- Data quality control
- Publishing and safeguarding access to data
- Archiving and ensuring the long-term interpretability of data

[ZB Med]



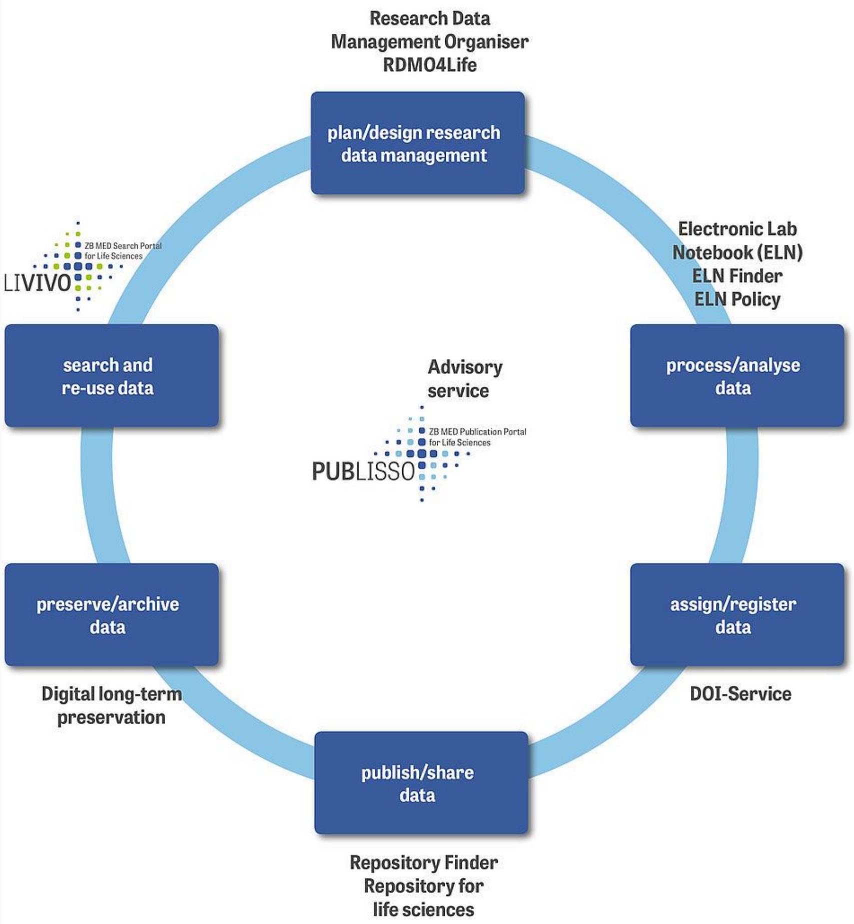
# Outline

- Icebreaker
- Introduction of our institutions
- Research data & Research data management
- **Research data lifecycle & ZB MED’s services**
- Requirements of funding organizations
- Good scientific practice
- Policies & guidelines on managing research data
- Q&A
- Feedback

# Research data life cycle

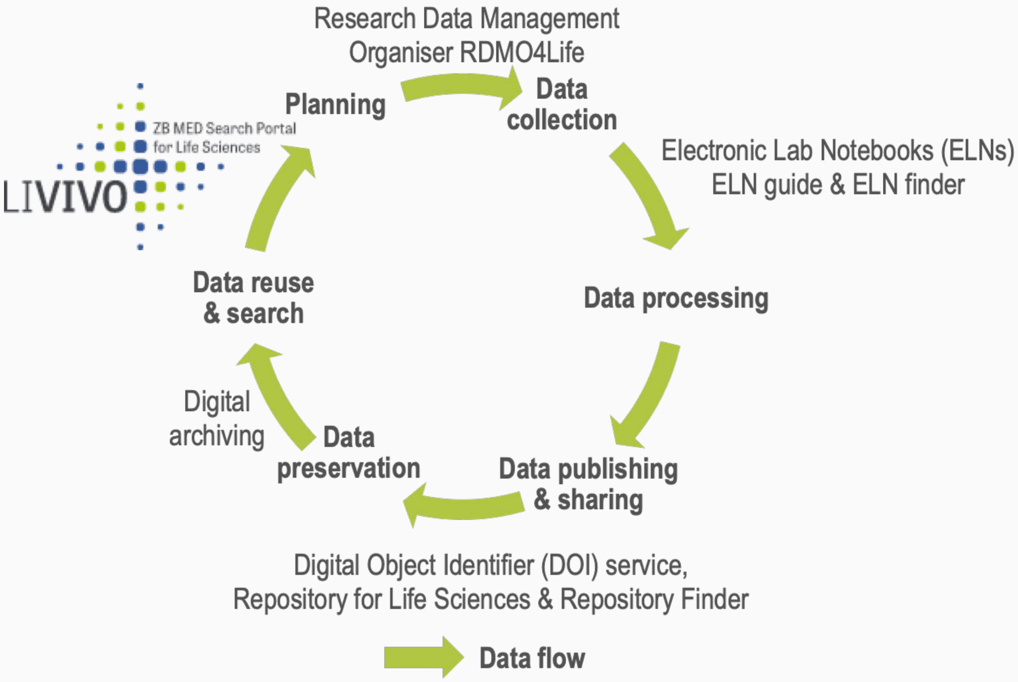
*"The research data lifecycle is a model that illustrates the stages of data management and describes how data flow through a research project from start to finish."*

- Princeton Research Data Service



Research Data Management Organiser RDMO4Life
plan/design research data management
Electronic Lab Notebook (ELN)
ELN Finder
ELN Policy
process/analyse data
assign/register data
DOI-Service
publish/share data
Repository Finder
Repository for life sciences
preserve/archive data
Digital long-term preservation
search and re-use data
LIVIVO
ZB MED Search Portal for Life Sciences
Advisory service
PUBLISSO
ZB MED Publication Portal for Life Sciences

# ZB MED's services



Research Data Management Organiser RDMO4Life

Planning

Data collection

Electronic Lab Notebooks (ELNs) ELN guide & ELN finder

Data processing

Data publishing & sharing

Digital Object Identifier (DOI) service, Repository for Life Sciences & Repository Finder

Data preservation

Digital archiving

Data reuse & search

LIVIVO ZB MED Search Portal for Life Sciences

Data flow

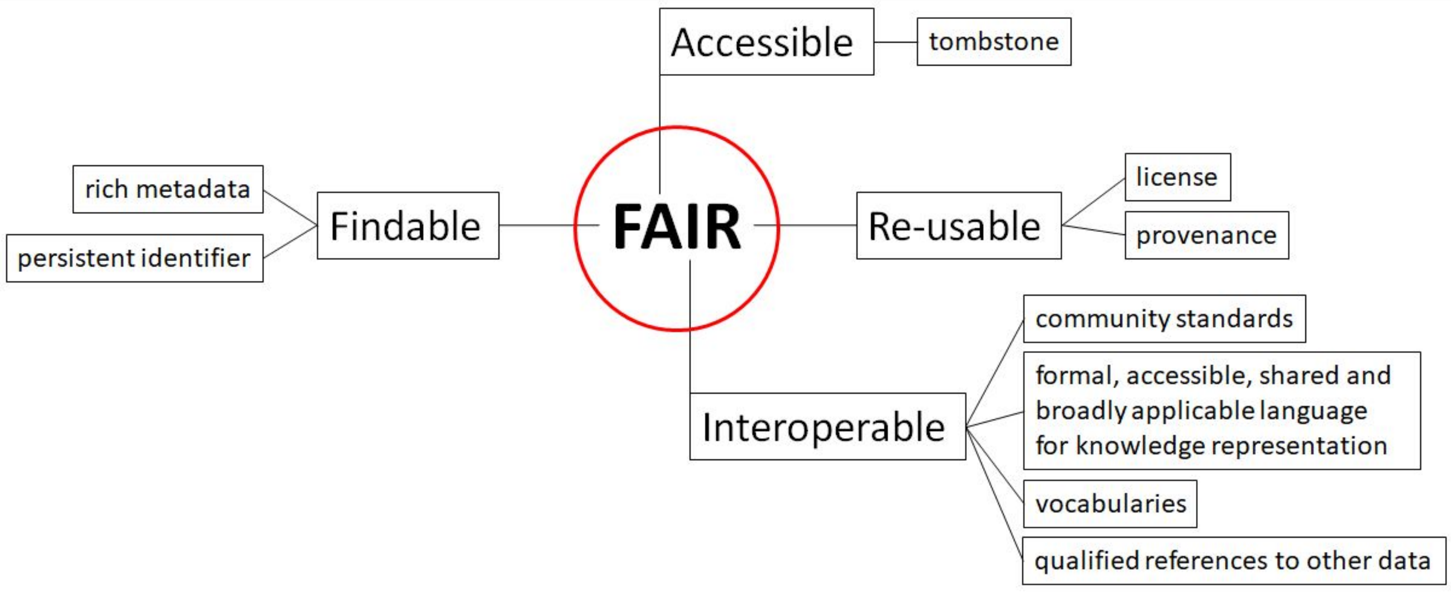
# Outline

- ➔ Icebreaker
- ➔ Introduction of our institutions
- ➔ Research data & Research data management
- ➔ Research data lifecycle & ZB MED’s services
- ➔ **Requirements of funding organizations**
- ➔ Good scientific practice
- ➔ Policies & guidelines on managing research data
- ➔ Q&A
- ➔ Feedback

# Requirements of funding organizations: Horizon Europe

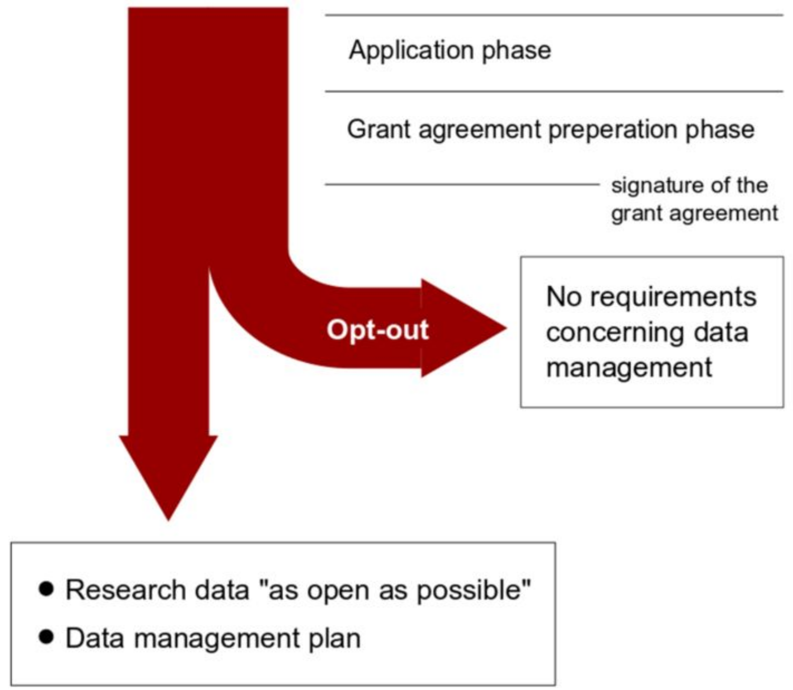
Principles:

- As open as possible, as closed as necessary
- FAIR



# Requirements of funding organizations: Horizon Europe: Opt-out

### Horizon 2020 project



Application phase

Grant agreement preperation phase

signature of the grant agreement

Opt-out

No requirements concerning data management

- Research data "as open as possible"
- Data management plan

### Horizon Europe project

Opt-out only possible with very good reasons. Examples:

- Innovative project
- Security-relevant
- Work is about vulnerable groups

# Requirements of funding organizations: Horizon Europe: 1 page about RDM already in the proposal

From the Standard Proposal Template:

**Types of data/research outputs** (e.g. experimental, observational, images, text, numerical) and their estimated size; if applicable, combination with, and provenance of, existing data.

F **Findability of data/research outputs:** Types of persistent and unique identifiers (e.g. digital object identifiers) and trusted repositories that will be used.

A **Accessibility of data/research outputs:** IPR considerations and timeline for open access (if open access not provided, explain why); provisions for access to restricted data for verification purposes.

I **Interoperability of data/research outputs:** Standards, formats and vocabularies for data and metadata.

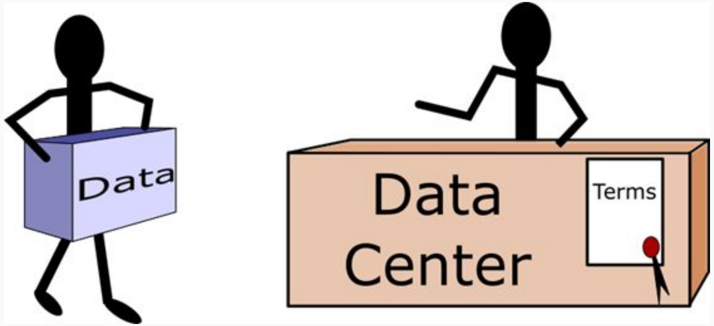
R **Reusability of data/research outputs:** Licenses for data sharing and re-use (e.g. Creative Commons, Open Data Commons); availability of tools/software/models for data generation and validation/interpretation /re-use.

**Curation and storage/preservation costs; person/team responsible for data management and quality assurance**.

# Requirements of funding organizations: Horizon Europe: Repositories

Research data shall be stored in a “trusted repository”:

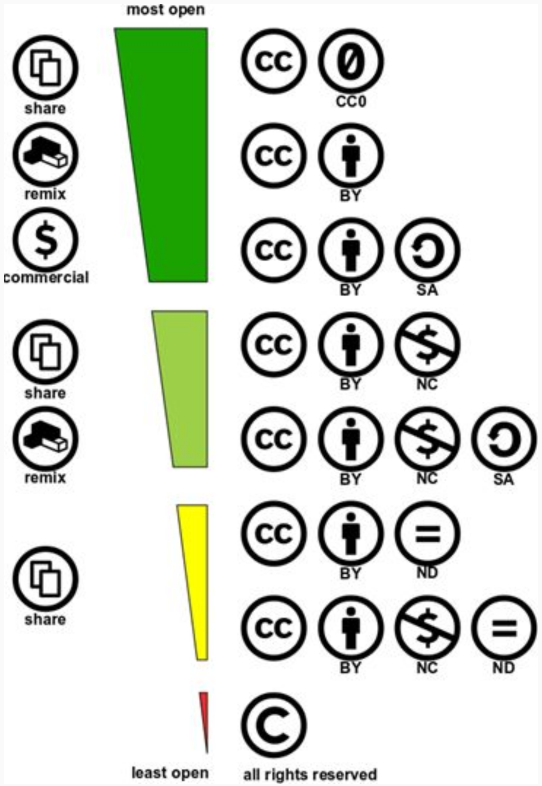
- certified or
- registered in www.opendoar.org or re3data.org or
- respected in professional circles or
- zenodo.org or figshare.org



# Requirements of funding organizations: Horizon Europe: Licenses

Explicitly mentioned:

- Creative Commons (CC)
- Open Data Commons (ODC)



for metadata

or equivalent for data

only allowed for long text formats

Image: Shaddim, [https://commons.wikimedia.org/wiki/File:Creative_commons_license_spectrum.svg](https://commons.wikimedia.org/wiki/File:Creative_commons_license_spectrum.svg) licensed under CC BY 4.0

# Requirements of funding organizations: DFG

Guidelines for Safeguarding Good Research Practice

Code of Conduct

Guideline 7: Cross-phase quality assurance

- “The nature and the scope of research data generated during the research process are described.”
- subject-specific standards

Guideline 12: Documentation

Guideline 13: Providing public access to research results

- FAIR

Guideline 17: Archiving

- ten years

Check also the third level of the Code for subject-specific extensions: https://wissenschaftliche-integritaet.de/en

# Requirements of funding organizations: DFG Guidelines on the Handling of Research Data

Make available research data

- as soon as possible
- if this does not conflict with privacy concerns or other rights of third parties
- at a stage of processing that allows it to be usefully reused by third parties

Consider

- relevance for other research contexts
- quality assurance
- data handling and long-term storage
- data types
- discipline-specific standards
- choice of suitable repositories
- third-party rights

# Requirements of funding organizations: DFG Checklist

Checklist Regarding the Handling of Research Data, for application under point 2.4

- data types, origin, how processed, data volume
- documentation, quality assurance, for re-use necessary software
- storage and access control during the course of the project
- legal peculiarities
- scientific codes and professional standards
- for re-use especially useful data
- data selection
- archiving, embargo period
- responsibilities
- necessary resources
- other research output

# Outline

- ➔ Icebreaker
- ➔ Introduction of our institutions
- ➔ Research data & Research data management
- ➔ Research data lifecycle & ZB MED’s services
- ➔ Requirements of funding organizations
- ➔ **Good scientific practice**
- ➔ Policies & guidelines on managing research data
- ➔ Q&A
- ➔ Feedback

# Good scientific practice: discussion

**Q1: what are examples of good scientific practice?**

**Good scientific practice** = ‘Good scientific practice sets out the principles, values and the standards of behaviour and practise for the Healthcare Science workforce. These standards and values must be achieved and maintained in the delivery of work activities, the provision of care and personal conduct.’ [Academy for Healthcare Science (AHCS)]

# Good scientific practice: discussion

Q2: what are examples of scientific misconduct?

# Good scientific practice: discussion

**Q3: how can we secure research integrity?**

# Good scientific practice: discussion

**Q1: what are examples of good scientific practice?**

**Suggested answer:**

- Documenting results
- Safeguarding and storing primary data
- Observing ethical standards when carrying out surveys

[Bosch, 2010;

Guidelines for Safeguarding Good Scientific Practice at the Friedrich Schiller University Jena]

# Good scientific practice: discussion

**Q2: what are examples of scientific misconduct?**

**Suggested answer:**

- Giving false information (e.g., fabrication and manipulation of raw data)
- Infringement of intellectual property (e.g., plagiarism)
- Compromising research activity of others (e.g., sabotaging research activity)

[Bosch, 2010;

Guidelines for Safeguarding Good Scientific Practice at the Friedrich Schiller University Jena]

# Good scientific practice: discussion

**Q3: how can we secure research integrity?**

**Suggested answer:**

- Establishing harmonize codes of good scientific practice (e.g., DFG's Guidelines for Safeguarding Good Research Practice)
- Regulating procedures for handling allegations of research misconduct

[Bosch, 2010;
Guidelines for Safeguarding Good Scientific Practice at the Friedrich Schiller University Jena]

# Outline

- ➔ Icebreaker
- ➔ Introduction of our institutions
- ➔ Research data & Research data management
- ➔ Research data lifecycle & ZB MED’s services
- ➔ Requirements of funding organizations
- ➔ Good scientific practice
- ➔ **Policies & guidelines on managing research data**
- ➔ Q&A
- ➔ Feedback

# Policies & guidelines on managing research data

**Policy**

‘a **definite course** or **method of action** selected from among alternatives and in light of given conditions to **guide** and **determine present** and **future decisions**’

**Guideline**

‘an **indication** or **outline** of **policy** or **conduct**’

[Merriam-Webster]

# Examples of policies and guidelines

- **General:** DFG Guidelines on the Handling of Research Data
- **Related to personal health data:**
  - FAIRDOM's Data Management Checklist
  - Medical informatics initiative (MII)'s set of standardised rules for broad access to and use of primary data from patient care
- **Institutional:** ZB MED‘s Research Data Policy (German only)

# Outline

- ➔ Icebreaker
- ➔ Introduction of our institutions
- ➔ Research data & Research data management
- ➔ Research data lifecycle & ZB MED’s services
- ➔ Requirements of funding organizations
- ➔ Good scientific practice
- ➔ Policies & guidelines on managing research data
- ➔ **Q&A**
- ➔ Feedback

# Q&A

Photo by Jon Tyson on Unsplash

# Outline

- ➔ Icebreaker
- ➔ Introduction of our institutions
- ➔ Research data & Research data management
- ➔ Research data lifecycle & ZB MED’s services
- ➔ Requirements of funding organizations
- ➔ Good scientific practice
- ➔ Policies & guidelines on managing research data
- ➔ Q&A
- ➔ **Feedback**

# Feedback

Thank you for attending our webinar. We now would like to ask you to fill in a 4-question survey to improve our webinars. In advance, thank you for your help.

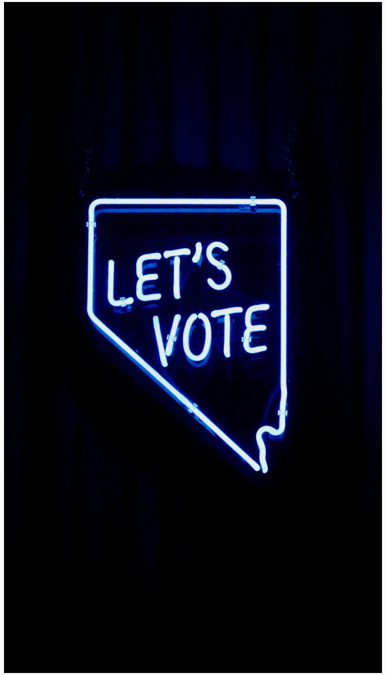


Photo by Manny Becerra
on Unsplash

# Thank you!

For further information we are at your disposal

**ZB MED –
Information Centre for
Life Sciences**
Gleueler Straße 60
50931 Köln

forschungsdaten@zbmed.de
www.zbmed.de

**Bergische Universität
Wuppertal –
Servicezentrum FDM**
Gaußstraße 20
42119 Wuppertal

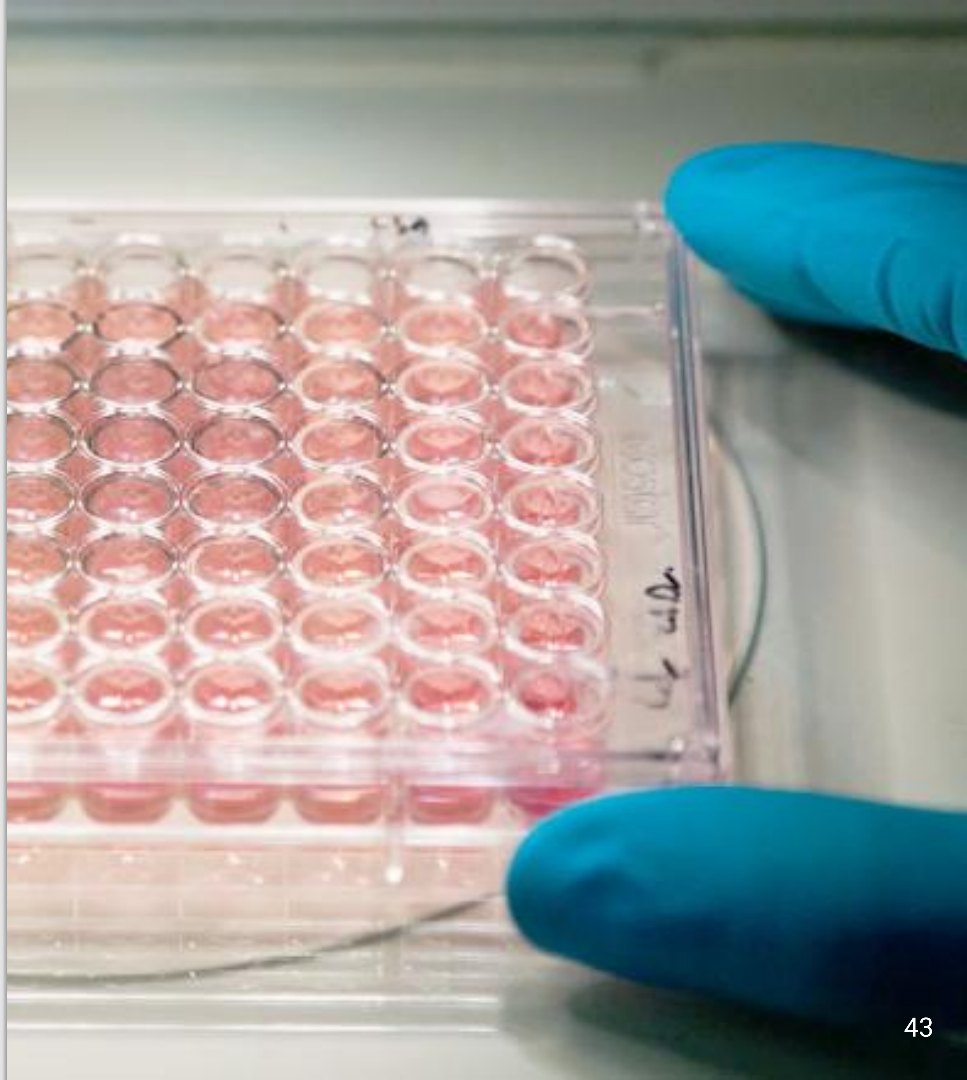
fdm@uni-wuppertal.de
fdm.uni-wuppertal.de

**HHU Düsseldorf**
Universitätsstraße 1
40225 Düsseldorf

fdm@hhu.de
https://www.fdm.hhu.de

**Universität Siegen –
e-Science-Service**

e-science-service@uni-siegen.de
https://e-science-service.uni-siegen.de/



# WORKSHOP ON RESEARCH DATA MANAGEMENT
# Introduction & Planning

Torsten Rathmann, Servicezentrum Forschungsdatenmanagement Wuppertal
Daniela Kastrup, ULB Düsseldorf
Bastian Weiß, UB Siegen
B. Lindstädt, A. Shutsko & J. Vandendorpe, ZB MED - Information Centre for Life Sciences

ZB MED Information Centre for Life Sciences
BERGISCHE UNIVERSITÄT WUPPERTAL
UNIVERSITÄT SIEGEN
hhu Heinrich Heine Universität Düsseldorf

Photo by ZB MED

# Outline

- ➔ Introduction
  - ◆ Metadata & metadata standards
  - ◆ FAIR data principles
- ➔ Planning
  - ◆ Data Management Plans (DMPs)
  - ◆ RDMO
  - ◆ Examples of DMPs
- ➔ Q&A
- ➔ Feedback

# Outline

- **Introduction**
  - **Metadata & metadata standards**
  - FAIR data principles
- Planning
  - Data Management Plans (DMPs)
  - RDMO
  - Examples of DMPs
- Q&A
- Feedback

# A definition and examples of metadata

## Definition

‘structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called **data about data** or information about information.’

## Examples

- Name
- Topic
- Features
- Categories
- Geospatial information

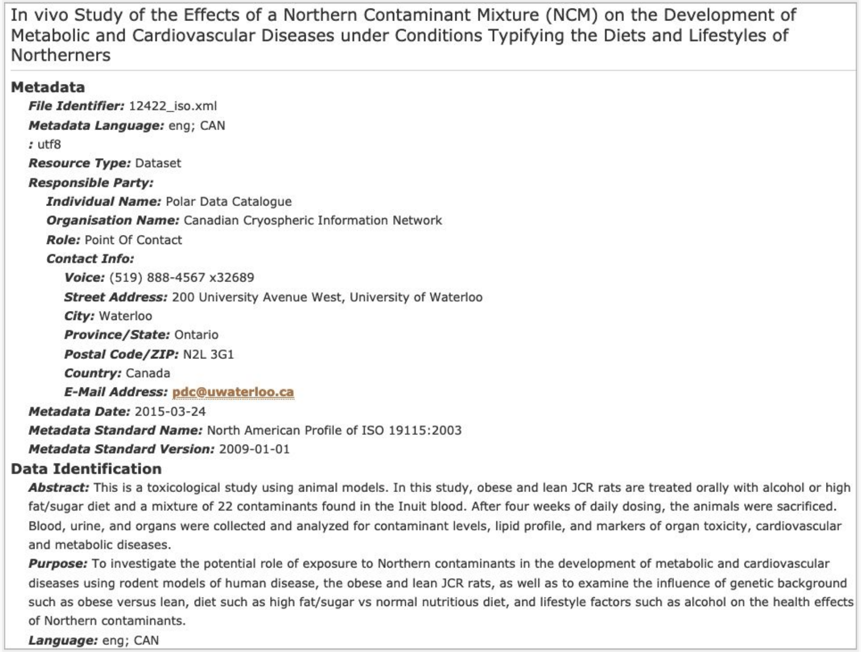
[NISO]

# A definition of metadata in the life sciences

## Biological metadata

« Metadata characterize biological resources by **core information** including a name, a description of its input and its output (parameters or format), its address, and various additional properties. »

[Springer Link]



In vivo Study of the Effects of a Northern Contaminant Mixture (NCM) on the Development of Metabolic and Cardiovascular Diseases under Conditions Typifying the Diets and Lifestyles of Northerners

**Metadata**

***File Identifier:*** 12422_iso.xml
***Metadata Language:*** eng; CAN
***:*** utf8
***Resource Type:*** Dataset
***Responsible Party:***
***Individual Name:*** Polar Data Catalogue
***Organisation Name:*** Canadian Cryospheric Information Network
***Role:*** Point Of Contact
***Contact Info:***
***Voice:*** (519) 888-4567 x32689
***Street Address:*** 200 University Avenue West, University of Waterloo
***City:*** Waterloo
***Province/State:*** Ontario
***Postal Code/ZIP:*** N2L 3G1
***Country:*** Canada
***E-Mail Address:*** pdc@uwaterloo.ca
***Metadata Date:*** 2015-03-24
***Metadata Standard Name:*** North American Profile of ISO 19115:2003
***Metadata Standard Version:*** 2009-01-01

**Data Identification**

***Abstract:*** This is a toxicological study using animal models. In this study, obese and lean JCR rats are treated orally with alcohol or high fat/sugar diet and a mixture of 22 contaminants found in the Inuit blood. After four weeks of daily dosing, the animals were sacrificed. Blood, urine, and organs were collected and analyzed for contaminant levels, lipid profile, and markers of organ toxicity, cardiovascular and metabolic diseases.

***Purpose:*** To investigate the potential role of exposure to Northern contaminants in the development of metabolic and cardiovascular diseases using rodent models of human disease, the obese and lean JCR rats, as well as to examine the influence of genetic background such as obese versus lean, diet such as high fat/sugar vs normal nutritious diet, and lifestyle factors such as alcohol on the health effects of Northern contaminants.

***Language:*** eng; CAN

[NWT Discovery Portal]

# Formats to document metadata

**Poll: in which format(s) are metadata documented in your research group?**

- In the **file** itself
- In **text files** (e.g., README, XML file)
- In **input fields**
- In your **own schemes**
- In **standardised schemes**
- In a **metadata repository**
- Metadata are **not documented** in my research group

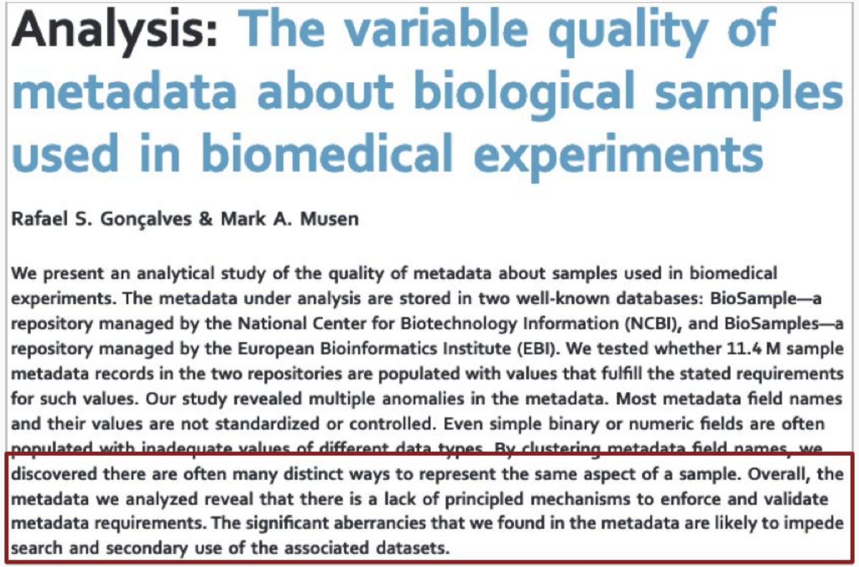
# Importance of metadata

- To make your data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (**FAIR**).
- To make your data **understandable**, **usable** and **shareable**.
- To facilitate the **long-term archiving** and **preservation** of data.
- To make your data **citable** by other researchers.
- To make the context for how your data was created/analysed/stored **reproducible**.
- To ensure consistency (i.e. **quality management**)

[Harvard University]

# Importance of metadata in the life sciences

« Overall, the metadata we analyzed reveal that there is a lack of principal mechanisms to enforce and validate metadata requirements. The significant aberrancies that we found in the metadata are likely to **impede search** and **secondary use** of the associated **datasets**. »



Analysis: The variable quality of metadata about biological samples used in biomedical experiments

Rafael S. Gonçalves & Mark A. Musen

We present an analytical study of the quality of metadata about samples used in biomedical experiments. The metadata under analysis are stored in two well-known databases: BioSample—a repository managed by the National Center for Biotechnology Information (NCBI), and BioSamples—a repository managed by the European Bioinformatics Institute (EBI). We tested whether 11.4 M sample metadata records in the two repositories are populated with values that fulfill the stated requirements for such values. Our study revealed multiple anomalies in the metadata. Most metadata field names and their values are not standardized or controlled. Even simple binary or numeric fields are often populated with inadequate values of different data types. By clustering metadata field names, we discovered there are often many distinct ways to represent the same aspect of a sample. Overall, the metadata we analyzed reveal that there is a lack of principled mechanisms to enforce and validate metadata requirements. The significant aberrancies that we found in the metadata are likely to impede search and secondary use of the associated datasets.

[Gonçalves and Musen 2019]

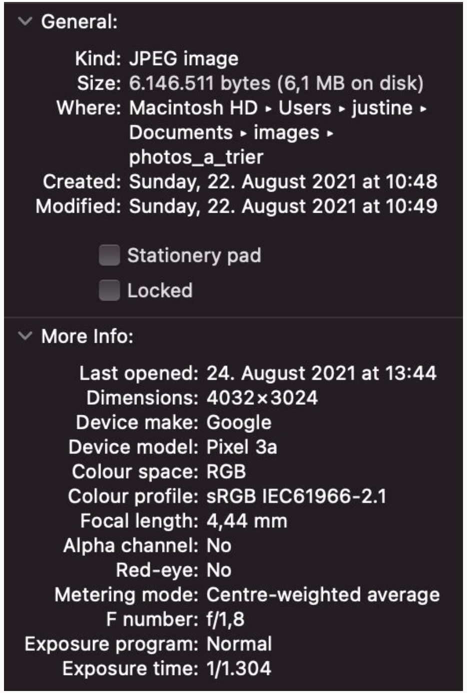
# Examples of metadata in the life sciences

| Type of metadata | Core information about… |
|---|---|
| **Reagent** | Clinical samples, biological or chemical reagents |
| **Technical** | Measurements made by the use of research instruments |
| **Experimental** | Experimental conditions, the experimental protocol, and the equipment used to generate the data |
| **Analytical** | Data analysis methods |
| **Dataset-level** | Objectives of the research project, participating investigators, recent publications, and funding sources |

**Source:** Harvard University

# Technical metadata

- **Automatically generated** by software associated to research instruments (e.g. metadata generated by cameras in images files)
- Metadata acquisition can be partly configured in the **software settings**
- Metadata **export** must sometimes be initiated deliberately



General:

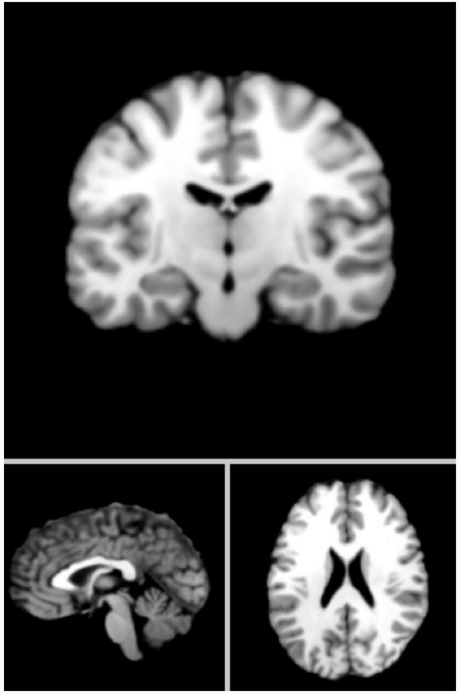
Kind: JPEG image  
Size: 6.146.511 bytes (6,1 MB on disk)  
Where: Macintosh HD ▸ Users ▸ justine ▸ Documents ▸ images ▸ photos_a_trier  
Created: Sunday, 22. August 2021 at 10:48  
Modified: Sunday, 22. August 2021 at 10:49

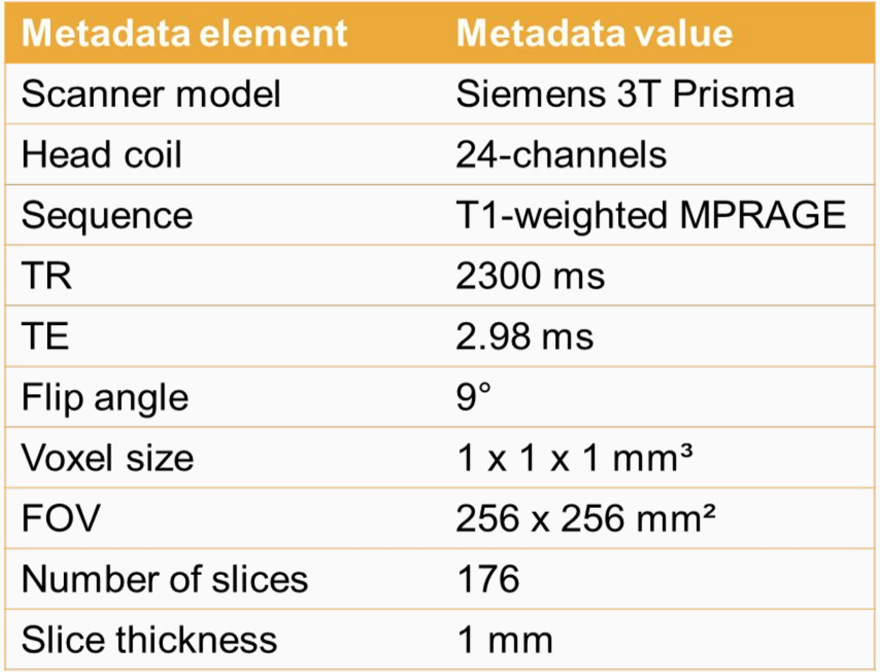
□ Stationery pad  
□ Locked

More Info:

Last opened: 24. August 2021 at 13:44  
Dimensions: 4032×3024  
Device make: Google  
Device model: Pixel 3a  
Colour space: RGB  
Colour profile: sRGB IEC61966-2.1  
Focal length: 4,44 mm  
Alpha channel: No  
Red-eye: No  
Metering mode: Centre-weighted average  
F number: f/1,8  
Exposure program: Normal  
Exposure time: 1/1.304

# Technical metadata

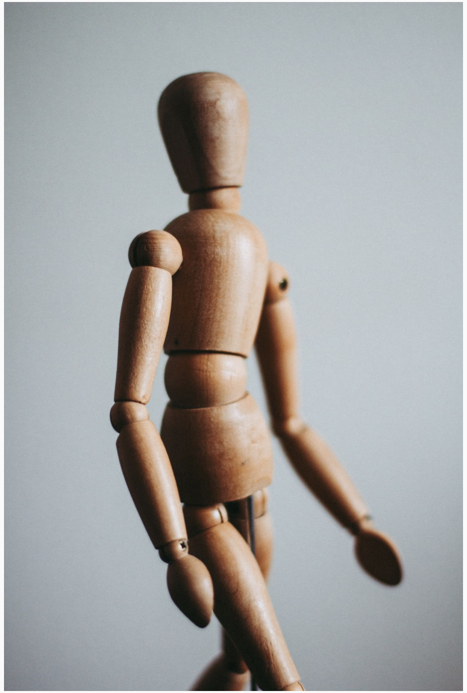




| Metadata element | Metadata value |
|---|---|
| Scanner model | Siemens 3T Prisma |
| Head coil | 24-channels |
| Sequence | T1-weighted MPRAGE |
| TR | 2300 ms |
| TE | 2.98 ms |
| Flip angle | 9° |
| Voxel size | 1 x 1 x 1 mm$^3$ |
| FOV | 256 x 256 mm$^2$ |
| Number of slices | 176 |
| Slice thickness | 1 mm |

# Importance of metadata standards in the life sciences

- Human / human
- *Homo sapiens / homo sapiens*
- *H. sapiens / h. sapiens*
- *Homo sapiens sapiens / homo sapiens sapiens*
- *H. sapiens sapiens / h. sapiens sapiens*



# Definitions of metadata standards & Cie

**Standards**

‘something set up and established by authority as a rule for the measure of quantity, weight, extent, value, or quality’

[Merriam-Webster]

**Controlled vocabulary**

‘organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching’

[Getty Center]

**Ontologies**

'set of concepts and categories in a subject area or domain that shows their properties and the relations between them’

[LEXICO]

# Relevant example in the life sciences

- **Corona Component Standards** (CoCos) initiative whose aim is to establish uniform data formats and standards for CoViD-19 and SARS-CoV-2-related data.
- **German Corona Consensus Data Set** (GECCO) is the core data set of the CoCos initiative:
  - Personal data: age, gender, height and weight
  - Laboratory values: blood pressure, cholesterol level, etc.
  - Risk factors
  - Medication intake
  - Symptoms
  - Initiated therapy procedures

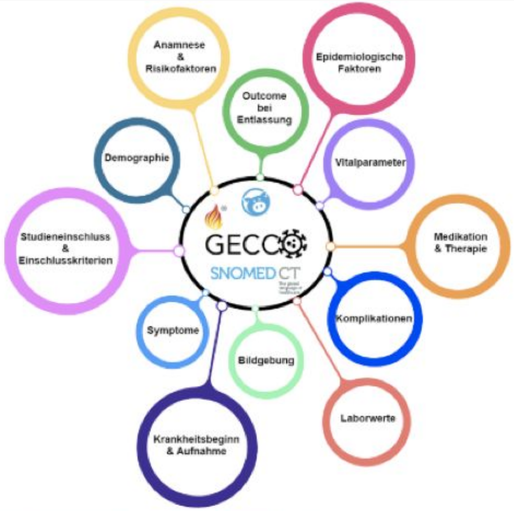


Abbildung 1 Komponenten von medizinischen Informationen

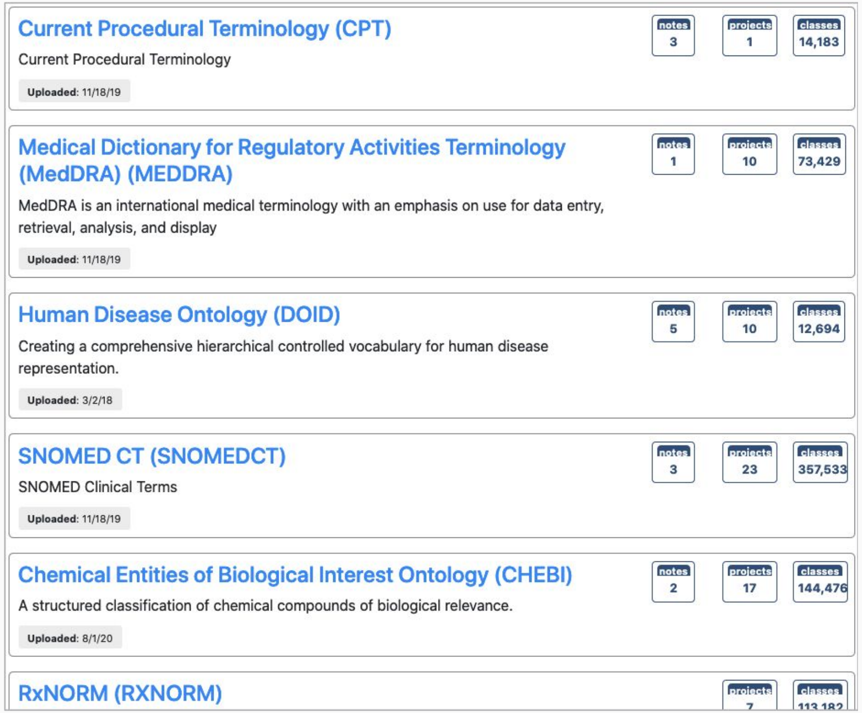
Sources: von Kalle and Thun 2020, German Biobank Node

# Examples of metadata standards in the life sciences

- **To report:**
  - Clinical data: SNOMED CT
  - Diseases and health conditions: ICD
  - Data derived by relevant methods in biosciences: MIBBI
- **To index** journal articles and books in the life sciences: MeSH
- **To exchange:**
  - Clinical and translational research data: CDISC - ODM-XML
  - Healthcare information electronically: HL7 FHIR
- **Formats:**
  - For neutron, x-ray, and muon science: NeXus
  - For storing microscopy information: OME-XML

# Examples of metadata standard directories, registries and repositories

- Basel Register of Thesauri, Ontologies & Classifications (BARTOC)
- DCC Disciplinary Metadata guide
- RDA Metadata Standards Directory
- BioPortal - Repository of biomedical ontologies
- Cancer Data Standards Registry and Repository (caDSR)



**Current Procedural Terminology (CPT)**
Current Procedural Terminology
**Uploaded:** 11/18/19
notes 3 · projects 1 · classes 14,183

**Medical Dictionary for Regulatory Activities Terminology (MedDRA) (MEDDRA)**
MedDRA is an international medical terminology with an emphasis on use for data entry, retrieval, analysis, and display
**Uploaded:** 11/18/19
notes 1 · projects 10 · classes 73,429

**Human Disease Ontology (DOID)**
Creating a comprehensive hierarchical controlled vocabulary for human disease representation.
**Uploaded:** 3/2/18
notes 5 · projects 10 · classes 12,694

**SNOMED CT (SNOMEDCT)**
SNOMED Clinical Terms
**Uploaded:** 11/18/19
notes 3 · projects 23 · classes 357,533

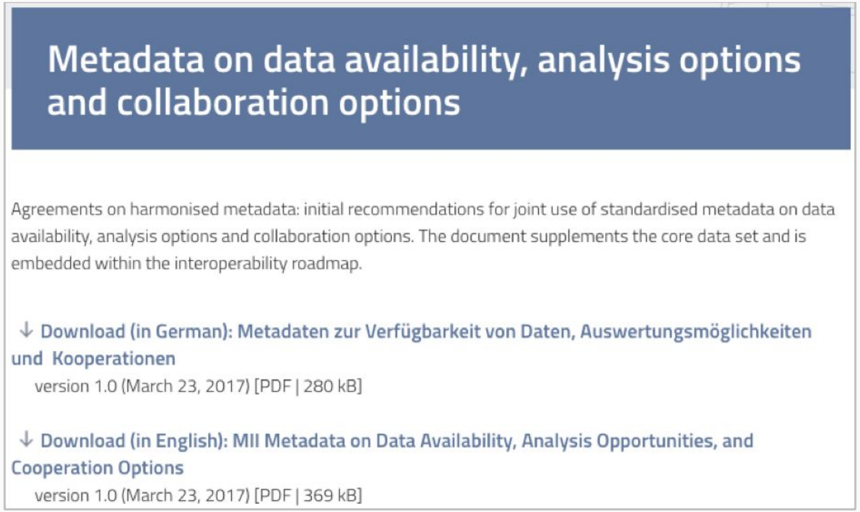
**Chemical Entities of Biological Interest Ontology (CHEBI)**
A structured classification of chemical compounds of biological relevance.
**Uploaded:** 8/1/20
notes 2 · projects 17 · classes 144,476

**RxNORM (RXNORM)**
projects 7 · classes 113,182

[BioPortal]

# Recommendations on using metadata standards in the life sciences

The German medical informatics initiative (MII)’s recommendations for the joint use of standardised metadata on data availability, analysis options and collaboration options.



## Metadata on data availability, analysis options and collaboration options

Agreements on harmonised metadata: initial recommendations for joint use of standardised metadata on data availability, analysis options and collaboration options. The document supplements the core data set and is embedded within the interoperability roadmap.

↓ Download (in German): Metadaten zur Verfügbarkeit von Daten, Auswertungsmöglichkeiten und Kooperationen
version 1.0 (March 23, 2017) [PDF | 280 kB]

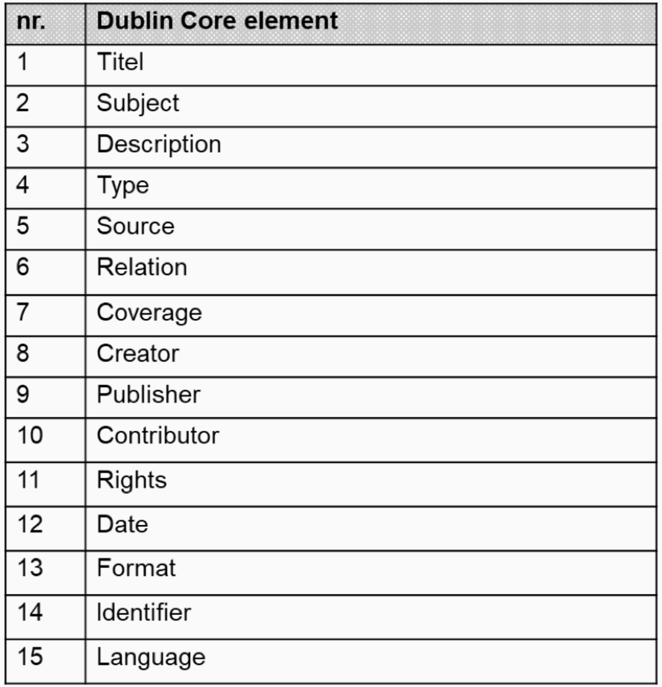
↓ Download (in English): MII Metadata on Data Availability, Analysis Opportunities, and Cooperation Options
version 1.0 (March 23, 2017) [PDF | 369 kB]

# Example of controlled vocabulary

Dublin Core Metadata Initiative:

« domain agnostic, basic and widely used metadata standard »
[Cornell University]

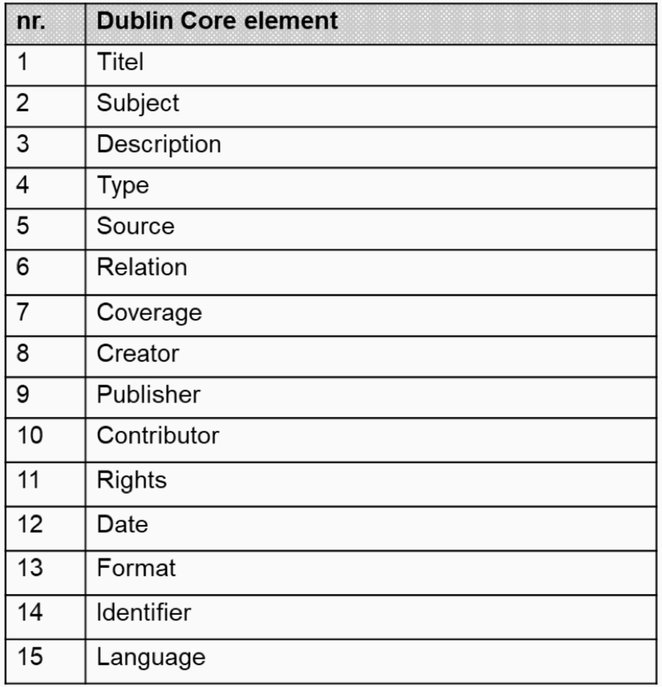
- International data **exchange** format
- 22 **elements** – 15 with an **ISO certificate**
- Refinements and encoding schemes for **subject-specification** applications



| nr. | Dublin Core element |
|---|---|
| 1 | Titel |
| 2 | Subject |
| 3 | Description |
| 4 | Type |
| 5 | Source |
| 6 | Relation |
| 7 | Coverage |
| 8 | Creator |
| 9 | Publisher |
| 10 | Contributor |
| 11 | Rights |
| 12 | Date |
| 13 | Format |
| 14 | Identifier |
| 15 | Language |

# Exercise

- Describe **yourself** using metadata (5 min.)
- Describe the **data** from your current research project with the help of the Dublin Core Metadata Initiative (10 min.)



| nr. | Dublin Core element |
|---|---|
| 1 | Titel |
| 2 | Subject |
| 3 | Description |
| 4 | Type |
| 5 | Source |
| 6 | Relation |
| 7 | Coverage |
| 8 | Creator |
| 9 | Publisher |
| 10 | Contributor |
| 11 | Rights |
| 12 | Date |
| 13 | Format |
| 14 | Identifier |
| 15 | Language |

# Outline

- **Introduction**
  - Metadata & metadata standards
  - **FAIR data principles**
- Planning
  - Data Management Plans (DMPs)
  - RDMO
  - Examples of DMPs
- Q&A
- Feedback

# FAIR data principles

- **Definition:** a concise and measurable set of principles that may act as a guideline for those wishing to enhance the reusability of their data holdings:
  - **Findability**
  - **Accessibility**
  - **Interoperability**
  - **Reusability**

- **Aims:**
  - Improving the **infrastructure** supporting the reuse of scholarly data
  - Enhancing the ability of **machines** to automatically find and use data
  - Supporting the reuse of data by **individuals**

[Wilkinson et al. 2016]

# To be Findable

- (Meta)data are assigned a globally unique and **persistent identifier**
- Data are described with rich **metadata**
- Metadata clearly and explicitly include the **identifier** of the data it describes
- (Meta)data are registered or indexed in a **searchable resource**

[Wilkinson et al. 2016]

# To be Accessible

- (Meta)data are retrievable by their identifier using a **standardized communications protocol** (e.g., http(s))
- The protocol is **open**, **free**, and **universally implementable**
- The protocol allows for an **authentication** and **authorization procedure**, where necessary
- **Metadata** are accessible, even when the data are no longer available

[Wilkinson et al. 2016, GO FAIR]

FAIR ≠ FOIR (O=Open)

# To be Interoperable

**Interoperability:** ‘each computer system at least has knowledge of the other system’s data exchange formats’

- (Meta)data use a formal, accessible, shared, and broadly applicable **language for knowledge representation** (e.g., controlled vocabularies/ontologies/thesauri, a good data model)
- (Meta)data use **vocabularies** that follow FAIR principles (e.g., using FAIR Data Point)
- (Meta)data include **qualified references** to other (meta)data (e.g., specifying if one datasets builds on another one, properly citing all datasets)

[Wilkinson et al. 2016, GO FAIR]

# To be Reusable

- Meta(data) are richly described with a plurality of accurate and relevant **attributes** (i.e. metadata that richly describes the context under which the data was generated such as the experimental protocols, the species used)
- (Meta)data are released with a clear and accessible **data usage license**
- (Meta)data are associated with detailed **provenance**

[Wilkinson et al. 2016, GO FAIR]

# FAIR or not FAIR?

**Case 1:** In chemistry, an international group from several universities is working closely with a company from the chemical industry to develop a new process. The findings from this collaboration are presented in overview articles and the data are stored and made accessible in a repository set up specifically for this project. The data is only accessible to people within the group, as several patent cases are pending. Externals can only access the metadata referencing this item.

[Bobrov et al. 2021]

# FAIR or not FAIR?

**Case 2:** For a master's thesis in geology soil samples are analyzed by machine. The student compiles extensive metadata for the publication of the data. To be sure that the correct version of the protocol of the analyzing machine is assumed, he takes a photo of the type plate and provides it under the item "protocol used".

[Bobrov et al. 2021]

# Outline

- ➔ Introduction
  - ◆ Metadata & metadata standards
  - ◆ FAIR data principles
- ➔ **Planning**
  - ◆ **Data Management Plans (DMPs)**
  - ◆ RDMO
  - ◆ Examples of DMPs
- ➔ Q&A
- ➔ Feedback

# Data management plan: Definition

What is a Data Management Plan?

Document that describes how data is handled during a research project and after the project is completed

*Wikipedia*

# Data management plan: Why?

- mandatory basis for common handling of research data
  - years of project duration + at least 10 years of storage
- coordination
- helps to avoid data loss and security holes
- compulsory in
  - Horizon Europe
  - DFG biodiversity research
  - DFG 101 Ancient Cultures: data intensive projects
  - BMBF and DFG: very often in educational projects
  - and at the University of Düsseldorf

# Data management plan: Why? Horizon Europe

Data Management Plan Template with 42 questions about

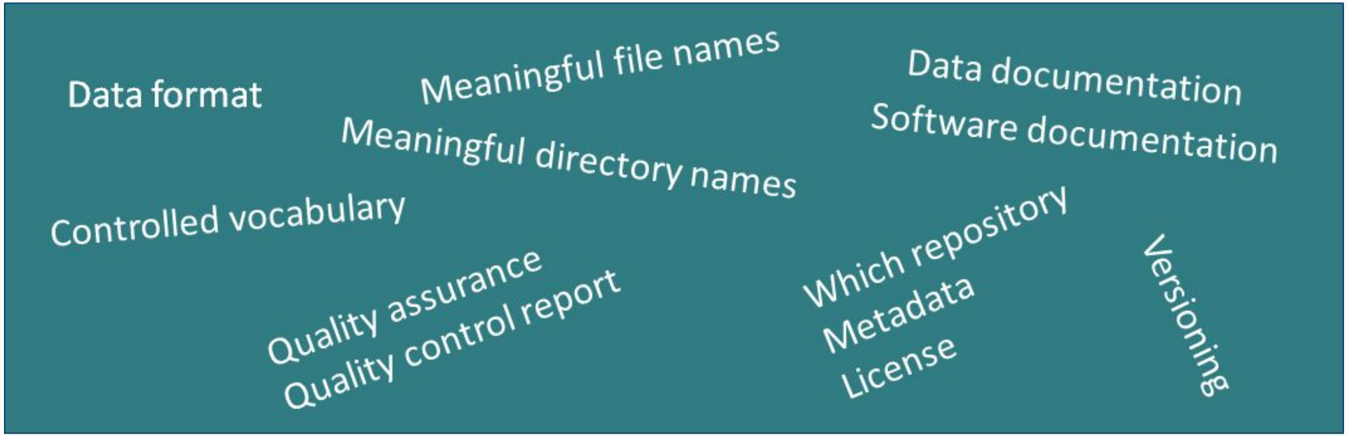
- the data
- the realization of the FAIR principles
- the costs for FAIR
- responsibility for research data management
- data security
- other research output
- other funders
- ethics

Deliverable

- for an initial data management plan at month 6 at the lastest
- living document
- evaluation takes place

# Data management plan: When?

Best before proposing the project or at the beginning



Data format

Meaningful file names

Meaningful directory names

Data documentation

Software documentation

Controlled vocabulary

Quality assurance

Quality control report

Which repository

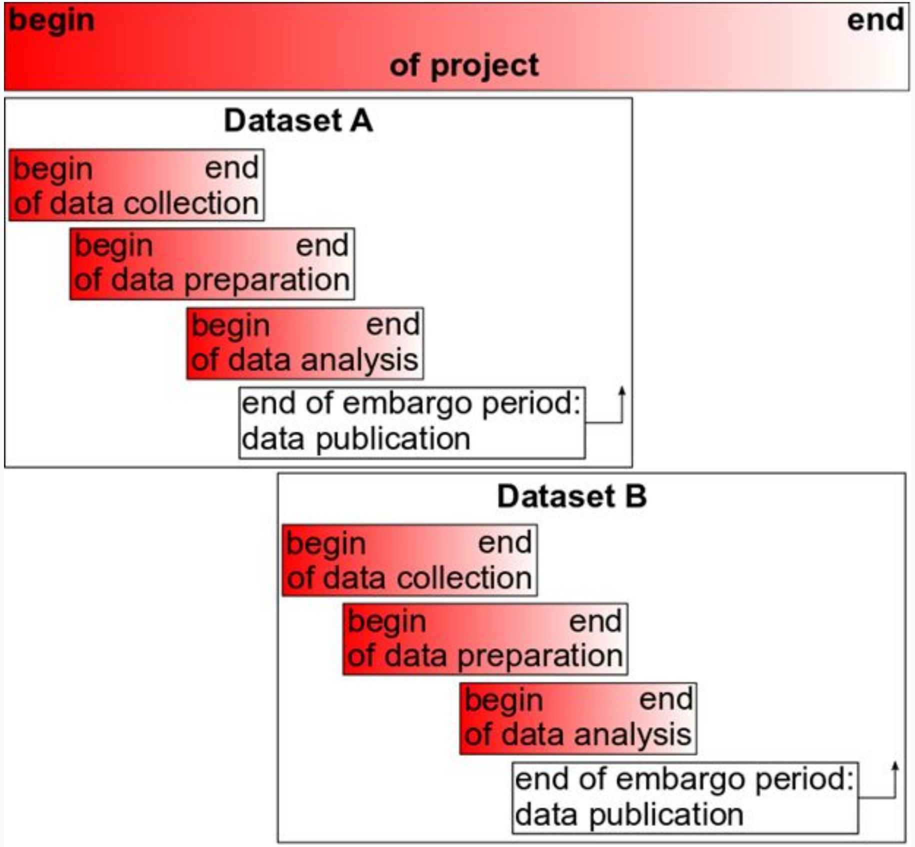
Metadata

License

Versioning

Avoid contradiction with the milestone and the cost plan

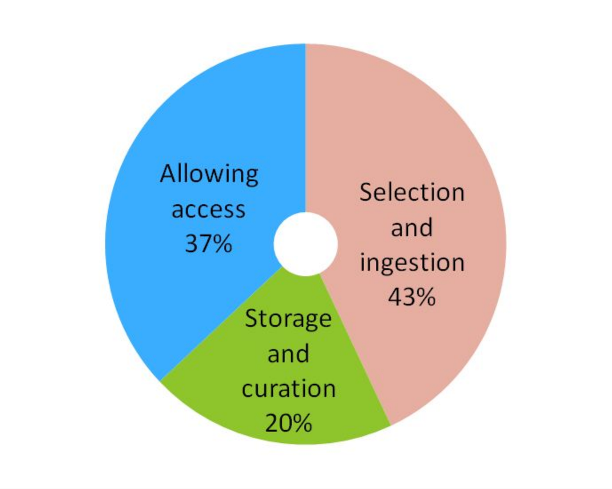
# Data management plan: + Milestone plan



begin end of project
Dataset A
begin end of data collection
begin end of data preparation
begin end of data analysis
end of embargo period: data publication
Dataset B
begin end of data collection
begin end of data preparation
begin end of data analysis
end of embargo period: data publication

# Data management plan: Costs

- Costs for research data management often reimbursed
- Labour costs much higher than material costs

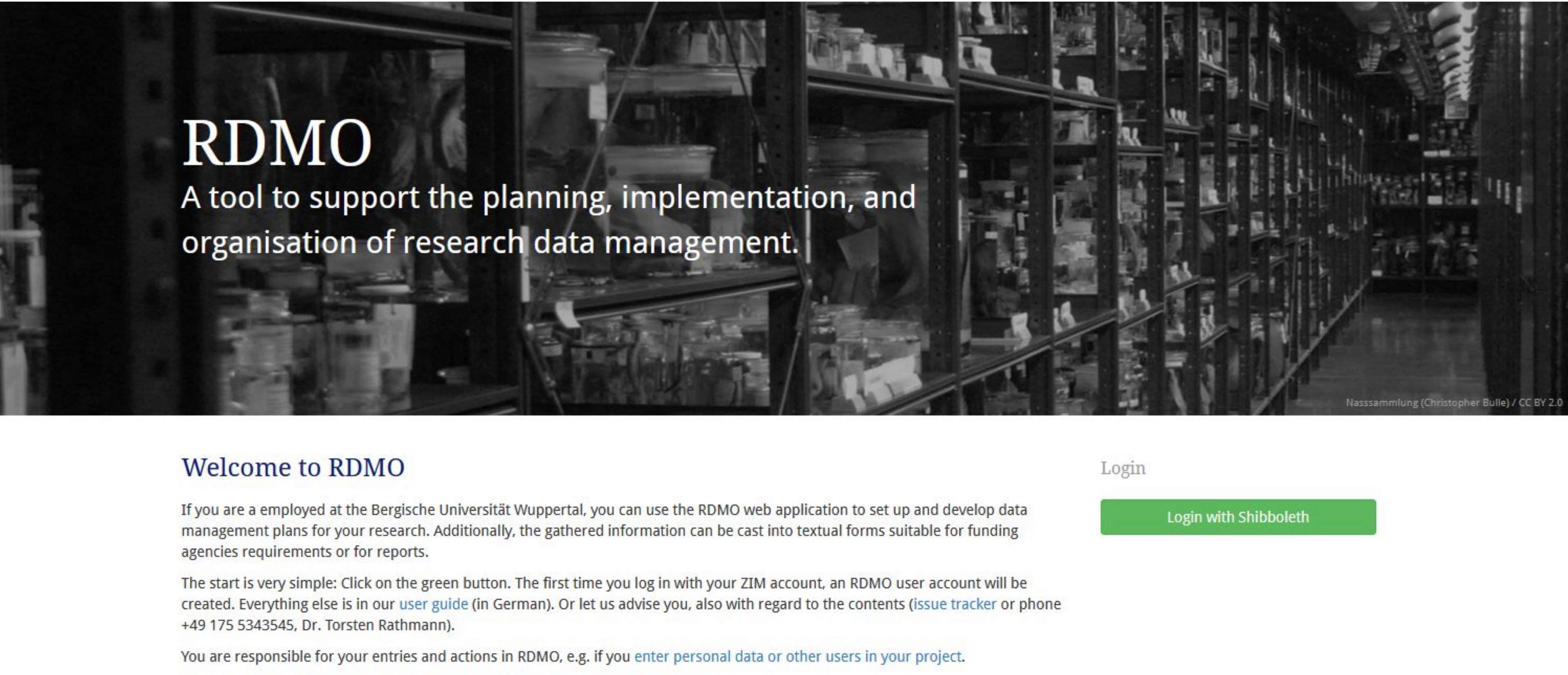


Cost structure from the view of the archive (result from Radieschen)

# Outline

- ➔ Introduction
  - ◆ Metadata & metadata standards
  - ◆ FAIR data principles
- ➔ **Planning**
  - ◆ Data Management Plans (DMPs)
  - ◆ **RDMO**
  - ◆ Examples of DMPs
- ➔ Q&A
- ➔ Feedback

# RDMO: Research Data Management Organiser



RDMO

A tool to support the planning, implementation, and organisation of research data management.

Nasssammlung (Christopher Bulle) / CC BY 2.0

## Welcome to RDMO

If you are a employed at the Bergische Universität Wuppertal, you can use the RDMO web application to set up and develop data management plans for your research. Additionally, the gathered information can be cast into textual forms suitable for funding agencies requirements or for reports.

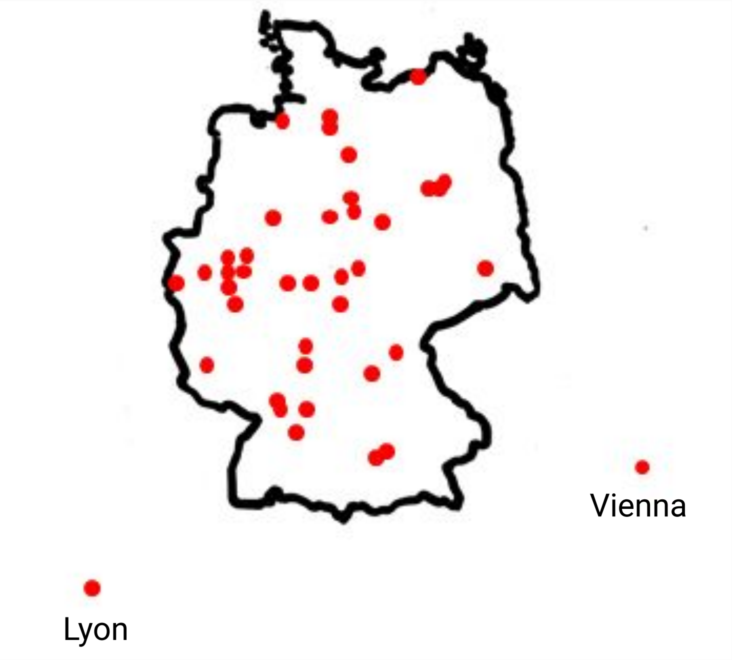
The start is very simple: Click on the green button. The first time you log in with your ZIM account, an RDMO user account will be created. Everything else is in our user guide (in German). Or let us advise you, also with regard to the contents (issue tracker or phone +49 175 5343545, Dr. Torsten Rathmann).

You are responsible for your entries and actions in RDMO, e.g. if you enter personal data or other users in your project.

# RDMO: Widespread in Germany

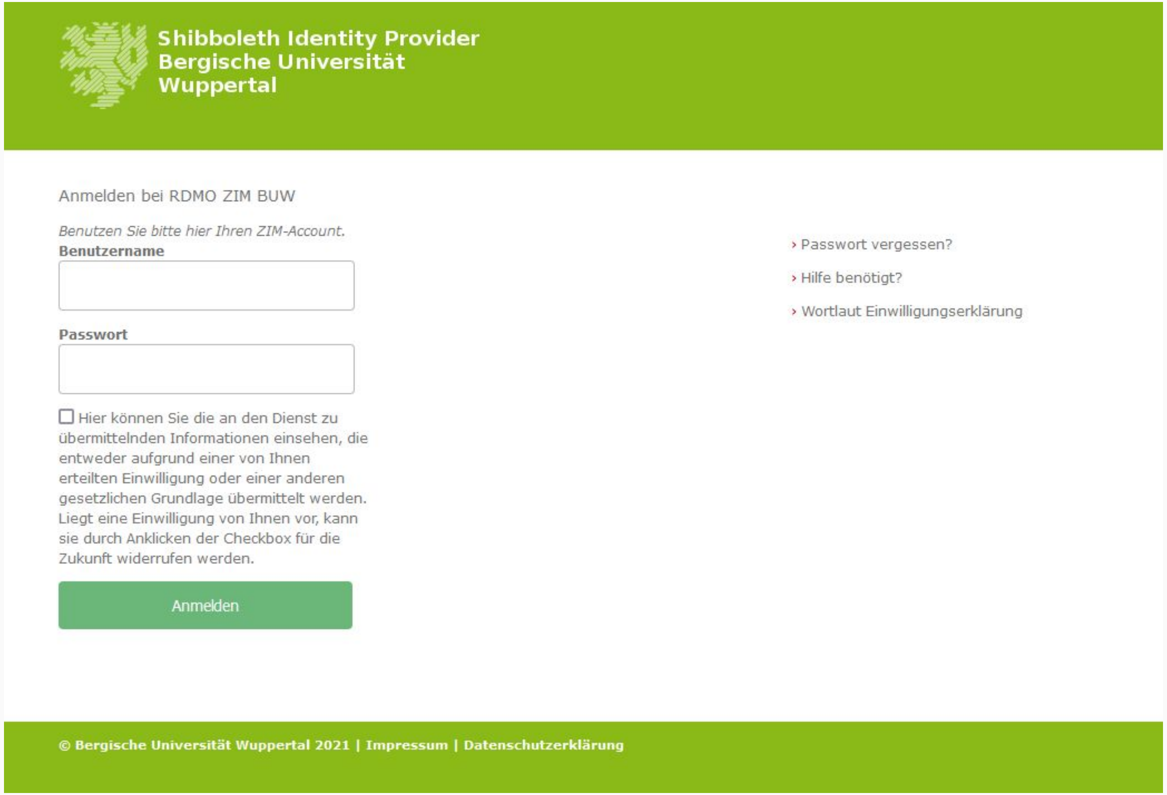
Available at the Universities of Düsseldorf, Siegen, Wuppertal and many other institutions

You can export your DMP as XML and re-import it, e.g. at your new institution.



# Software tool RDMO: Login

Login with your ZIM(T) username and password



Shibboleth Identity Provider
Bergische Universität
Wuppertal

Anmelden bei RDMO ZIM BUW

*Benutzen Sie bitte hier Ihren ZIM-Account.*

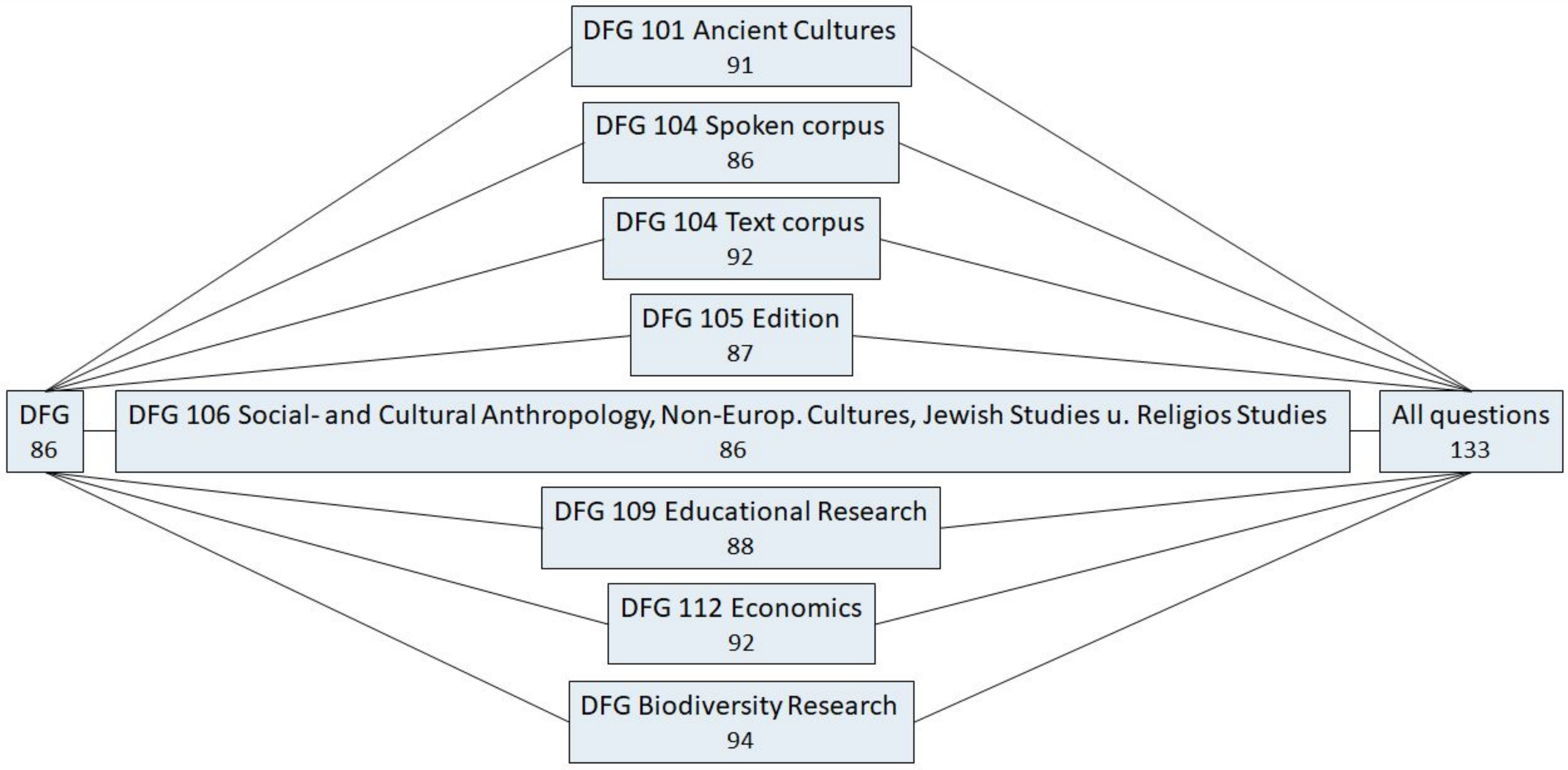
**Benutzername**

**Passwort**

□ Hier können Sie die an den Dienst zu übermittelnden Informationen einsehen, die entweder aufgrund einer von Ihnen erteilten Einwilligung oder einer anderen gesetzlichen Grundlage übermittelt werden. Liegt eine Einwilligung von Ihnen vor, kann sie durch Anklicken der Checkbox für die Zukunft widerrufen werden.

© Bergische Universität Wuppertal 2021 | Impressum | Datenschutzerklärung

# RDMO: FoDaKo question catalogs



DFG 101 Ancient Cultures
91

DFG 104 Spoken corpus
86

DFG 104 Text corpus
92

DFG 105 Edition
87

DFG
86

DFG 106 Social- and Cultural Anthropology, Non-Europ. Cultures, Jewish Studies u. Religios Studies
86

All questions
133

DFG 109 Educational Research
88

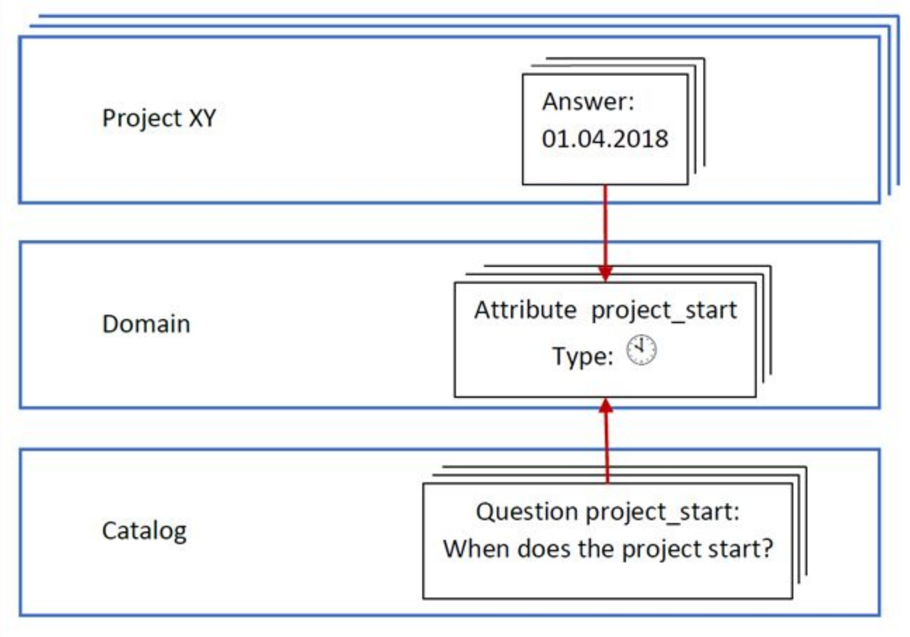
DFG 112 Economics
92

DFG Biodiversity Research
94

# RDMO: Question-answer link

All questions and answers (Q&A) are stored in a database

- Question catalogs can be changed without information loss
- If you change to a smaller question catalog, Q&A missing in that catalog are not shown
- To show them switch back or to catalog “All questions”

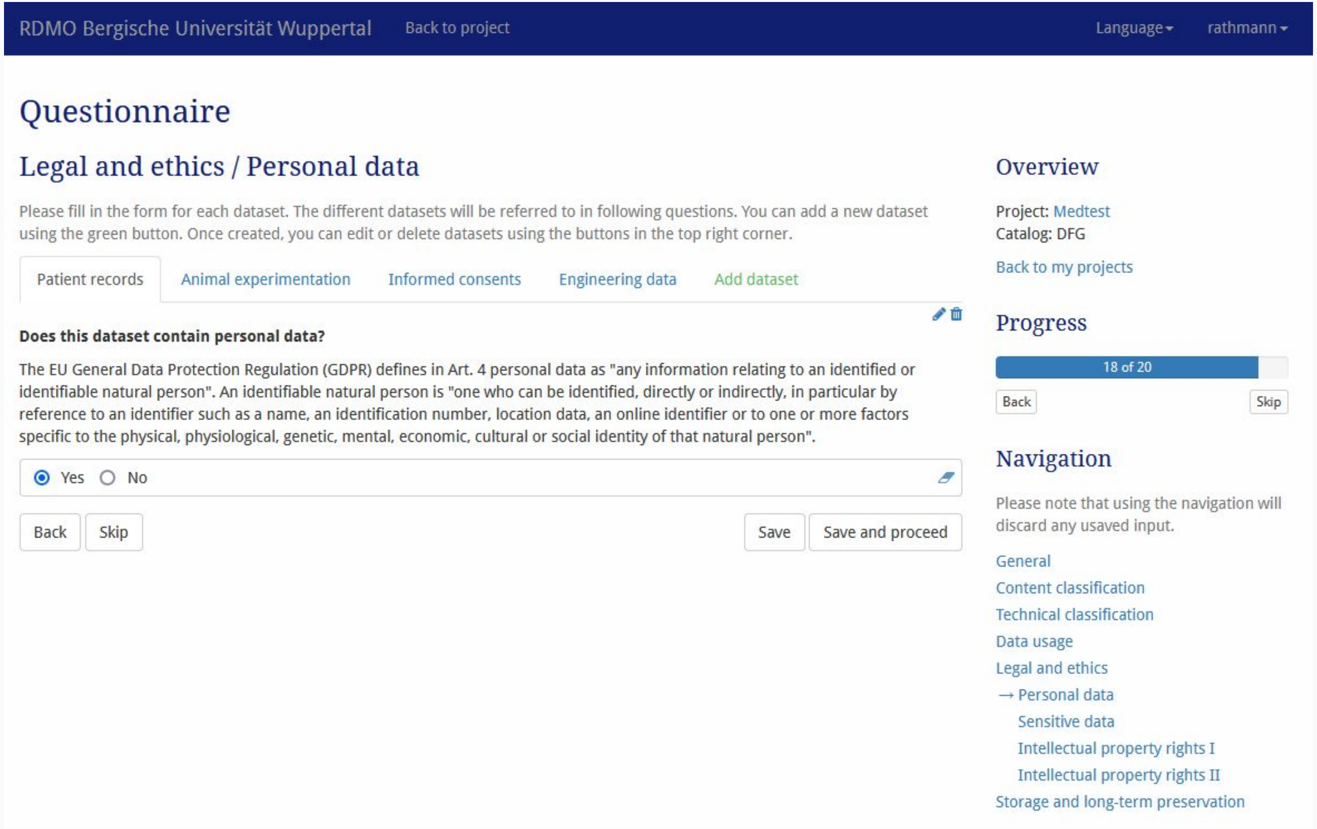


# RDMO: Answer questions

Many questions allow dataset-specific answers.

Help information

- specific if catalog is funder- or subject-specific



## Questionnaire

### Legal and ethics / Personal data

Please fill in the form for each dataset. The different datasets will be referred to in following questions. You can add a new dataset using the green button. Once created, you can edit or delete datasets using the buttons in the top right corner.

**Does this dataset contain personal data?**

The EU General Data Protection Regulation (GDPR) defines in Art. 4 personal data as "any information relating to an identified or identifiable natural person". An identifiable natural person is "one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".

● Yes ○ No

### Overview

Project: Medtest
Catalog: DFG

Back to my projects

### Progress

18 of 20

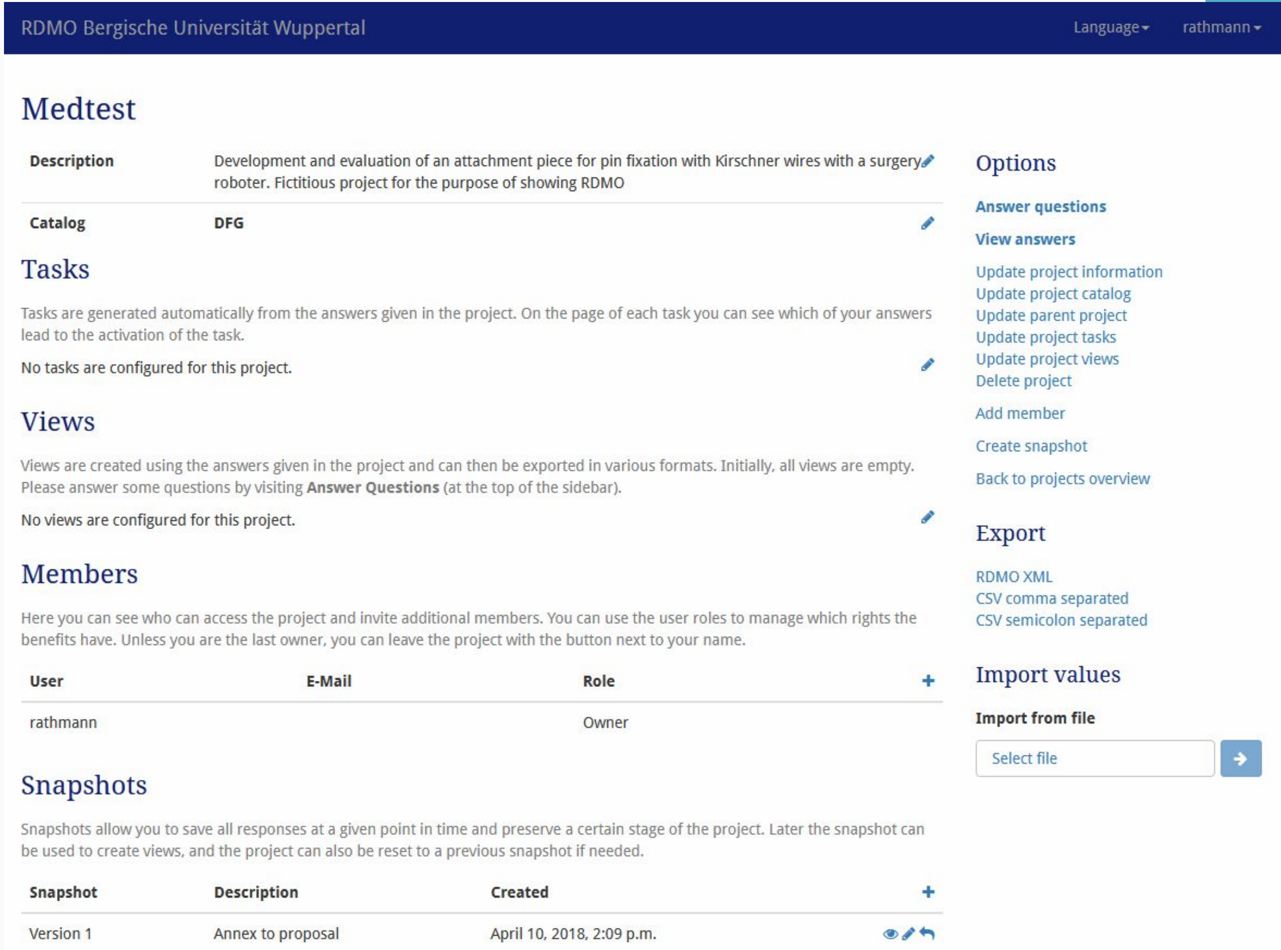
### Navigation

Please note that using the navigation will discard any usaved input.

- General
- Content classification
- Technical classification
- Data usage
- Legal and ethics
  - → Personal data
  - Sensitive data
  - Intellectual property rights I
  - Intellectual property rights II
- Storage and long-term preservation

# RDMO: More

- many export formats under “View answers”
  - Word
  - Open Office
  - LaTeX
  - ...
- collaborate work:
  - invitation via email
  - only inside your university
- snapshots



RDMO Bergische Universität Wuppertal

## Medtest

| | |
|---|---|
| **Description** | Development and evaluation of an attachment piece for pin fixation with Kirschner wires with a surgery roboter. Fictitious project for the purpose of showing RDMO |
| **Catalog** | **DFG** |

## 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 tasks are configured for this project.

## 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).

No views are configured for this project.

## Members

Here you can see who can access the project and invite additional members. You can use the user roles to manage which rights the benefits have. Unless you are the last owner, you can leave the project with the button next to your name.

| User | E-Mail | Role |
|---|---|---|
| rathmann | | Owner |

## Snapshots

Snapshots allow you to save all responses at a given point in time and preserve a certain stage of the project. Later the snapshot can be used to create views, and the project can also be reset to a previous snapshot if needed.

| Snapshot | Description | Created |
|---|---|---|
| Version 1 | Annex to proposal | April 10, 2018, 2:09 p.m. |

## Options

## Export

## Import values

**Import from file**

# Outline

- ➔ Introduction
  - ◆ Metadata & metadata standards
  - ◆ FAIR data principles
- ➔ **Planning**
  - ◆ Data Management Plans (DMPs)
  - ◆ RDMO
  - ◆ **Examples of DMPs**
- ➔ Q&A
- ➔ Feedback

# Examples of DMPs from the University of Minnesota

- **Roles** and **responsibilities** of project/institutional staff in the **management/retention** of data
- **Types** of data to be collected and shared
- Metadata **documentation**
- Data **preparation** for transformations/sharing/preservation and **format** of the final dataset
- Data **sharing** (prevention or agreement) and data **confidentiality**
- Method of data **access** (e.g. repository, archiving)
- **Expected schedule** for data access
- Data **secondary use** and associated **limitations**

# Outline

- ➔ Introduction
  - ◆ Metadata & metadata standards
  - ◆ FAIR data principles
- ➔ Planning
  - ◆ Data Management Plans (DMPs)
  - ◆ RDMO
  - ◆ Examples of DMPs
- ➔ **Q&A**
- ➔ Feedback

# Q&A

Photo by Jon Tyson on Unsplash

# Outline

- ➔ Introduction
  - ◆ Metadata & metadata standards
  - ◆ FAIR data principles
- ➔ Planning
  - ◆ Data Management Plans (DMPs)
  - ◆ RDMO
  - ◆ Examples of DMPs
- ➔ Q&A
- ➔ **Feedback**

# Feedback

Thank you for attending our webinar. We now would like to ask you to fill in a 4-question survey to improve our webinars. In advance, thank you for your help.

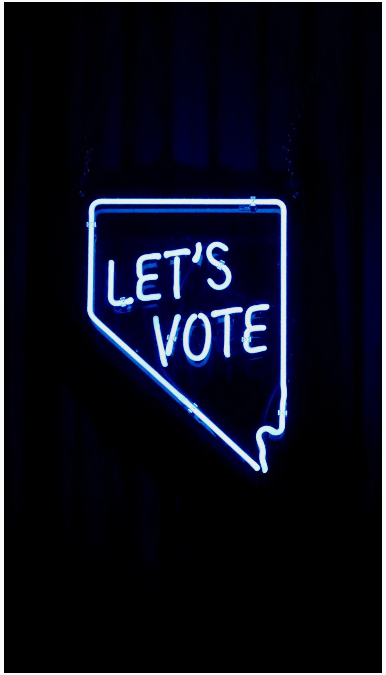


Photo by Manny Becerra
on Unsplash

# Thank you!

For further information we are at
your disposal

**ZB MED –
Information Centre for
Life Sciences**
Gleueler Straße 60
50931 Köln

forschungsdaten@zbmed.de
www.zbmed.de

**HHU Düsseldorf**
Universitätsstraße 1
40225 Düsseldorf

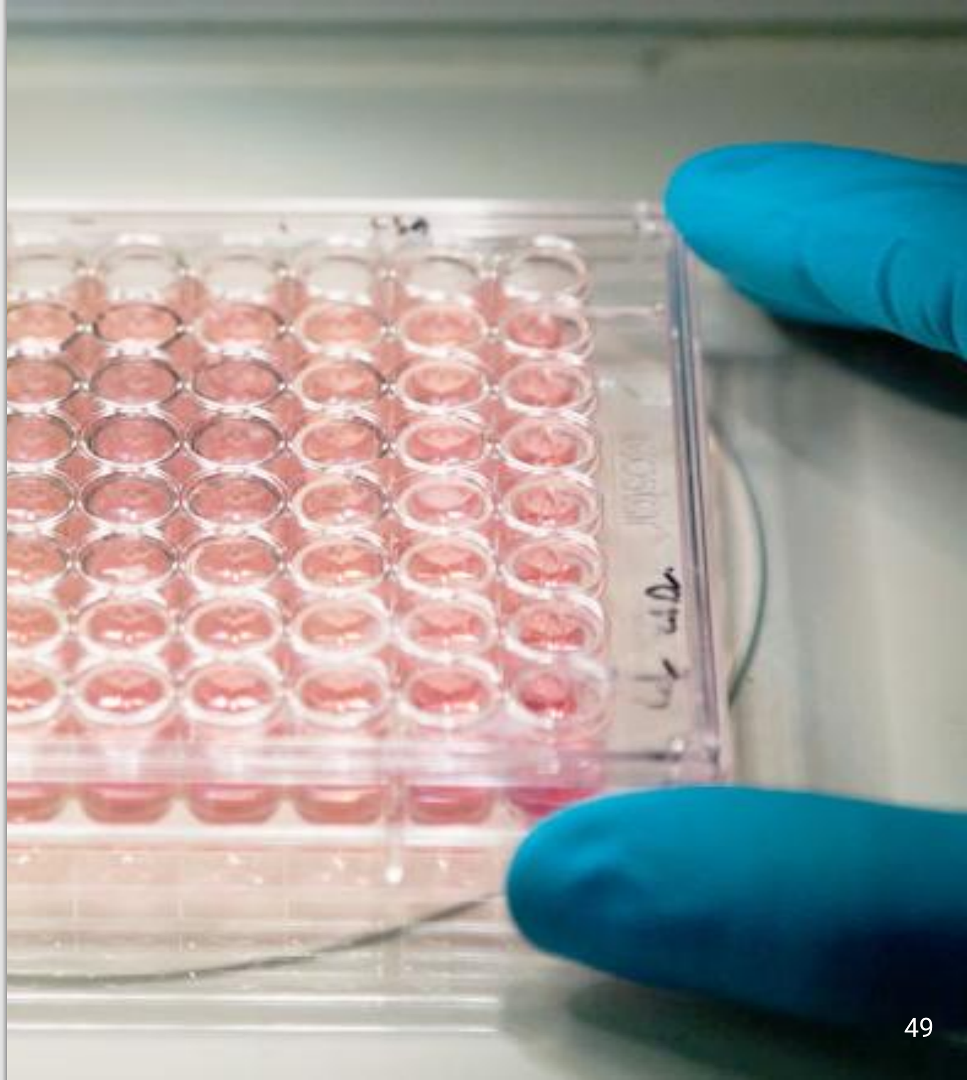
fdm@hhu.de
https://www.fdm.hhu.de

**Bergische Universität
Wuppertal –
Servicezentrum FDM**
Gaußstraße 20
42119 Wuppertal

fdm@uni-wuppertal.de
fdm.uni-wuppertal.de

**Universität Siegen –
e-Science-Service**

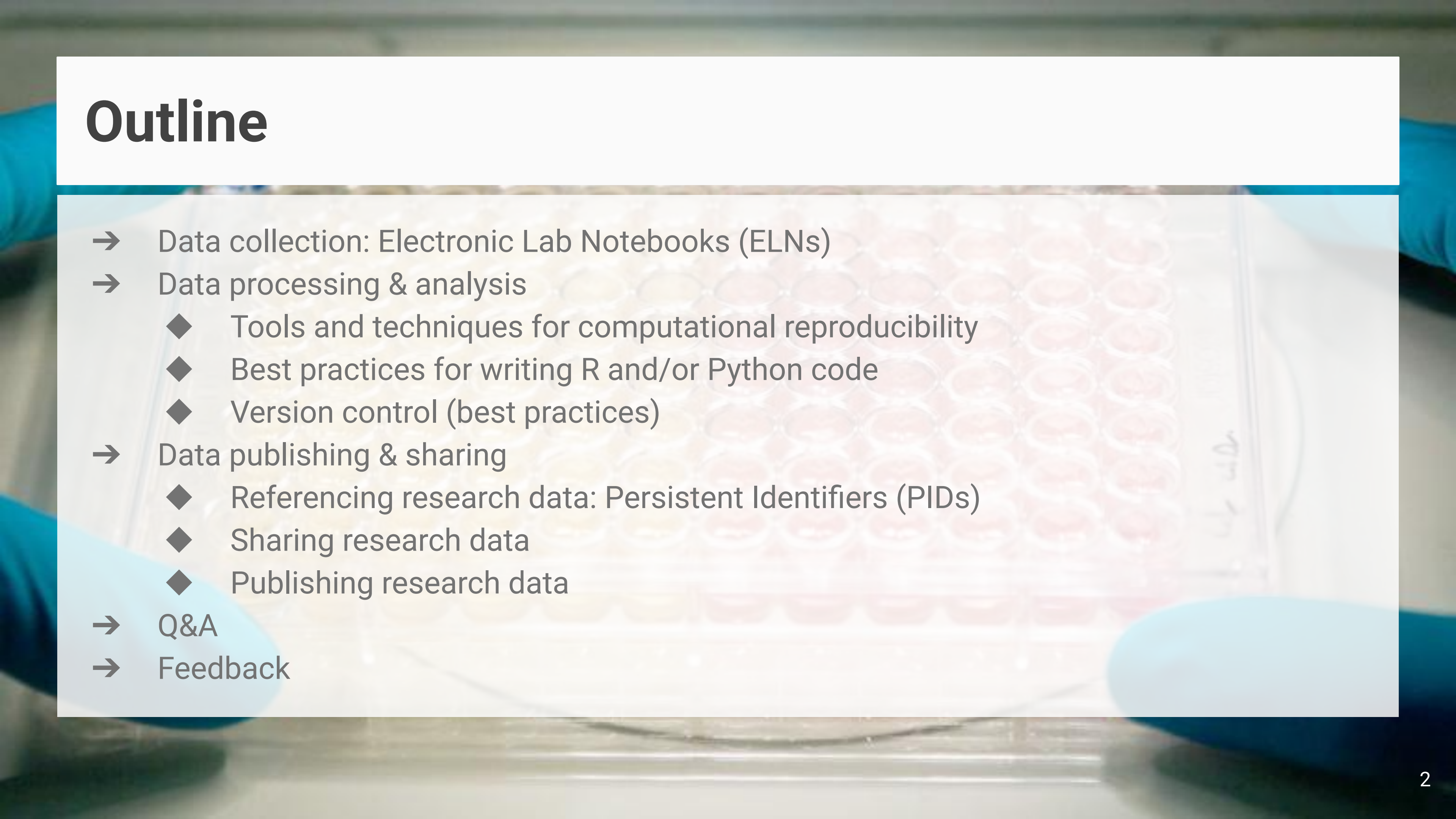
e-science-service@uni-siegen.de
https://e-science-service.uni-siegen.de/



# WORKSHOP ON RESEARCH DATA MANAGEMENT
# Data collection, processing and publishing

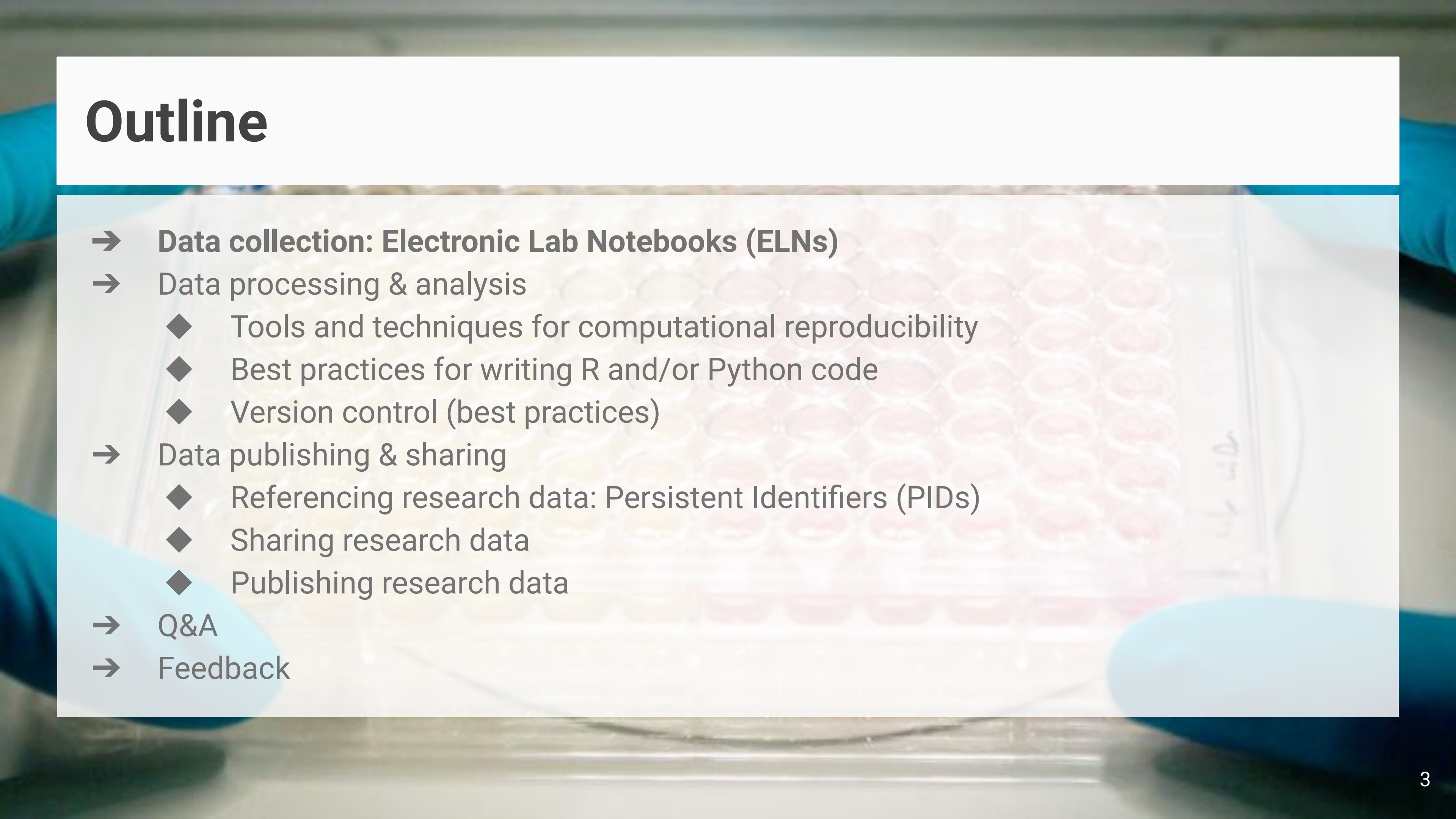
Torsten Rathmann, Servicezentrum Forschungsdatenmanagement Wuppertal
Daniela Kastrup, ULB Düsseldorf
Bastian Weiß, UB Siegen
B. Lindstädt, A. Shutsko & J. Vandendorpe, ZB MED - Information Centre for Life Sciences

Photo by ZB MED



# Outline

- Data collection: Electronic Lab Notebooks (ELNs)
- Data processing & analysis
  - Tools and techniques for computational reproducibility
  - Best practices for writing R and/or Python code
  - Version control (best practices)
- Data publishing & sharing
  - Referencing research data: Persistent Identifiers (PIDs)
  - Sharing research data
  - Publishing research data
- Q&A
- Feedback



# Outline

- ➔ **Data collection: Electronic Lab Notebooks (ELNs)**
- ➔ Data processing & analysis
  - ◆ Tools and techniques for computational reproducibility
  - ◆ Best practices for writing R and/or Python code
  - ◆ Version control (best practices)
- ➔ Data publishing & sharing
  - ◆ Referencing research data: Persistent Identifiers (PIDs)
  - ◆ Sharing research data
  - ◆ Publishing research data
- ➔ Q&A
- ➔ Feedback

# Electronic Lab Notebooks (ELNs) in the research data life cycle

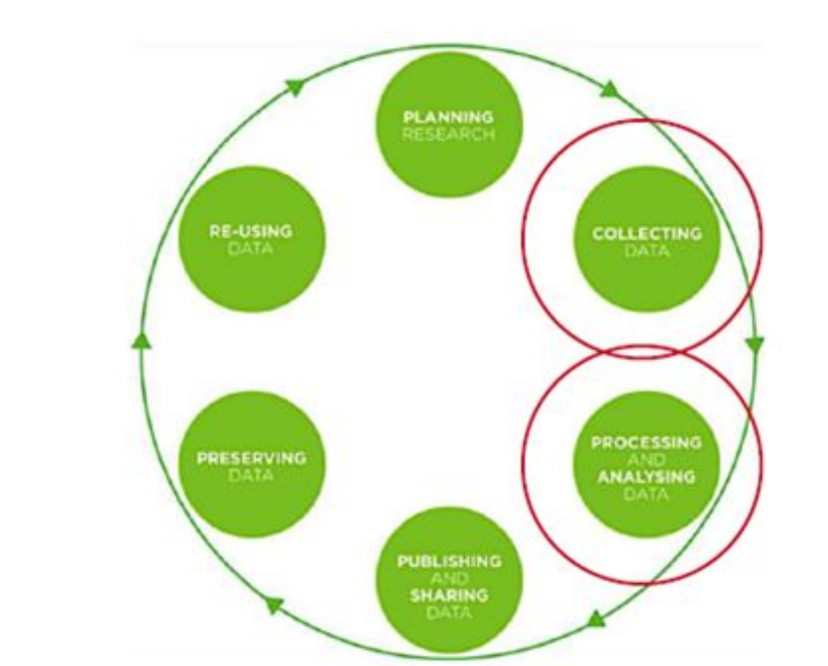


*Figure 1.* Data life cycle model (UK Data Archive).

[Mosconi et al. 2019]

# Electronic Lab Notebooks (ELNs)

A definition:

“software that helps researchers to document experiments, and that often has features such as protocol templates, collaboration tools, support for electronic signatures and the ability to manage the lab inventory“ [nature]

# Electronic Lab Notebooks (ELNs): benefits

- Create templates for logs, processes and workflows
- Save time by taking advantage of standardization
- Use search features and filters
- Log measurement results automatically
- Avoid the loss of information caused by illegible entries
- Structure and visualise processes and workflows
- Easily create backups
- Support in creating metadata
- Import and export functions
- Enables researchers to take their research work with them if they move to a different institute

[ZB Med]

# Electronic Lab Notebooks: types

| **Basic systems** | **Dedicated, commercial ELNs** | **High end systems** |
|---|---|---|
| Ability to enter text | All features from the basic systems | All features from the dedicated, commercial systems |
| Notes can be made available on multiple devices | Freehand drawing | Inventory management:<br>complete tracking of samples/reagents through all experiments |
| Attach files to notes | Complex rights management | Workflows for certain samples, tasks, experiments |
| Visualization of attachments in the note | Extensions/API for customization available | Direct link to laboratory equipment:<br>Automatic delivery of raw data by device<br>Delivery of metadata (e.g. date of last calibration) from device |
| Search within the written text | Inventory management:<br>Only amount and location of samples/ reagents | Analysis of raw data within the system |
| Possibly: Annotation of attachments, Search in attachments | 21CFR 11 compliance | Data mining (aggregate and cluster structured data) |

[Dirnagl/Przesdzing 2016]

# Example: Electronic Lab Notebook

eLabFTW

- Web application (Open Source Software)
- The service is available to all employees and the institutes and facilities of HHU
- Documentation of results, logging of work steps, no data deletion, immutability due to time stamp, searchability
- elabftw@hhu.de

[HHU_eLabFTW]

# Electronic Lab Notebook (ELN) guide

ELN guide

- **Content:** criteria for choosing an ELN
- **Target audience:**
  - Information infrastructures
  - Researchers
- **Languages:**
  - German
  - English

ELN-Wegweiser

Elektronische Laborbücher im Kontext von Forschungsdatenmanagement und guter wissenschaftlicher Praxis – ein Wegweiser für die Lebenswissenschaften

2. aktualisierte und erweiterte Fassung 2020

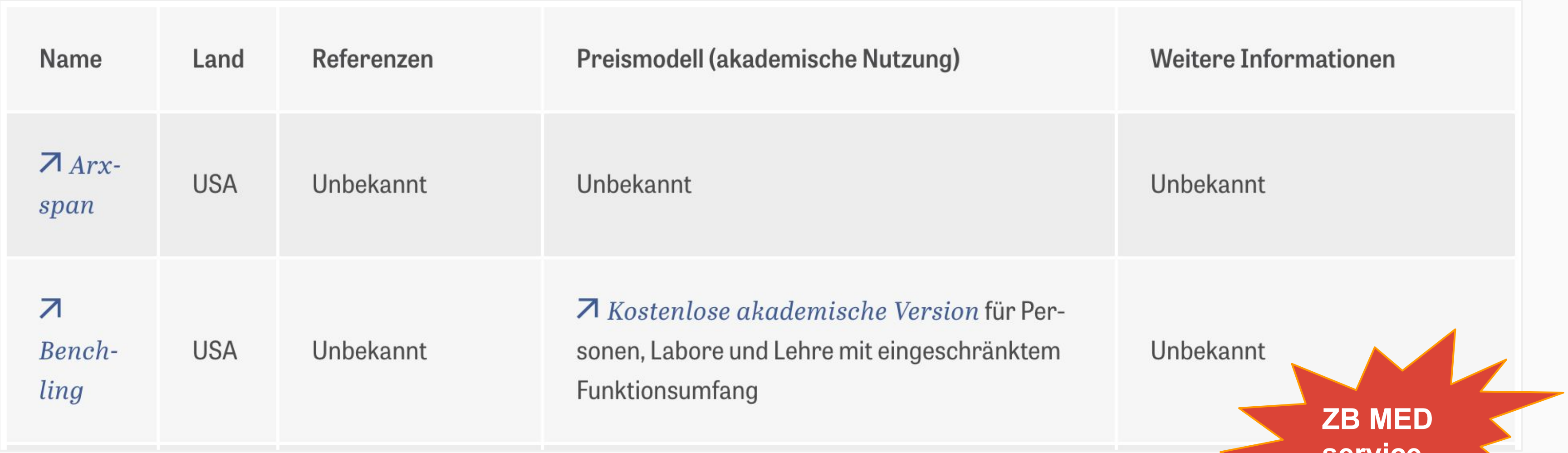
ZB MED-Publikationsportal Lebenswissenschaften

PUBLISSO

ZB MED service

# Electronic Lab Notebook (ELN) finder & filter

- **ELN finder:** interactive tool for **filtering** ELNs based on different criteria (under development in collaboration between ZB MED and HeFDI).
- **ELN filter** (in German only): step towards the ELN finder.

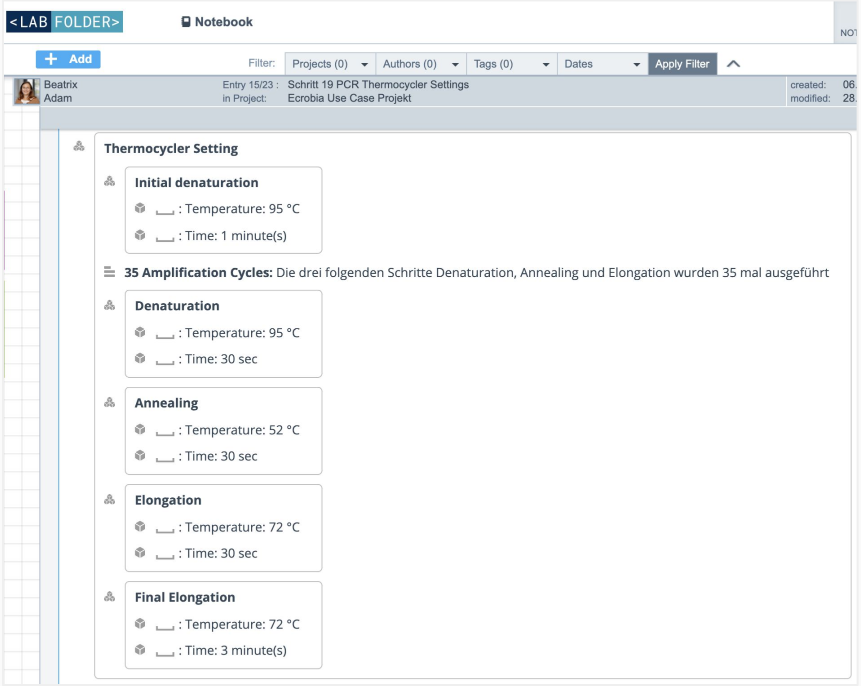


| Name | Land | Referenzen | Preismodell (akademische Nutzung) | Weitere Informationen |
|---|---|---|---|---|
| ↗ *Arxspan* | USA | Unbekannt | Unbekannt | Unbekannt |
| ↗ *Benchling* | USA | Unbekannt | ↗ *Kostenlose akademische Version* für Personen, Labore und Lehre mit eingeschränktem Funktionsumfang | Unbekannt |

ZB MED service

# Examples of ELNs in molecular biology

- eLABJournal
- LabCollector
- Labfolder
- LabWare ELN
- Limsophy LIMS



Beatrix Adam
Entry 15/23 : Schritt 19 PCR Thermocycler Settings
in Project: Ecrobia Use Case Projekt
created: 06.
modified: 28.

Thermocycler Setting

Initial denaturation
: Temperature: 95 °C
: Time: 1 minute(s)

35 Amplification Cycles: Die drei folgenden Schritte Denaturation, Annealing und Elongation wurden 35 mal ausgeführt

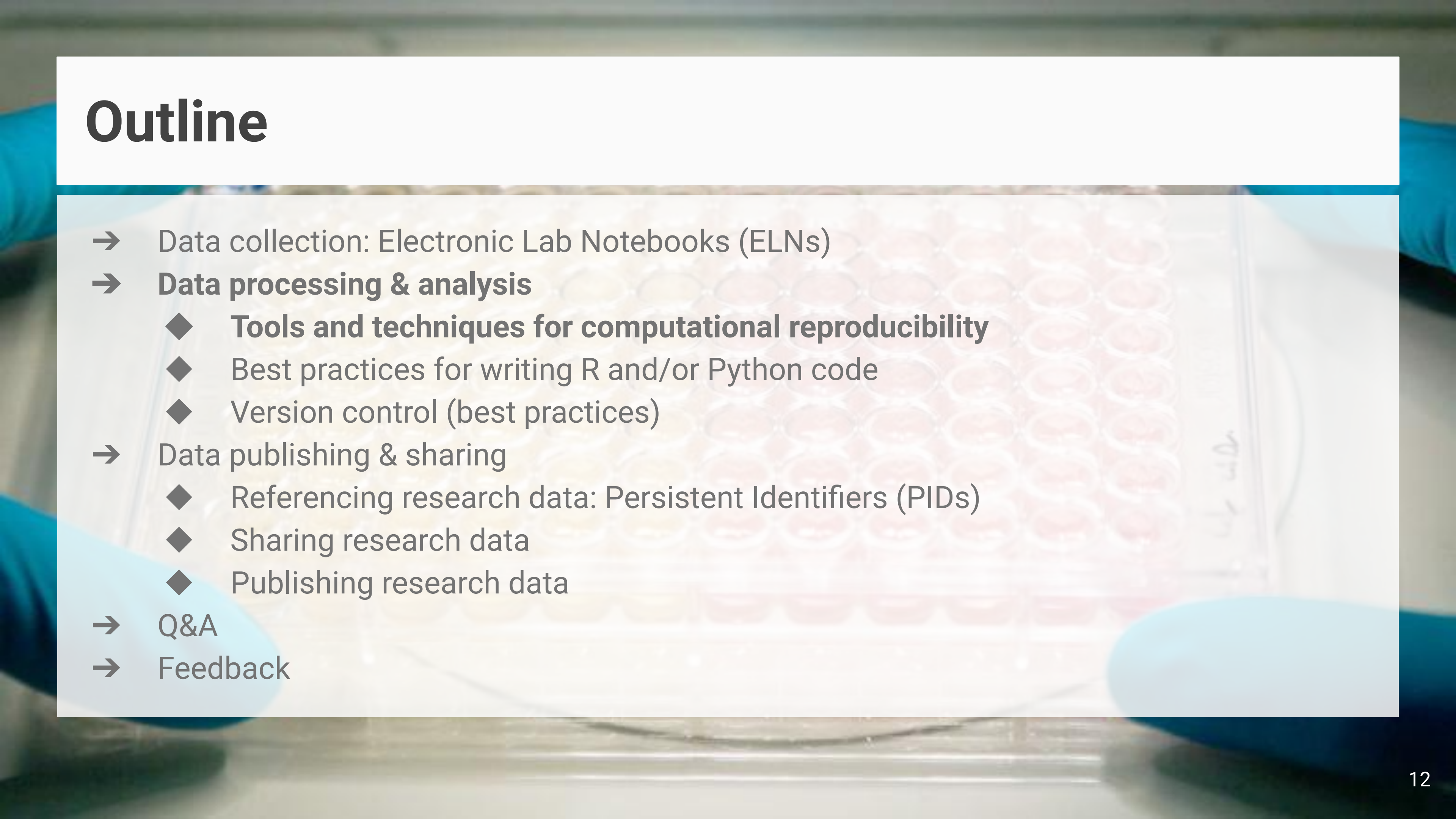
Denaturation
: Temperature: 95 °C
: Time: 30 sec

Annealing
: Temperature: 52 °C
: Time: 30 sec

Elongation
: Temperature: 72 °C
: Time: 30 sec

Final Elongation
: Temperature: 72 °C
: Time: 3 minute(s)

Labfolder

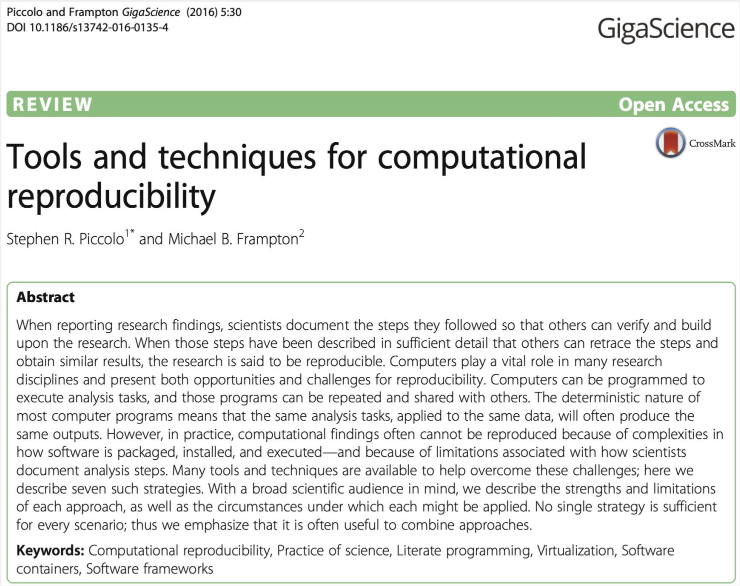


# Outline

- ➔ Data collection: Electronic Lab Notebooks (ELNs)
- ➔ **Data processing & analysis**
  - ◆ **Tools and techniques for computational reproducibility**
  - ◆ Best practices for writing R and/or Python code
  - ◆ Version control (best practices)
- ➔ Data publishing & sharing
  - ◆ Referencing research data: Persistent Identifiers (PIDs)
  - ◆ Sharing research data
  - ◆ Publishing research data
- ➔ Q&A
- ➔ Feedback

# Tools and techniques for computational reproducibility

**Reproducible research:** systematic investigation whose steps have been documented in sufficient detail that others can retrace the steps and obtain similar results



Piccolo and Frampton *GigaScience* (2016) 5:30
DOI 10.1186/s13742-016-0135-4

GigaScience

REVIEW Open Access

## Tools and techniques for computational reproducibility

Stephen R. Piccolo[1*] and Michael B. Frampton[2]

**Abstract**

When reporting research findings, scientists document the steps they followed so that others can verify and build upon the research. When those steps have been described in sufficient detail that others can retrace the steps and obtain similar results, the research is said to be reproducible. Computers play a vital role in many research disciplines and present both opportunities and challenges for reproducibility. Computers can be programmed to execute analysis tasks, and those programs can be repeated and shared with others. The deterministic nature of most computer programs means that the same analysis tasks, applied to the same data, will often produce the same outputs. However, in practice, computational findings often cannot be reproduced because of complexities in how software is packaged, installed, and executed—and because of limitations associated with how scientists document analysis steps. Many tools and techniques are available to help overcome these challenges; here we describe seven such strategies. With a broad scientific audience in mind, we describe the strengths and limitations of each approach, as well as the circumstances under which each might be applied. No single strategy is sufficient for every scenario; thus we emphasize that it is often useful to combine approaches.

**Keywords:** Computational reproducibility, Practice of science, Literate programming, Virtualization, Software containers, Software frameworks

[Piccolo and Frampton 2016]

# Tools and techniques for computational reproducibility

**Opportunities**

- Ability of computers to be **programmed** to execute analysis tasks
- Possibility to **repeat** and **share** programs
- **Same analysis tasks + same data** = **same outputs**

**Challenges**

- **Packaging**, **installation**, and **execution** of software = **complex**
- **Documentation** of analysis steps provided by scientists = **limited**

→ **Limited possibilities to reproduce computational findings**

[Piccolo and Frampton 2016]

# Tool/technique 1: narrative descriptions

- Narrative description = **detailed**, written **description** of computational analyses
- **Content**:
  - Operating system(s)
  - Software dependencies
  - Analytical software
  - Software version
  - Order
  - All non-default parameters
- **When?** Throughout the research process

# Tool/technique 2: custom scripts and code

- Using **text-based commands** via a command-line interface to **automate** research analyses, indicating:
  - **Software program(s)** to be executed
  - **Parameter(s)** to be used
- Compiling commands into **scripts** specifying the **order** in which they should be executed
  - Including commands for **installing** and **configuring software**
  - Documenting **software dependencies** and **input data**
- Creating **new software**
- **Publishing**/**storing** scripts and code:
  - Alongside a manuscript as **supplementary material**
  - In a **public repository** with a permanent URL
  - In a **Version Control System** (VCS)

# Tool/technique 2: custom scripts and code

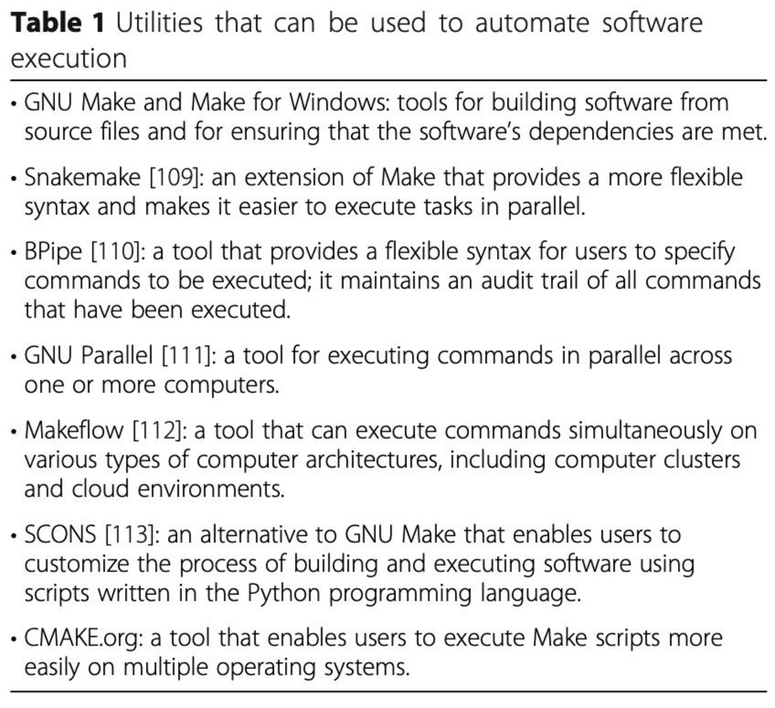


**Table 1** Utilities that can be used to automate software execution

- GNU Make and Make for Windows: tools for building software from source files and for ensuring that the software's dependencies are met.
- Snakemake [109]: an extension of Make that provides a more flexible syntax and makes it easier to execute tasks in parallel.
- BPipe [110]: a tool that provides a flexible syntax for users to specify commands to be executed; it maintains an audit trail of all commands that have been executed.
- GNU Parallel [111]: a tool for executing commands in parallel across one or more computers.
- Makeflow [112]: a tool that can execute commands simultaneously on various types of computer architectures, including computer clusters and cloud environments.
- SCONS [113]: an alternative to GNU Make that enables users to customize the process of building and executing software using scripts written in the Python programming language.
- CMAKE.org: a tool that enables users to execute Make scripts more easily on multiple operating systems.

[Piccolo and Frampton 2016]

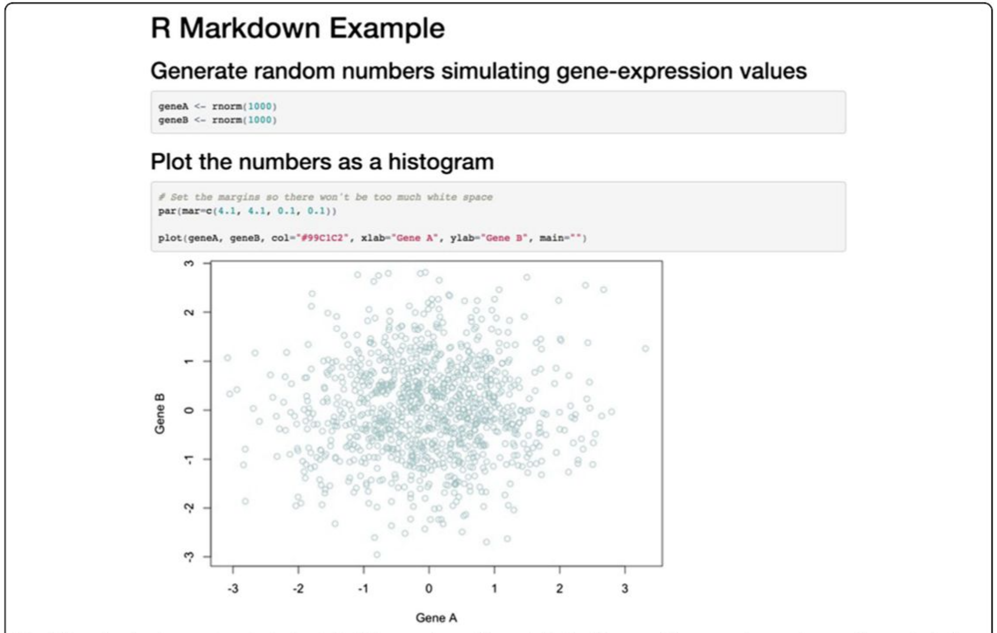
# Tool/technique 3: software frameworks

- **Software framework** = ‘abstraction in which software, providing generic functionality, can be selectively changed by additional user-written code, thus providing application-specific software’ [Wikipedia]
- Benefits of building on a pre-existing **software framework**: easily…
  - … accessing **software libraries**
  - … downloading and installing **software dependencies**
  - … ensuring that the **versions** are **compatible** with each other
- **Examples** of:
  - Software framework in biology: Bioconductor
  - General purpose tools for managing software dependencies: Apache Ivy, Puppet
  - Tools to make it easy to download and install previous versions of a software tool and dependencies: software container, virtual machines, aRchive project

# Tool/technique 4: literate programming

- **Literate programming** = code intermingled within a narrative of the scientific analysis
- When **executing** the code → generation of a **document** including:
  - Code
  - Narratives
  - Output (e.g. figures, plots)
- **Benefit:** reducing barriers of understanding → greater trust in computational findings
- **Examples of tools:** Jupyter, knitr

# Tool/technique 4: literate programming



## R Markdown Example

### Generate random numbers simulating gene-expression values

```
geneA <- rnorm(1000)
geneB <- rnorm(1000)
```

### Plot the numbers as a histogram

```
# Set the margins so there won't be too much white space
par(mar=c(4.1, 4.1, 0.1, 0.1))

plot(geneA, geneB, col="#99C1C2", xlab="Gene A", ylab="Gene B", main="")
```

**Fig. 4** Example of a document created using knitr. This example contains code (in the R language) for generating random numbers and plotting them on a graph. The knitr tool was used to generate the document, which combines the code and the output object (figure). See Additional file 4 for an executable version of this document

[Piccolo and Frampton 2016]

# Tool/technique 5: workflow management systems

- **Workflow** = series of commands resulting from using the output from one tool as input to additional tools.
- **Workflow management systems:**
  - Typically managed via a **graphical user interface**
  - Enabling scientists to **upload data** and **process** them using existing tools
  - Facilitating the **execution** of scientific software

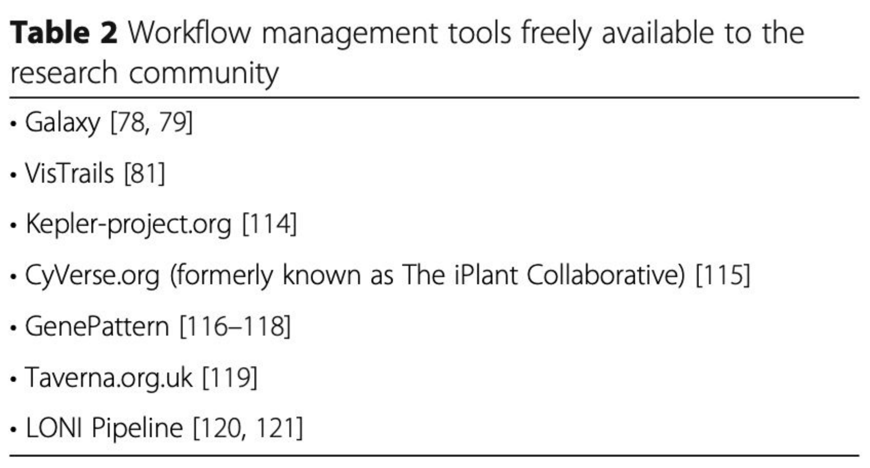


**Table 2** Workflow management tools freely available to the research community

- Galaxy [78, 79]
- VisTrails [81]
- Kepler-project.org [114]
- CyVerse.org (formerly known as The iPlant Collaborative) [115]
- GenePattern [116–118]
- Taverna.org.uk [119]
- LONI Pipeline [120, 121]

[Piccolo and Frampton 2016]

# Tool/technique 6: virtual machines

- Encapsulating everything necessary to execute a computational analysis:
  - **Operating system**
  - **Software**
  - **Scripts and code**
  - **Data**
- Benefits: can be…
  - … executed **anywhere**
  - … constrained to use specific amounts of **computational resources**
  - … exported to a single **binary file**

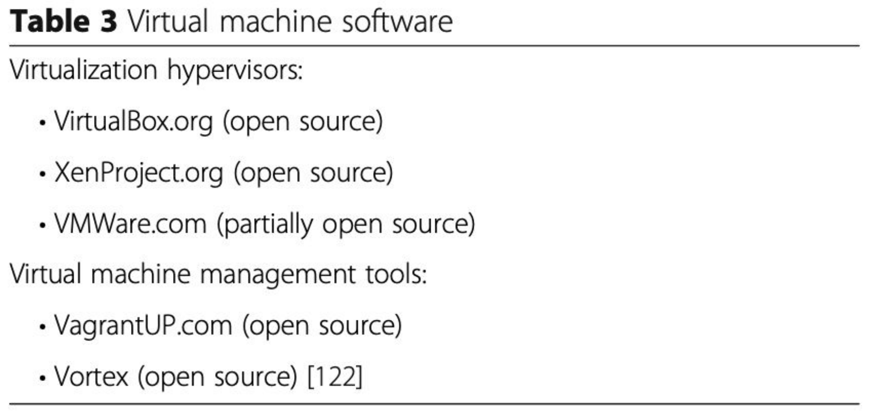


**Table 3** Virtual machine software

| |
|---|
| Virtualization hypervisors: |
| • VirtualBox.org (open source) |
| • XenProject.org (open source) |
| • VMWare.com (partially open source) |
| Virtual machine management tools: |
| • VagrantUP.com (open source) |
| • Vortex (open source) [122] |

[Piccolo and Frampton 2016]

# Tool/technique 7: software containers

- Encapsulating into a single package that can be shared:
  - **Operating system components**
  - **Scripts and code**
  - **Data**
- Benefits:
  - Easing the **installation** and **configuration** of dependencies
  - **Simultaneous execution** of multiple containers on a single computer
  - Containing **different** software versions and configurations

[Piccolo and Frampton 2016]

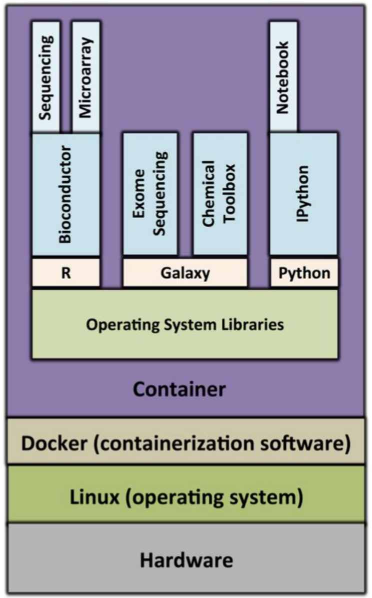
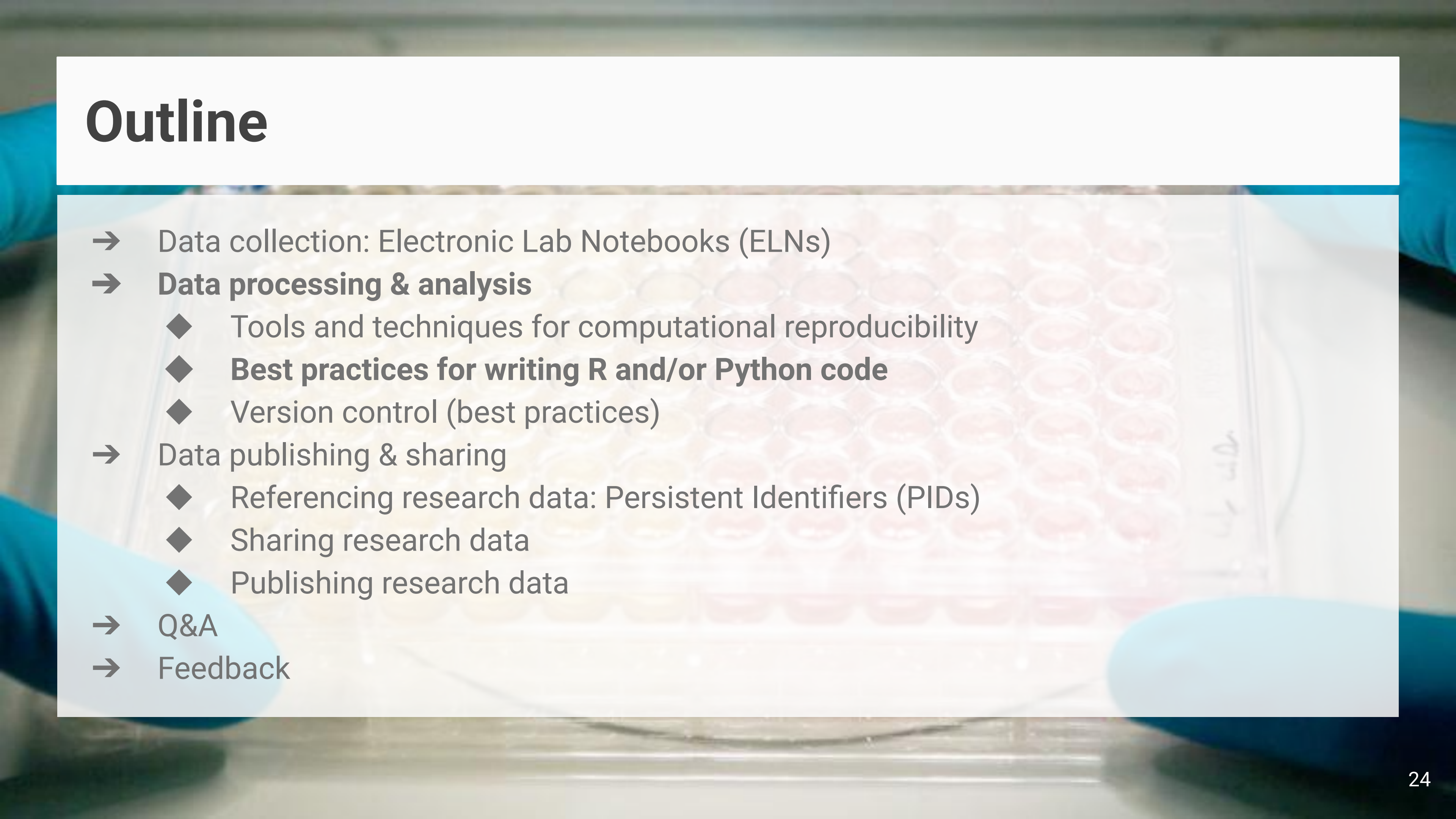


**Fig. 7** Example of a Docker container for genomics research. This container would enable researchers to preprocess various types of molecular data, using tools from Bioconductor and Galaxy, and to analyze the resulting data within a Jupyter notebook. Each box within the container represents a distinct Docker image. These images are layered such that some images depend on others (for example, the Bioconductor image depends on R). At its base, the container includes operating system libraries, which may not be present (or may be configured differently) on the computer's main operating system



# Outline

- Data collection: Electronic Lab Notebooks (ELNs)
- **Data processing & analysis**
  - Tools and techniques for computational reproducibility
  - **Best practices for writing R and/or Python code**
  - Version control (best practices)
- Data publishing & sharing
  - Referencing research data: Persistent Identifiers (PIDs)
  - Sharing research data
  - Publishing research data
- Q&A
- Feedback

# Best practices for writing R and/or Python code

- **Coding best practices:** ‘set of informal rules that the software development community employs to help improve software quality’ [Wikipedia].

- **Why?** Code is read much more often than it is written.

- **Benefits:**
    - Readability
    - Verifiability
    - Shareability
    - Reusability

# Best practices for writing R code

No widely accepted best practices.

**Resources:**

- Advanced R by Hadley Wickham
- Google’s R Style Guide
- QuantInsti
- R-bloggers
- Software carpentry

# Best practices for writing R code

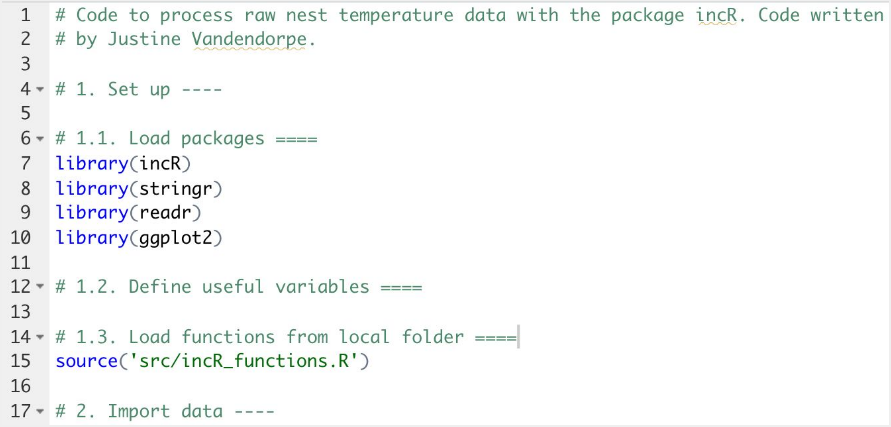
## Style

- Use a **consistent** style within your code.
- Agree on a common style up-front with your collaborators.
- Keep your code in **bite-sized chunks**.
- Use **snake case** (e.g. “my_variable”) and end data frame names by “_df”.
- Avoid naming variables after **base R functions** (e.g. “cat”).
- Variable name = **noun**; function name = **verb**.
- Use two spaces when **indenting** your code.
- **Modularize** your code and name the modules in ways that indicate the order in which they should be used (e.g. “1_...”).
- Use **sections** (1: ----, 2: ====, 3: ####).
- The code tells you “how”, **comments** tell you “why”.

# Best practices for writing R code

## Libraries

- Use “library()” instead of “**require()**” → Avoid using functions that change someone’s computer.
- Put all “library()” calls (and any hard-coded variables) at the **top of the script**.

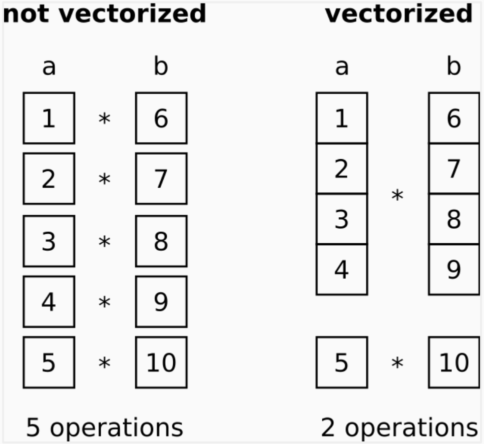


```r
# Code to process raw nest temperature data with the package incR. Code written
# by Justine Vandendorpe.

# 1. Set up ----

# 1.1. Load packages ====
library(incR)
library(stringr)
library(readr)
library(ggplot2)

# 1.2. Define useful variables ====

# 1.3. Load functions from local folder ====
source('src/incR_functions.R')

# 2. Import data ----
```

# Best practices for writing R code

## Loops

- Use **vectorized functions** instead of loops.
- Create a **matrix of 0s** with the right dimensions to hold the results of the loop instead of growing objects during the loop.



Multiplication vectorized and not vectorized
[WZB Data Science Blog]

# Best practices for writing R code

## Others

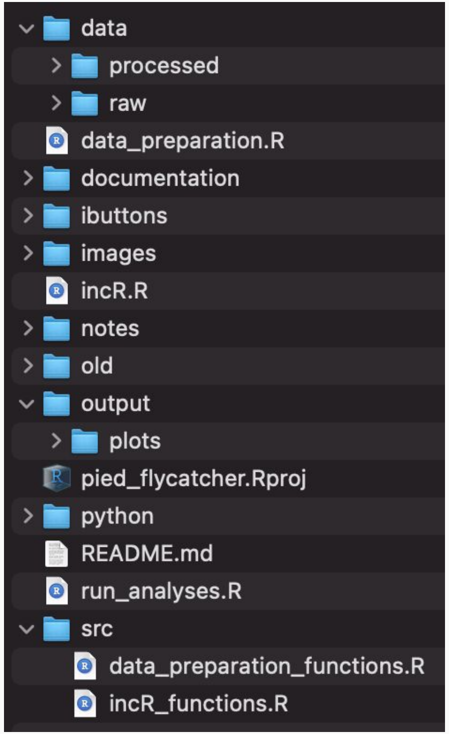
- Fill in objects with 0s instead of defining **empty variables**.
- Avoid saving your workspace and working directory (i.e. **.RData**).
- Use functions and the apply family to avoid **repetitions**.
- Avoid **storing** variables in the global environment.
- Save **figures** to files with R code.

# Best practices for writing R code

**RStudio project:** organisation of your scripts, data and output.

Structure:

- data (folder)
  - processed (folder)
  - raw (folder)
- output (folder)
  - plots (folder)
- name.R
- name.Rproj
- README.md
- src (folder)
  - functions.R



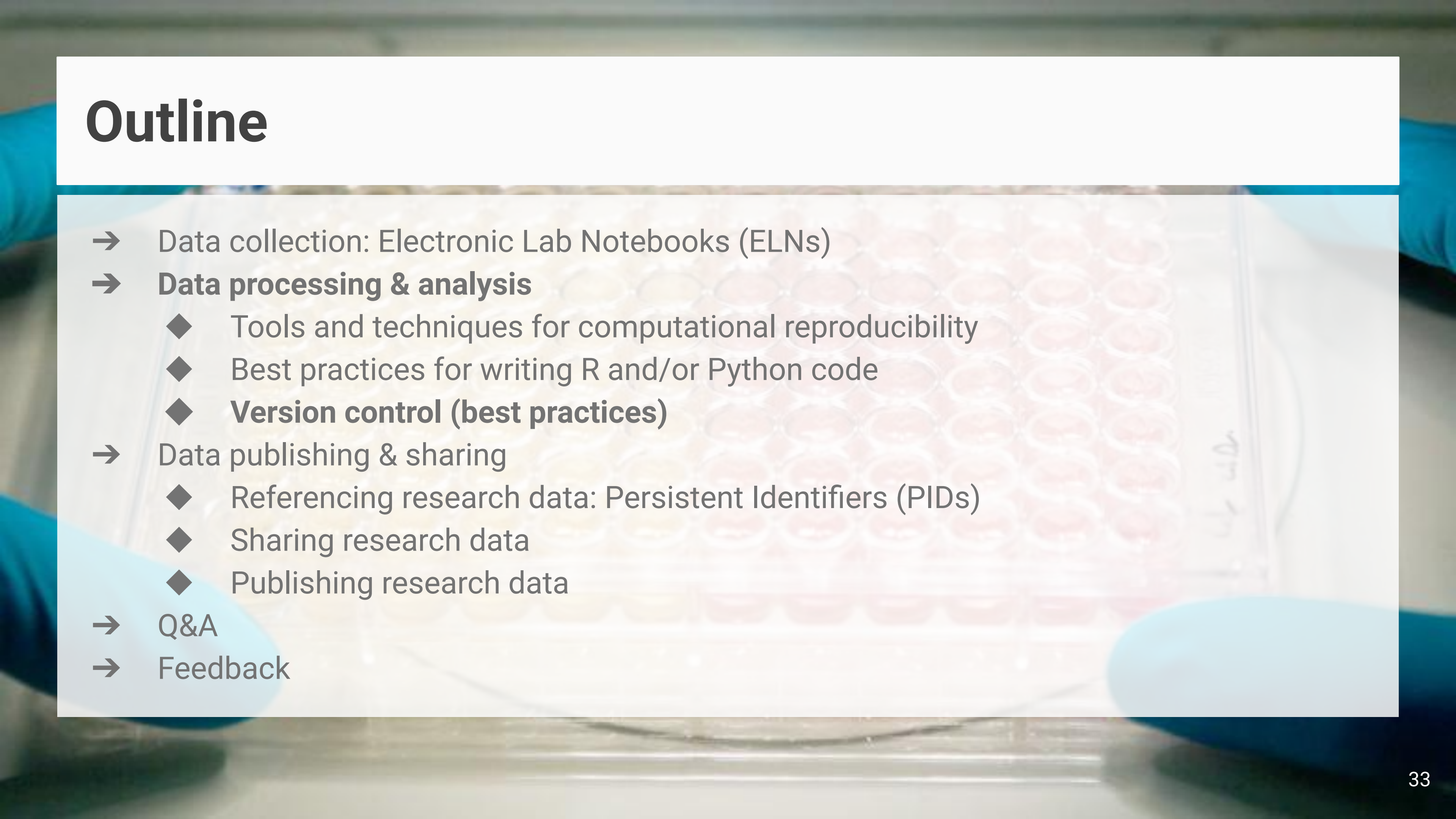
```
data
    processed
    raw
data_preparation.R
documentation
ibuttons
images
incR.R
notes
old
output
    plots
pied_flycatcher.Rproj
python
README.md
run_analyses.R
src
    data_preparation_functions.R
    incR_functions.R
```

# Best practices for writing Python code

Relatively complete set of Code Style guidelines and ‘Pythonic’ idioms.

**Resources:**

- PEP 8 -- Style Guide for Python Code
- The Hitchhiker’s Guide to Python!



# Outline

- ➔ Data collection: Electronic Lab Notebooks (ELNs)
- ➔ **Data processing & analysis**
  - ◆ Tools and techniques for computational reproducibility
  - ◆ Best practices for writing R and/or Python code
  - ◆ **Version control (best practices)**
- ➔ Data publishing & sharing
  - ◆ Referencing research data: Persistent Identifiers (PIDs)
  - ◆ Sharing research data
  - ◆ Publishing research data
- ➔ Q&A
- ➔ Feedback

# Version control

## Definitions

- **Version control:** practice of tracking and managing changes to a file (e.g. software code, lab protocol) or set of files over time so that you can recall specific versions later.
- **Version Control Systems (VCSs):** software tools that help teams manage changes to file(s) over time.
- **File tree:** folder structure in which files are arranged.
- **Branch:** independent stream of changes that can be merged back to the main branch, thus enabling work in parallel by separating work-in-progress from tested and stable code.

# Version Control Systems (VCSs)

| Version Control System (VCS) | Advantages | Drawbacks |
|---|---|---|
| **(Time-stamped) directories** | Simple | Error prone |
| **Local VCSs (e.g. RCS)** | Less error prone | Not easy to (1) collaborate with developers on other systems, (2) deal with local databases on every client |
| **Centralized VCSs (e.g. CVS, subversion)** | (1) Everyone knows what everyone else is doing, (2) admins have control over who can do what, (3) easy to admin | Centralized server = single point of failure |
| **Distributed VCSs (e.g. Git)** | (1) Every clone = full backup of all the data, (2) easy to collaborate | |

[Git]

# Version control

**Benefits of Version Control Systems (VCSs)**

- **Change history:**
  Keeping a complete long-term change history of every file, giving the possibility to go back to previous versions.

- **Branching & Merging:**
  Each team member may make their changes in several parts of the file tree, helping prevent concurrent work from conflicting.

- **Traceability & Annotation:**
  Being able to trace each change made to the file(s) and being able to annotate each change with a message describing the purpose and intent of the change.

# Version control best practices

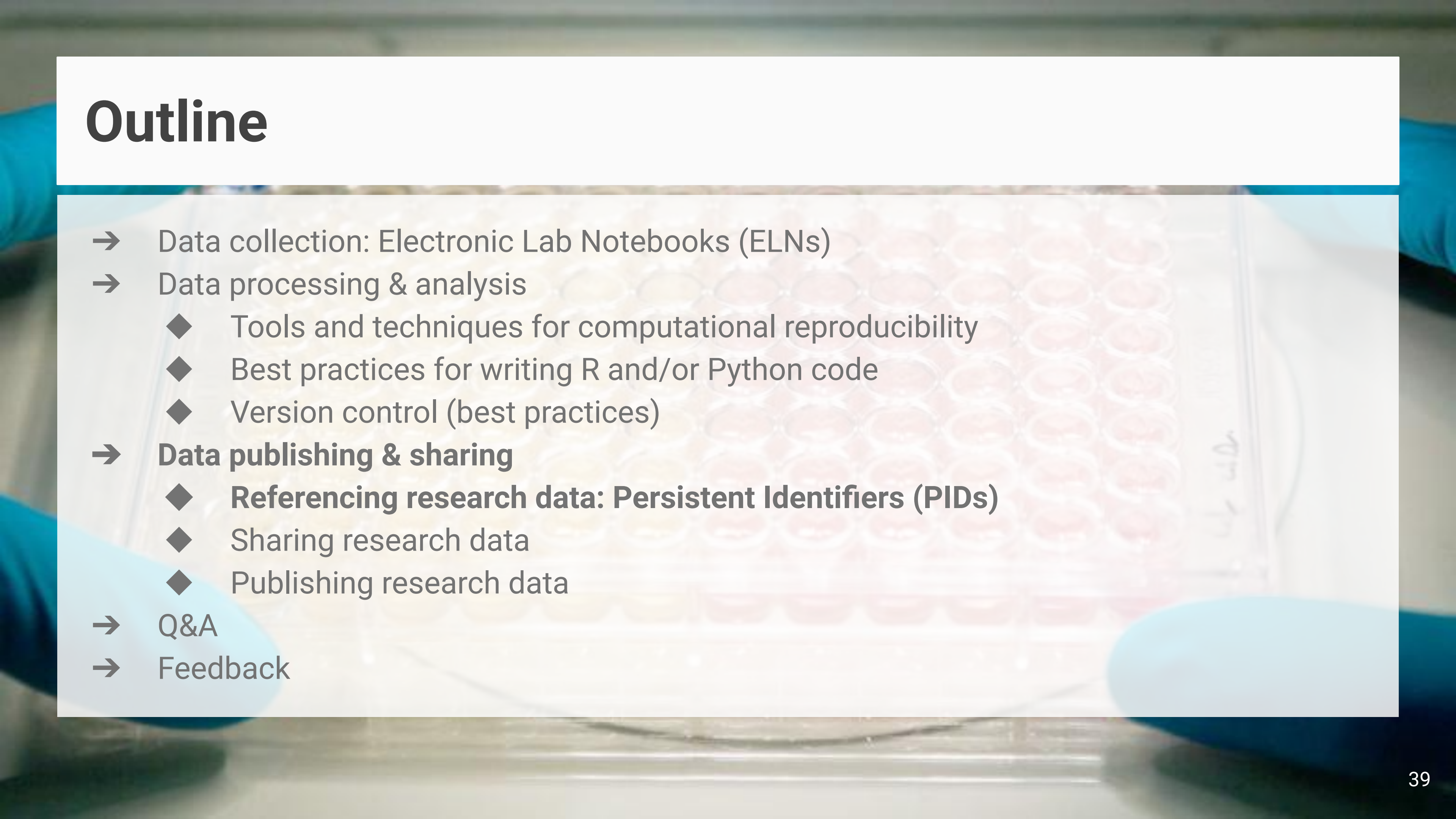
### Best practices

- Use a **Version Control System** (VCS)
- Break up commits into **related changes**
- Commit **early** and **often**
- Only commit **completed work**
- Avoid **breaking builds**
- **Test** and **review** before committing to a shared repository
- Write **descriptive commit messages**
- Incorporate **others' changes** frequently
- Ensure **traceability**
- Use **branches**
- Agree on a **workflow**
- Use **.gitignore** wisely

# Version control

Free courses on Version Control with Git

- Software Carpentry
- Udacity



# Outline

- ➔ Data collection: Electronic Lab Notebooks (ELNs)
- ➔ Data processing & analysis
  - ◆ Tools and techniques for computational reproducibility
  - ◆ Best practices for writing R and/or Python code
  - ◆ Version control (best practices)
- ➔ **Data publishing & sharing**
  - ◆ **Referencing research data: Persistent Identifiers (PIDs)**
  - ◆ Sharing research data
  - ◆ Publishing research data
- ➔ Q&A
- ➔ Feedback

# The Digital Object Identifier (DOI)

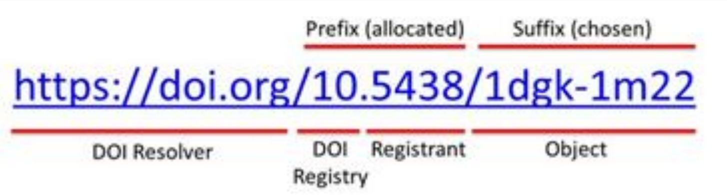
Persistent identifiers (PIDs): “a long-lasting reference to a digital resource” [ORCID]

Digital Object Identifier (DOI)

- unique and permanent identifier for digital objects
- DOIs identify the actual object, and not, like the URL, a current location

Advantage: Making research data accessible and citable in the long term

Citation capability is ensured by mandatory fields for metadata



Prefix (allocated) Suffix (chosen)

https://doi.org/10.5438/1dgk-1m22

DOI Resolver DOI Registry Registrant Object

[TIB Hannover]

# ZB MED’s Digital Object Identifier (DOI) Service

- Digital content that ZB MED can assign DOIs to:
  - **Research data** (e.g. observational data, statistical data, images, videos)
  - **Text publication** (e.g. journal articles, research reports, conference publications, posters)
- Target audience: **academic repositories**, **Open Access journals**

**ZB MED service**

# Other examples of PIDs

ORCID:

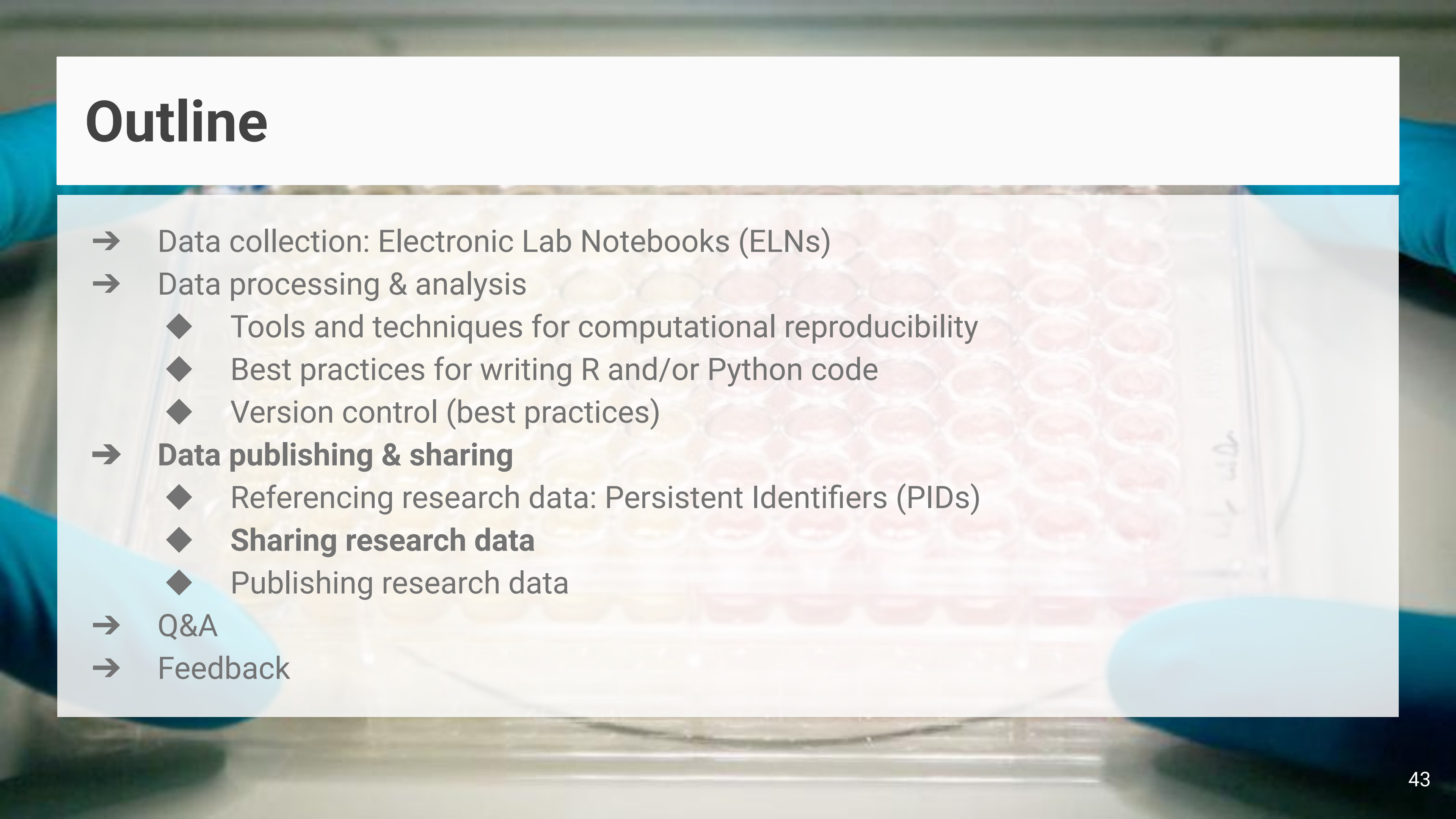
- Open Researcher Contributor Identification Initiative
- Unique identification for people
- Independent of name changes and institutional changes
- Self-registration

ROR:

- Research Organization Registry
- unique identifier for organization

Other persistent identifiers:

- Uniform Resource Name (URN), Research Activity Identifier (RAiD), International Geo Sample Number (IGSN)



# Outline

- ➔ Data collection: Electronic Lab Notebooks (ELNs)
- ➔ Data processing & analysis
  - ◆ Tools and techniques for computational reproducibility
  - ◆ Best practices for writing R and/or Python code
  - ◆ Version control (best practices)
- ➔ **Data publishing & sharing**
  - ◆ Referencing research data: Persistent Identifiers (PIDs)
  - ◆ **Sharing research data**
  - ◆ Publishing research data
- ➔ Q&A
- ➔ Feedback

# Data sharing

Data type options:

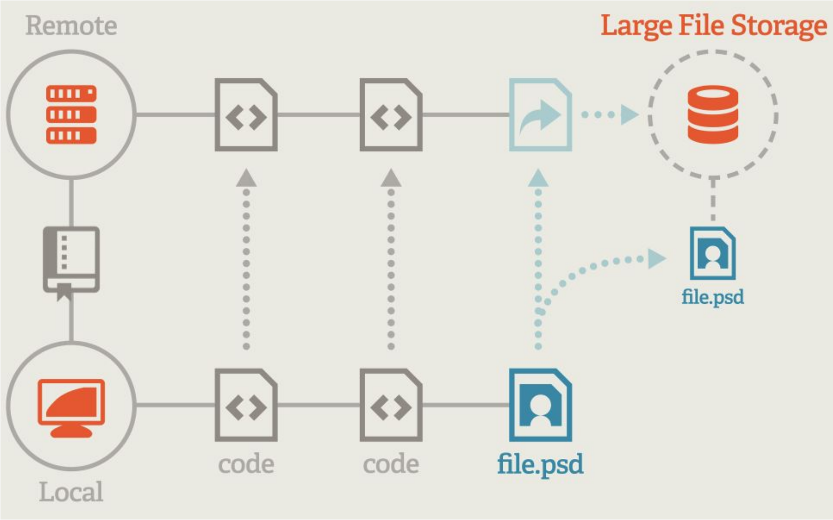
- Primary data
- Microdata
- Aggregate data
- Metadata, Semantic metadata

Data access options:

- Full access / Open access
- Partical access
- Access on request
- On-site data access
- Remote access

# General purpose collaboration tools

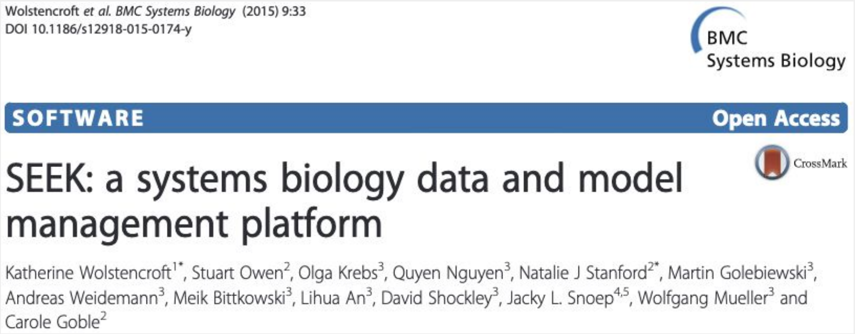
- **SharePoint:** web-based platform that integrate with Microsoft Office.
- **Git-based tools:**
  - **GitHub** providing hosting for software development and version control.
  - **GitLab** providing wiki, issue-tracking and a deployment platform.
  - **git-annex** ang **Git Large File Storage** providing file managing/versioning systems without checking the file contents into git.



Git Large File Storage

# Discipline-specific example of collaboration tool

SEEK: ‘web-based cataloguing and commons platform, for sharing heterogeneous scientific research datasets, models or simulations, processes and research outcomes’



Wolstencroft *et al. BMC Systems Biology* (2015) 9:33
DOI 10.1186/s12918-015-0174-y

BMC Systems Biology

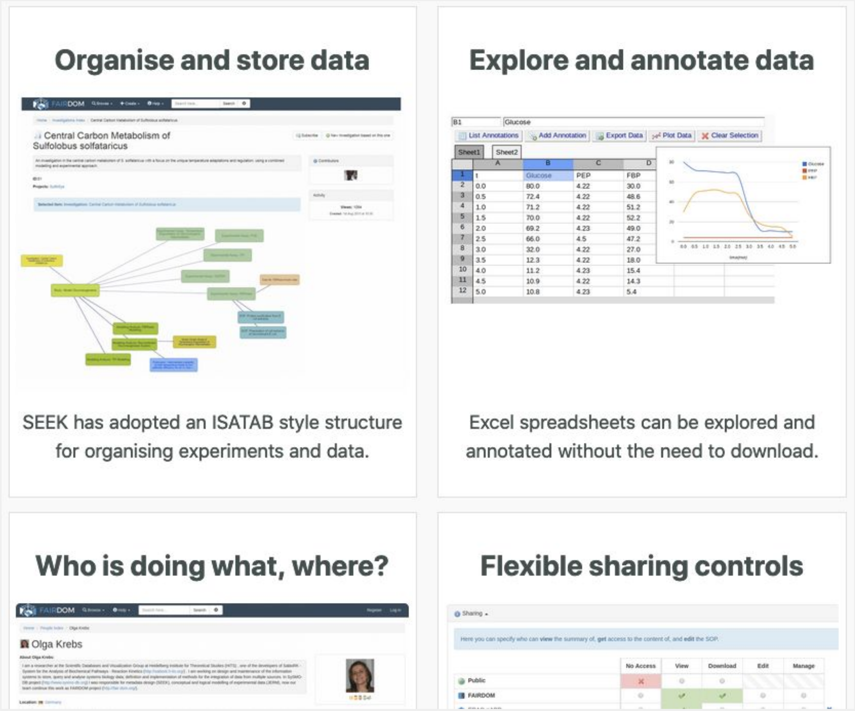
**SOFTWARE** **Open Access**

CrossMark

## SEEK: a systems biology data and model management platform

Katherine Wolstencroft[1*], Stuart Owen[2], Olga Krebs[3], Quyen Nguyen[3], Natalie J Stanford[2*], Martin Golebiewski[3], Andreas Weidemann[3], Meik Bittkowski[3], Lihua An[3], David Shockley[3], Jacky L. Snoep[4,5], Wolfgang Mueller[3] and Carole Goble[2]

[Wolstencroft *et al.* 2015]



**Organise and store data**

SEEK has adopted an ISATAB style structure for organising experiments and data.

**Explore and annotate data**

|  | A | B | C | D |
|---|---|---|---|---|
| 1 | t | Glucose | PEP | FBP |
| 2 | 0.0 | 80.0 | 4.22 | 30.0 |
| 3 | 0.5 | 72.4 | 4.22 | 48.6 |
| 4 | 1.0 | 71.2 | 4.22 | 51.2 |
| 5 | 1.5 | 70.0 | 4.22 | 52.2 |
| 6 | 2.0 | 69.2 | 4.23 | 49.0 |
| 7 | 2.5 | 66.0 | 4.5 | 47.2 |
| 8 | 3.0 | 32.0 | 4.22 | 27.0 |
| 9 | 3.5 | 12.3 | 4.22 | 18.0 |
| 10 | 4.0 | 11.2 | 4.23 | 15.4 |
| 11 | 4.5 | 10.9 | 4.22 | 14.3 |
| 12 | 5.0 | 10.8 | 4.23 | 5.4 |

Excel spreadsheets can be explored and annotated without the need to download.

**Who is doing what, where?**

**Flexible sharing controls**

Here you can specify who can **view** the summary of, **get** access to the content of, and **edit** the SOP.

|  | No Access | View | Download | Edit | Manage |
|---|---|---|---|---|---|
| Public | ✗ | ○ | ○ |  |  |
| FAIRDOM | ○ | ✓ | ✓ | ○ | ○ |

# Outline

- ➔ Data collection: Electronic Lab Notebooks (ELNs)
- ➔ Data processing & analysis
  - ◆ Tools and techniques for computational reproducibility
  - ◆ Best practices for writing R and/or Python code
  - ◆ Version control (best practices)
- ➔ **Data publishing & sharing**
  - ◆ Referencing research data: Persistent Identifiers (PIDs)
  - ◆ Sharing research data
  - ◆ **Publishing research data**
- ➔ Q&A
- ➔ Feedback

# Publication

Why is it important to publish your research data?

- Reusability
- Funding organization requirements
- Research data policy of an institution
- Possibility of meta-analyses
- Long-term availability
- Increased visibility, transparencey and accountability
- Clear citability (persistent identifiers)
- Data production as an independent scientific result

# Publication

Examples of funding organization requirements:

DFG:
Data should be made accessible at a stage of processing that allows it to be usefully reused by third parties (raw data or structured data)

EU:
Open Access to research data (as open as possible, as closed as necessary)

# Open Access

OA: Practice of providing online access to scientific information that is free of charge to the end-user

OA for articles: free online access for any user:

- Gold OA: the article is immediately published in open access mode
- Green OA: a free copy of the article is deposited in an online repository

OA for research data: right to access and reuse digital research data under the terms and conditions set out in the Grant Agreement

- Access and use free of charge
- Restricted access and/or use
  [European Comission]

# Publication models

- Research data can be published as an independent information object in a data repository
- As a data supplement in an enhanced publication: publication that is enriched with three categories of information: research data, extra materials, post-publication data [forschungsdaten.org]
- Documented in a data report: technical document that details whatever data you have collected and shows how it was analyzed [Chron]
- Documented in a data paper published in a data journal

# Data repositories

Storage locations for digital objects that make them available to a public or restricted group of users. Repositories can be distinguished:

- According to the type of objects to be stored (publications or research data)
- According to the domain of the contained data (institutional, technical or generic)
- According to the storage period of the data (e.g. 10 years to comply with the rules of good scientific practice, or permanently)
- According to the policies with which the data may be retrieved and reused [forschungsdaten.info]

# Data repositories

Examples of data repositories:

Institutional repository:

- HHU ResearchData: is available to all HHU employees and researchers free of charge. Each research group can publish their research data up to a limit of 1 TB with automatic DOI creation

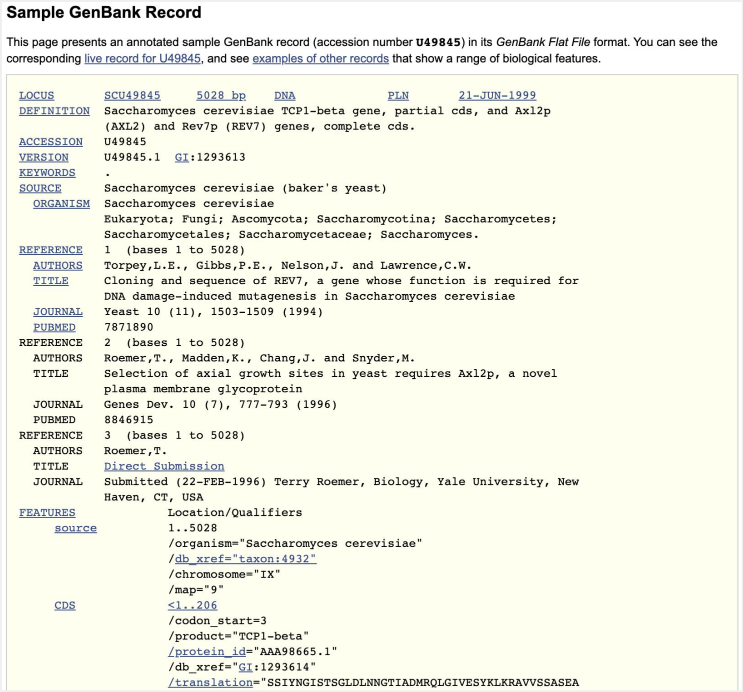
Interdisciplinary repositories:

- Zenodo: makes the sharing, curation and publication of data and software a reality for all researchers
- Figshare: repository where users can make all of their research outputs available in a citable, shareable and discoverable manner

# Examples of discipline-specific repositories

**GenBank**: ‘an annotated collection of all publicly available DNA sequences’

- **Submission tools**:
  - Web-based submission tools (BankIt, Submission Portal)
  - Submission preparation tools (tbl2asn, Genome Workbench)
- **Submission types**:
  - mRNA or genomic sequence data
  - Complete Microbial Genomes
  - Whole Genome Shotgun (WGS) Sequences
  - …



**Sample GenBank Record**

This page presents an annotated sample GenBank record (accession number **U49845**) in its *GenBank Flat File* format. You can see the corresponding live record for U49845, and see examples of other records that show a range of biological features.

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
```

Annotated sample GenBank record for a *Saccharomyces cerevisiae* gene

# Examples of discipline-specific repositories

ZB MED's Repository for Life Sciences

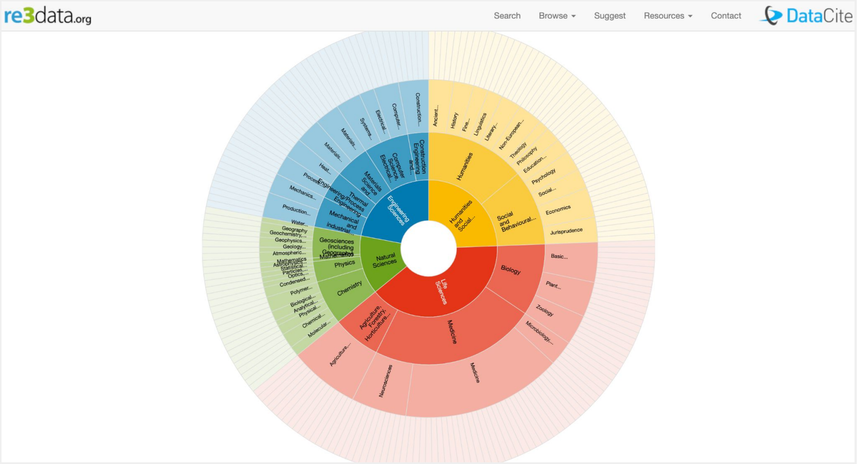
- Permanent publishing and archiving of data from the life sciences:
  - **Raw research data** = singular research data
  - **Enhanced publication** = research data linked to a full text
- **Requirements:**
  - Licensing of the data in the sense of **Open Data** to give the possibility of subsequent use
  - Providing a detailed **description** to ensure that the published research data can be clearly interpreted and reused in the future
  - Giving **essential information** (e.g. title, author(s), format)
- Information for authors and institutions

ZB MED service

# Examples of repository finders

DataCite’s **re**gistry of **re**search data **re**positories (**re3data**): global **registry** of research data **repositories**:

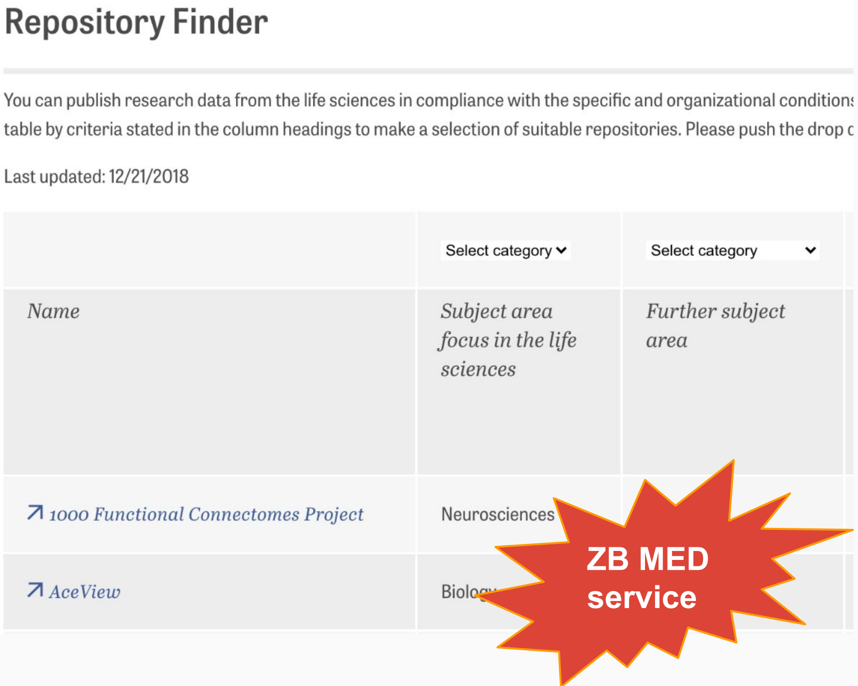
- from **different academic disciplines**
- that enable permanent **storage** of and **access** to data sets



# Examples of repository finders

**Repository Finder**: ZB MED's curated selection of repositories from re3data

- **Target audience**: researchers who would like to publish their research data
- **Criteria:**
  - **Subject:** Life Sciences
  - **Data access:** open
  - **Data upload:** open (registration at most)



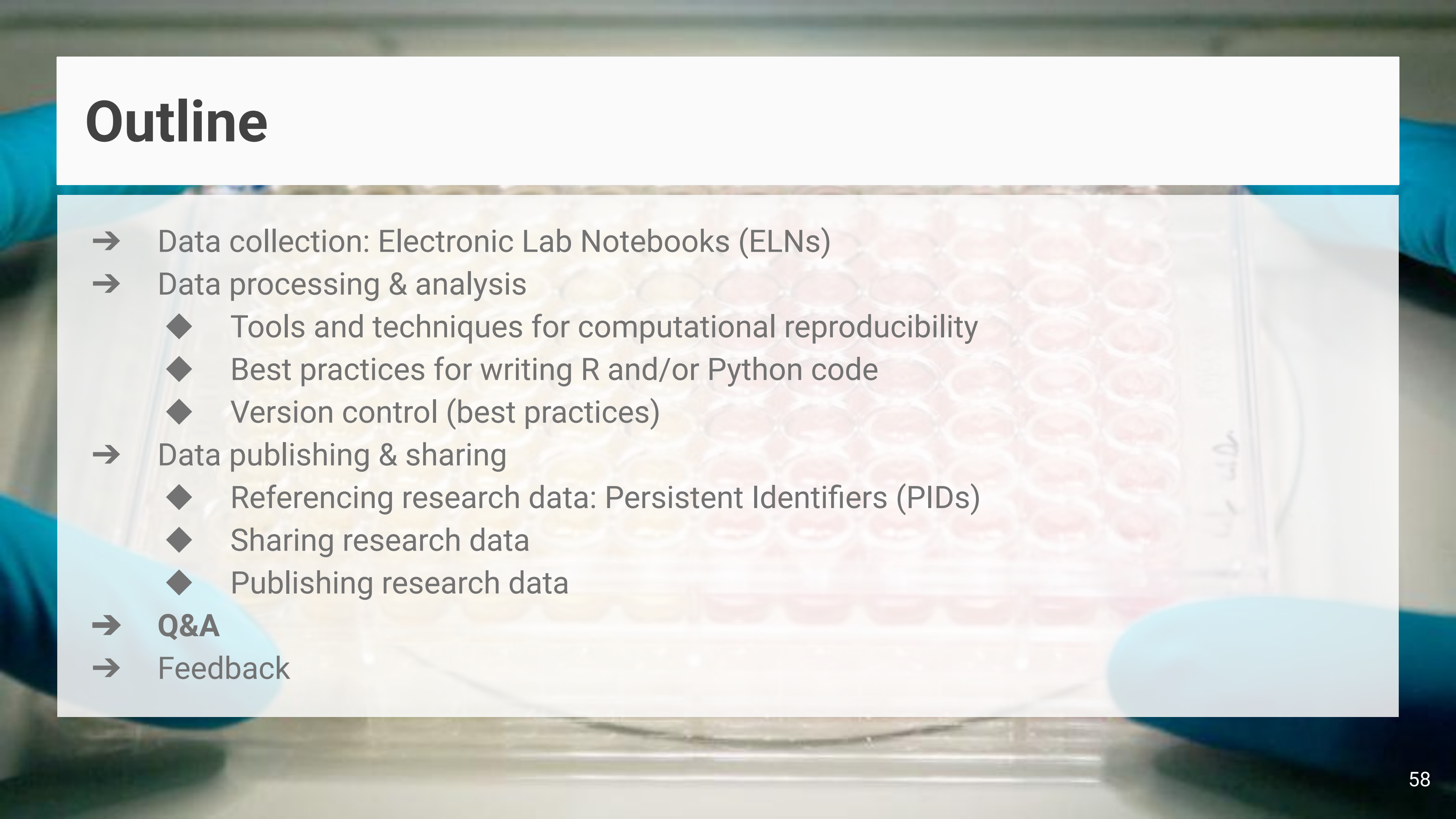
## Repository Finder

You can publish research data from the life sciences in compliance with the specific and organizational condition
table by criteria stated in the column headings to make a selection of suitable repositories. Please push the drop d

Last updated: 12/21/2018

| | Select category ⌄ | Select category ⌄ |
|---|---|---|
| *Name* | *Subject area focus in the life sciences* | *Further subject area* |
| ↗ *1000 Functional Connectomes Project* | Neurosciences | |
| ↗ *AceView* | Biolog | |

ZB MED service

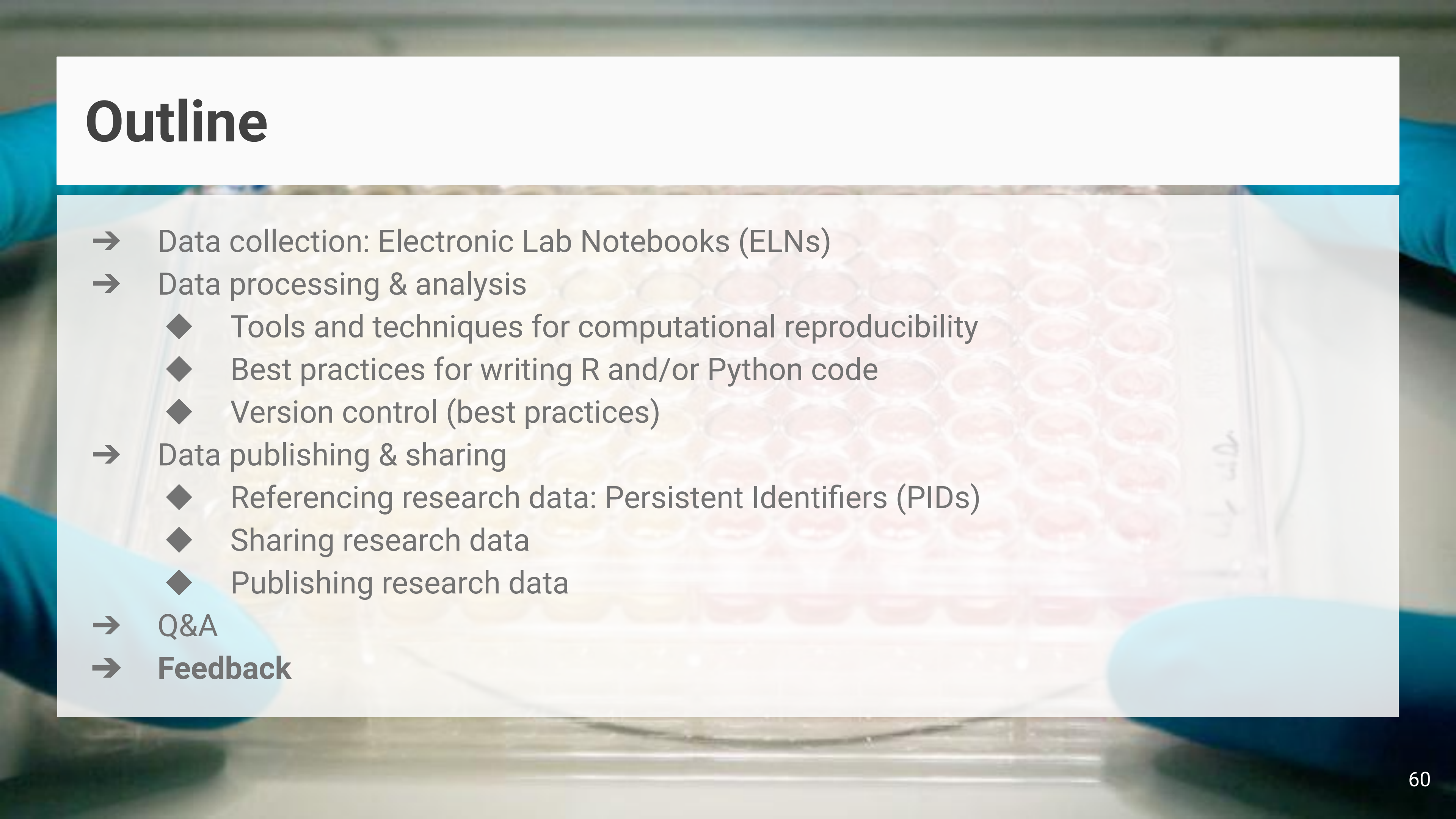


# Outline

- Data collection: Electronic Lab Notebooks (ELNs)
- Data processing & analysis
  - Tools and techniques for computational reproducibility
  - Best practices for writing R and/or Python code
  - Version control (best practices)
- Data publishing & sharing
  - Referencing research data: Persistent Identifiers (PIDs)
  - Sharing research data
  - Publishing research data
- **Q&A**
- Feedback

# Q&A

Photo by Jon Tyson on Unsplash



# Outline

- ➔ Data collection: Electronic Lab Notebooks (ELNs)
- ➔ Data processing & analysis
    - ◆ Tools and techniques for computational reproducibility
    - ◆ Best practices for writing R and/or Python code
    - ◆ Version control (best practices)
- ➔ Data publishing & sharing
    - ◆ Referencing research data: Persistent Identifiers (PIDs)
    - ◆ Sharing research data
    - ◆ Publishing research data
- ➔ Q&A
- ➔ **Feedback**

# Feedback

Thank you for attending our webinar. We now would like to ask you to fill in a 4-question survey to improve our webinars. In advance, thank you for your help.

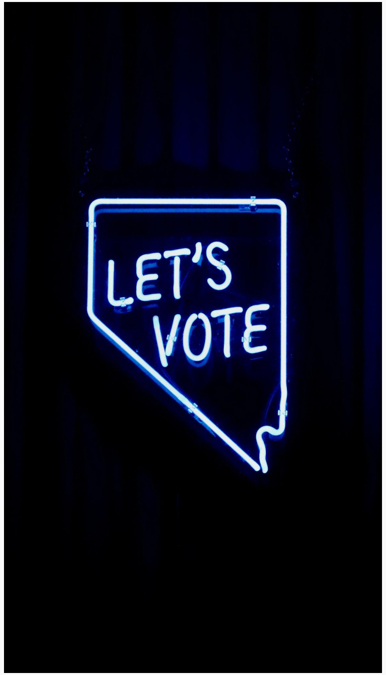


Photo by Manny Becerra
on Unsplash

# Thank you!

For further information we are at your disposal

**ZB MED –
Information Centre for
Life Sciences**
Gleueler Straße 60
50931 Köln

forschungsdaten@zbmed.de
www.zbmed.de

**HHU Düsseldorf**
Universitätsstraße 1
40225 Düsseldorf

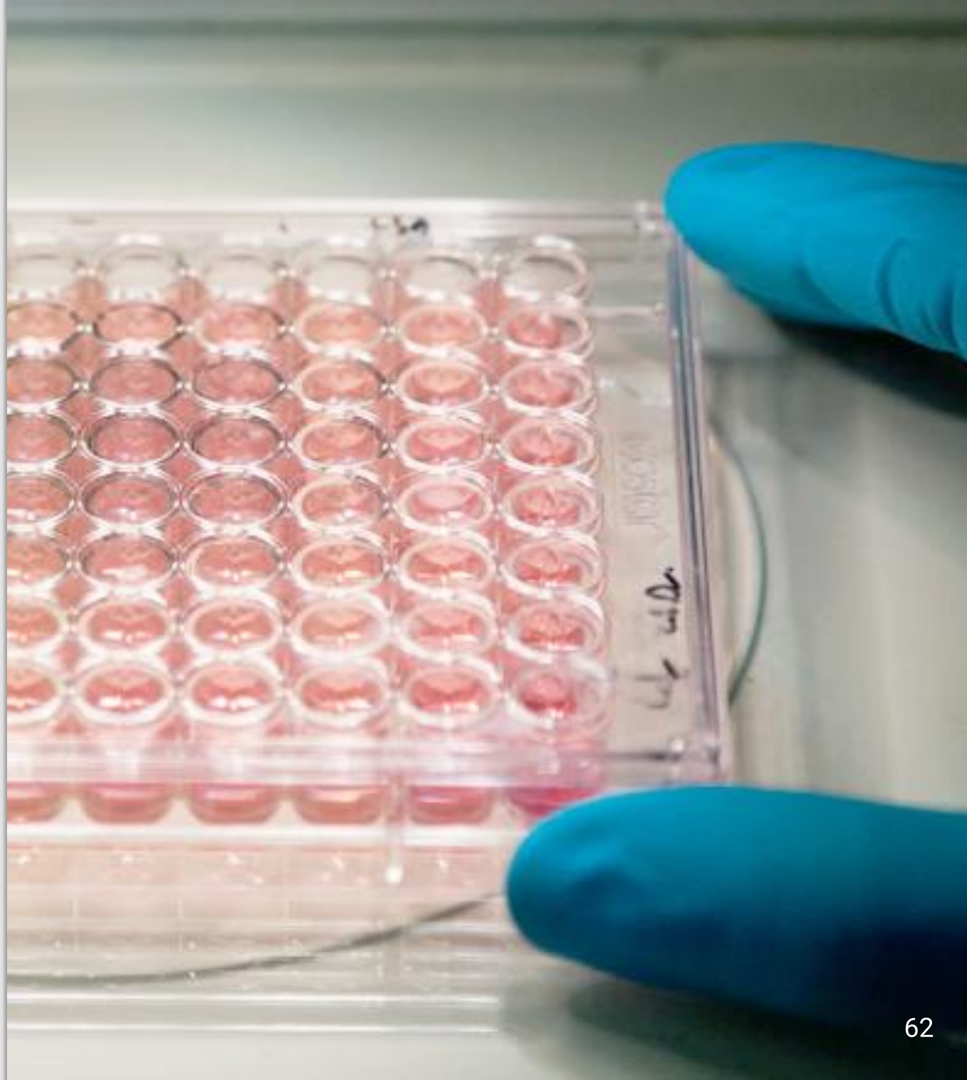
fdm@hhu.de
https://www.fdm.hhu.de

**Bergische Universität
Wuppertal –
Servicezentrum FDM**
Gaußstraße 20
42119 Wuppertal

fdm@uni-wuppertal.de
fdm.uni-wuppertal.de

**Universität Siegen –
e-Science-Service**

e-science-service@uni-siegen.de
https://e-science-service.uni-siegen.de/

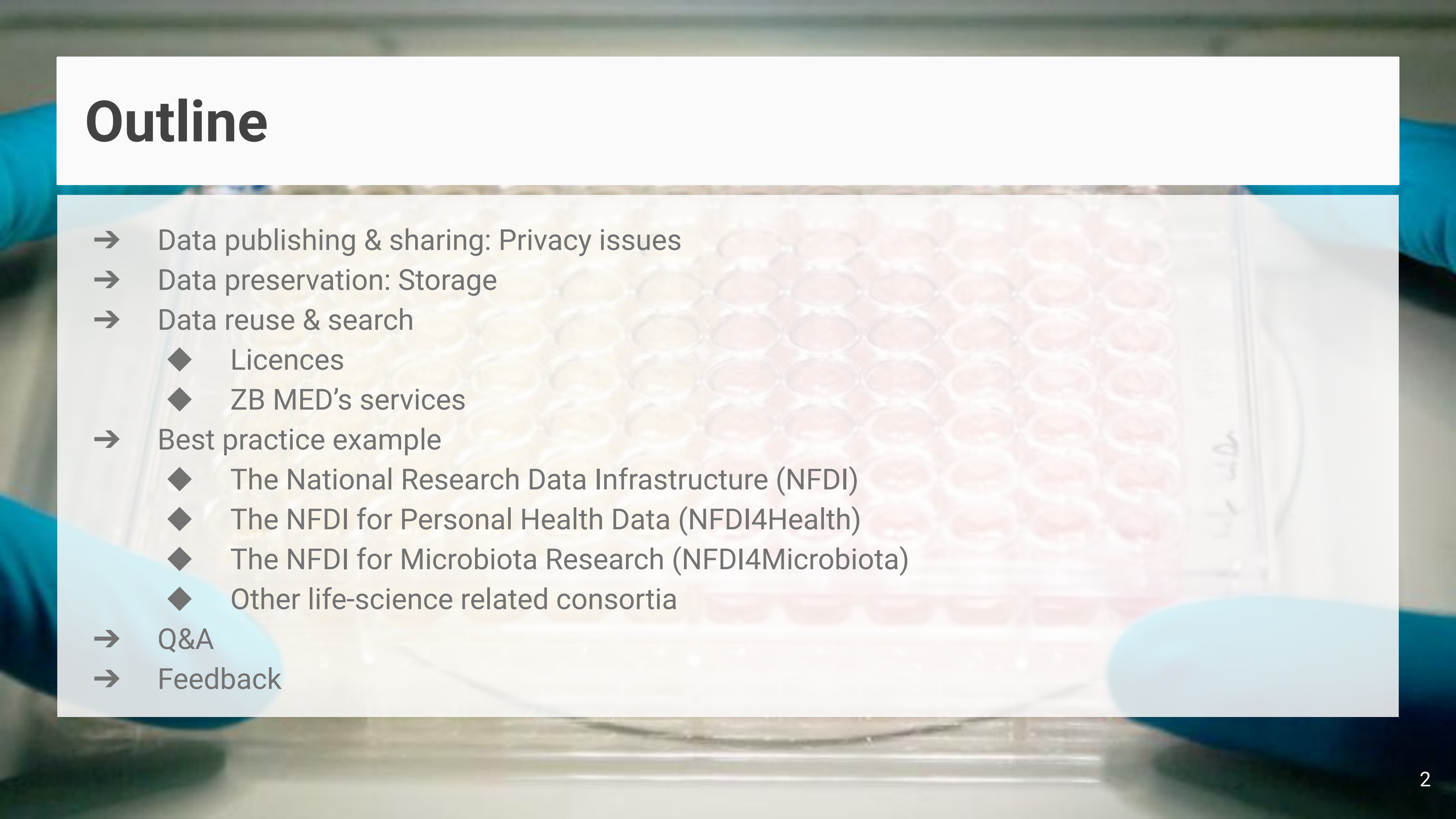


# WORKSHOP ON RESEARCH DATA MANAGEMENT
# Data publishing, preservation and reuse & NFDI

Torsten Rathmann, Servicezentrum Forschungsdatenmanagement Wuppertal
Daniela Kastrup, ULB Düsseldorf
Bastian Weiß, UB Siegen
B. Lindstädt, A. Shutsko & J. Vandendorpe, ZB MED - Information Centre for Life Sciences

ZB MED Information Centre for Life Sciences
BERGISCHE UNIVERSITÄT WUPPERTAL
UNIVERSITÄT SIEGEN
hhu Heinrich Heine Universität Düsseldorf

Photo by ZB MED



# Outline

- ➔ Data publishing & sharing: Privacy issues
- ➔ Data preservation: Storage
- ➔ Data reuse & search
  - ◆ Licences
  - ◆ ZB MED's services
- ➔ Best practice example
  - ◆ The National Research Data Infrastructure (NFDI)
  - ◆ The NFDI for Personal Health Data (NFDI4Health)
  - ◆ The NFDI for Microbiota Research (NFDI4Microbiota)
  - ◆ Other life-science related consortia
- ➔ Q&A
- ➔ Feedback

# Outline

- ➔ **Data publishing & sharing: Privacy issues**
- ➔ Data preservation: Storage
- ➔ Data reuse & search
  - ◆ Licences
  - ◆ ZB MED’s services
- ➔ Best practice example
  - ◆ The National Research Data Infrastructure (NFDI)
  - ◆ The NFDI for Personal Health Data (NFDI4Health)
  - ◆ The NFDI for Microbiota Research (NFDI4Microbiota)
  - ◆ Other life-science related consortia
- ➔ Q&A
- ➔ Feedback

# Privacy issues

**Informed consent** = ‘process for getting permission before conducting a healthcare intervention on a person, or for disclosing personal information’ [Wikipedia].

The German medical informatics initiative (MII) offers a template text for patient consent forms (only in German).

**Medizininformatik-Initiative**

Begleitstruktur – Koordinationsstelle des Nationalen Steuerungsgremiums

MEDIZIN
INFORMATIK
INITIATIVE

**Arbeitsgruppe Consent**
**Mustertext Patienteneinwilligung**

(Stand 26.04.2019)

Version 1.6a

**bestehend aus Patienteninformation und -einwilligung**

# Privacy issues: discussion

**Q1: why is health data privacy important (theoretically / in practice)?**

**Data privacy** = ‘right of a citizen to have control over how personal information is collected and used’ [EMOTIV].

# Privacy issues: discussion

**Q2: what are the risks of NOT sharing clinical trial data (including personal data)?**

**Personal data** = ‘all information associated with an identified or identifiable natural person’ [Federal Ministry of Health].

# Privacy issues: discussion

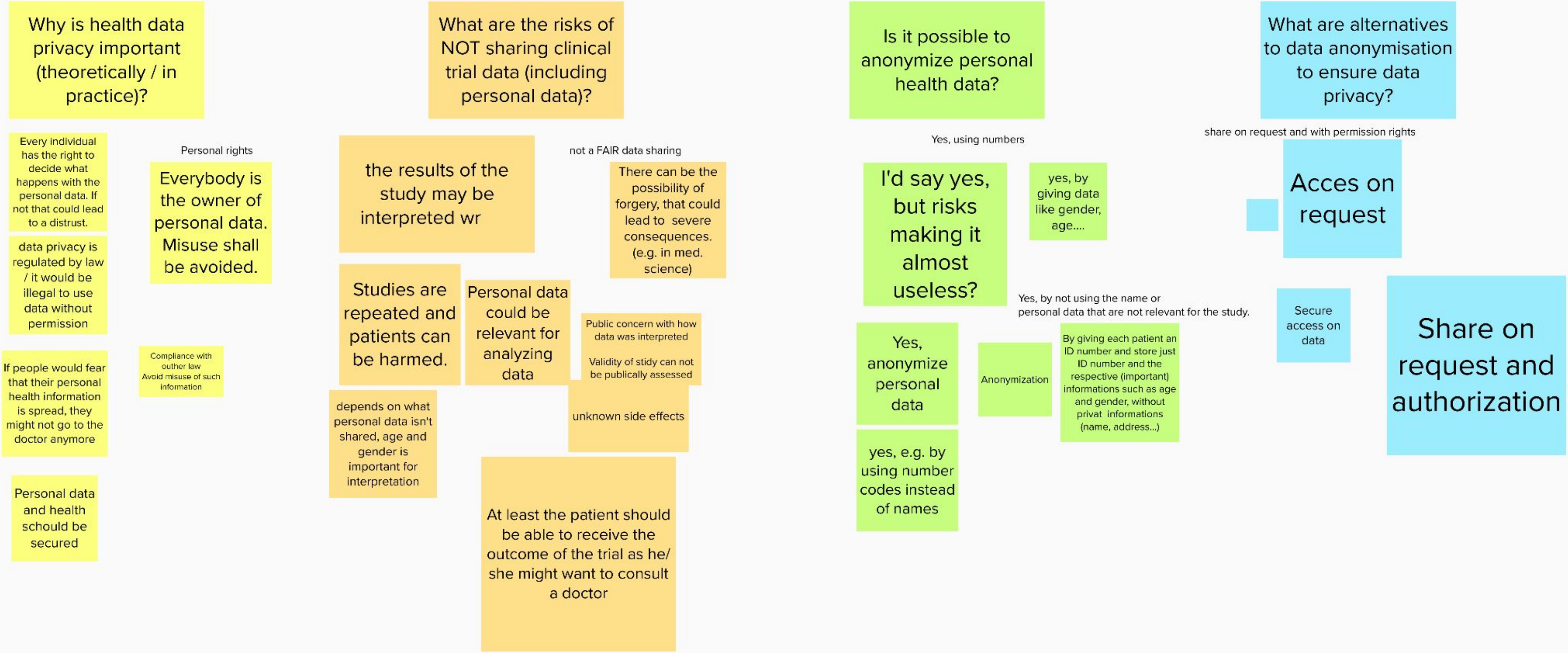
**Q3: is it possible to anonymize personal health data?**

- **Pseudonymisation:** ‘processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information’ [intersoft consulting].
- **Anonymisation:** ‘complete and irreversible removal of any information that could lead to an individual being identified, either from the removed information itself or this information combined with other data’ [University of Edinburgh].

# Privacy issues: discussion

**Q4: what are alternatives to data anonymisation to ensure data privacy?**

# Privacy issues: discussion



Why is health data privacy important (theoretically / in practice)?

Every individual has the right to decide what happens with the personal data. If not that could lead to a distrust.

Personal rights

Everybody is the owner of personal data. Misuse shall be avoided.

data privacy is regulated by law / it would be illegal to use data without permission

If people would fear that their personal health information is spread, they might not go to the doctor anymore

Compliance with outher law
Avoid misuse of such information

Personal data and health schould be secured

What are the risks of NOT sharing clinical trial data (including personal data)?

the results of the study may be interpreted wr

not a FAIR data sharing

There can be the possibility of forgery, that could lead to severe consequences. (e.g. in med. science)

Studies are repeated and patients can be harmed.

Personal data could be relevant for analyzing data

Public concern with how data was interpreted

Validity of stidy can not be publically assessed

depends on what personal data isn't shared, age and gender is important for interpretation

unknown side effects

At least the patient should be able to receive the outcome of the trial as he/she might want to consult a doctor

Is it possible to anonymize personal health data?

Yes, using numbers

I'd say yes, but risks making it almost useless?

yes, by giving data like gender, age....

Yes, by not using the name or personal data that are not relevant for the study.

Yes, anonymize personal data

Anonymization

By giving each patient an ID number and store just ID number and the respective (important) informations such as age and gender, without privat informations (name, address...)

yes, e.g. by using number codes instead of names

What are alternatives to data anonymisation to ensure data privacy?

share on request and with permission rights

Acces on request

Secure access on data

Share on request and authorization

# Privacy issues: discussion

**Q1: why is health data privacy important (theoretically / in practice)?**

**Suggested answer:**

- Privacy is a **basic human right** that promotes other **fundamental values** (e.g., individuality)
- Privacy furthers the existence of a **free society**
- Privacy is required for developing **interpersonal relationships** and promoting more **effective communication** between physician and patient
- Privacy can foster **socially beneficial activities** such as health research

[IOM (Institute of Medicine) 2009]

# Privacy issues: discussion

**Q2: what are the risks of NOT sharing clinical trial data (including personal data)?**

**Suggested answer:**

- Unnecessary **duplication** of trials and exposing additional participants to experimentation
- Increased **unwillingness** of individuals to participate in clinical trials if the data resulting from those trials are withheld
- **Bias** in the body of evidence
- Inability of investigators to build on previous work, thus **slowing scientific progress**

[IOM (Institute of Medicine) 2015]

# Privacy issues: discussion

**Q3: is it possible to anonymize personal health data?**

**Suggested answer:**

‘even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR* and seriously challenge the technical and legal adequacy of the de-identification release-and- forget model’ [Rocher *et al.* 2019].

*General Data Protection Regulation: ‘regulation in EU law on data protection and privacy in the European Union and the European Economic Area’ [Wikipedia].

# Privacy issues: discussion

**Q4: what are alternatives to data anonymisation to ensure data privacy?**

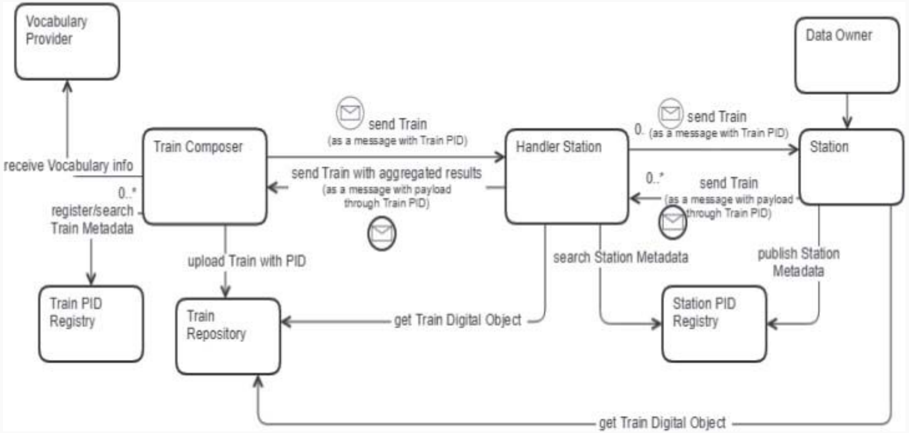
**Suggested answer:**

Personal data remain in their original location, and data owners enable analytical tasks to visit data sources and execute the task, leading to data being (re)used [Beyan *et al.* 2020].

# Privacy issues: discussion

**Q4: what are alternatives to data anonymisation to ensure data privacy?**

**Personal Health Train (PHT) Approach:** 'distributed infrastructure that enables the use and reuse of health data for the benefit of individuals and society' [GO FAIR].

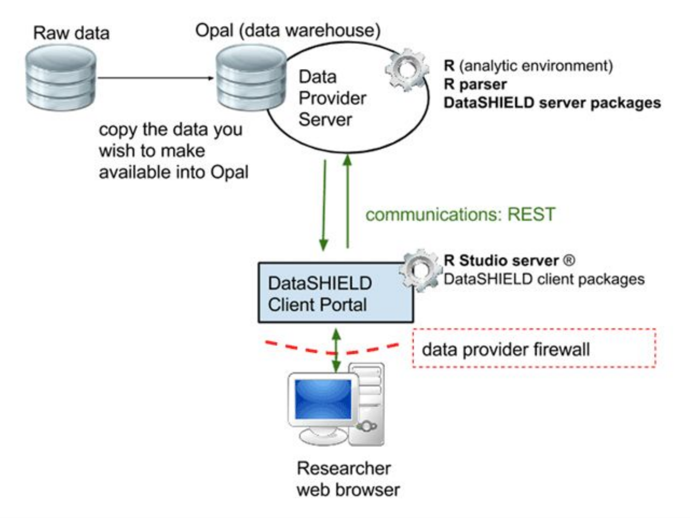


Main components of the PHT architecture [Bezan *et al.* 2020].

# Privacy issues: discussion

**Q4: what are alternatives to data anonymisation to ensure data privacy?**

**DataSHIELD:** ‘distributed approach that allows the analysis of sensitive individual- level data from one study, and the co- analysis of such data from several studies simultaneously without physically pooling them or disclosing any data’ [Wilson *et al.* 2017].



An example infrastructure for single site DataSHIELD.

# Outline

- ➔ Data publishing & sharing: Privacy issues
- ➔ **Data preservation: Storage**
- ➔ Data reuse & search
  - ◆ Licences
  - ◆ ZB MED’s services
- ➔ Best practice example
  - ◆ The National Research Data Infrastructure (NFDI)
  - ◆ The NFDI for Personal Health Data (NFDI4Health)
  - ◆ The NFDI for Microbiota Research (NFDI4Microbiota)
  - ◆ Other life-science related consortia
- ➔ Q&A
- ➔ Feedback

# Storage

- Why?
  - FAIR-Principles - Make Data Findable, Accessible, Reusable
  - Good Scientific Practice (ensure data-archival for at least 10 years)
  - Funder requirements

# Storage - Backup, Archival, Publication

- Backup
    - Keep data that is still being worked on safe
    - Maybe co-working solutions
- Archival
    - Long-term preservation
    - Not necessarily to allow access to others
- Publication
    - Allow ‘*others*’ to ‘*use*’ your data

# Storage - What to keep?

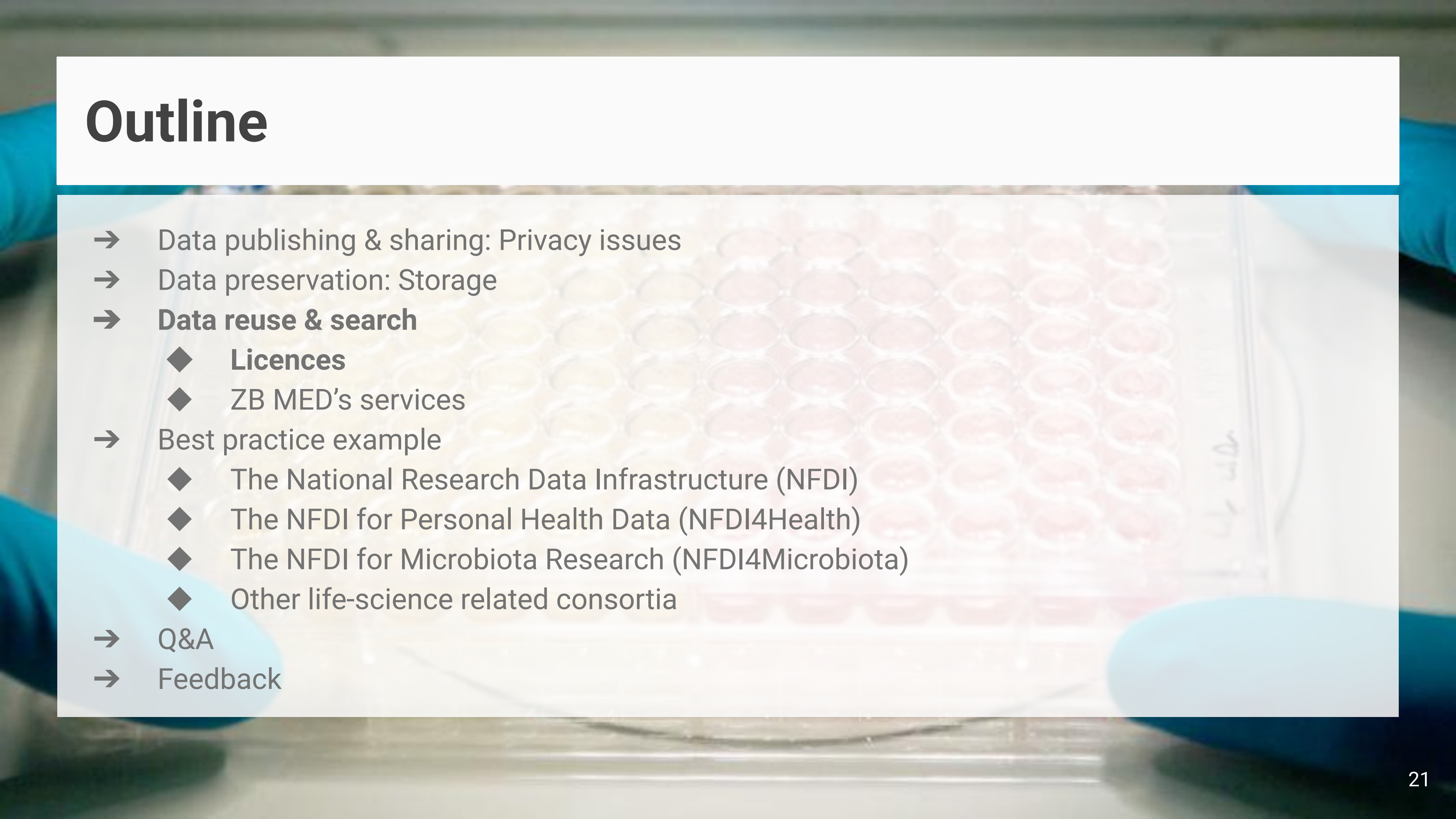
How to select data you want/need to keep? 5 steps:

1. Identify data that must be kept considering funders demands, legal or policy compliance risks
2. Identify reuse purposes that the data could fulfil
3. Identify data that should be kept as it may have long-term value
4. Weigh up the costs
5. Complete the data appraisal, including how to prepare the data for deposit or the justification for not keeping them

# Storage - Where to keep it?

- Select a repository!
  - How to find one? re3data; Help of experts in your discipline (NFDI, ZBMED)
  - Categories: **Discipline-specific**, Interdisciplinary, Institutional
  - Things to consider:
    - Costs?
    - Metadata
    - Visibility
    - Access(-restriction) options
- FoDaKo runs a joint infrastructure with repositories in Düsseldorf, Siegen, and Wuppertal (under construction)

**->** Demo of FoDaSi (Siegen)



# Outline

- ➔ Data publishing & sharing: Privacy issues
- ➔ Data preservation: Storage
- ➔ **Data reuse & search**
  - ◆ **Licences**
  - ◆ ZB MED’s services
- ➔ Best practice example
  - ◆ The National Research Data Infrastructure (NFDI)
  - ◆ The NFDI for Personal Health Data (NFDI4Health)
  - ◆ The NFDI for Microbiota Research (NFDI4Microbiota)
  - ◆ Other life-science related consortia
- ➔ Q&A
- ➔ Feedback

# Licences

- Why?
  - Provide legal certainty => Reusability
  - FAI**R** - Reusable
  - No Licence ≠ Free Licence/Free Use
- Which?
  - Different possibilities, we focus an **Creative Commons (CC)**
  - Others: Open Data Commons, Software: GNU GPL
- Example: https://zenodo.org/record/1469705#.YUGQUedCSUm

# Licences - Creative Commons

CC Licences are combinations of five components:

- BY: Author must always be named
- NC: Non-Commercial
- SA: Share Alike, same licence must be used
- ND: No Derivatives, no changes can be made
- Public Domain, none of the above limitations

# Licences - Creative Commons

- CC0 (Public Domain)
- CC BY (Name author)
- CC BY-SA (Name author - Share under same licence)
- CC BY-ND (Name author - No changes)
- CC BY-NC (Name author - No commercial use)
- CC BY-NC-SA (Name author - No commercial use - Share under same licence)
- CC BY-NC-ND (Name author - No commercial use - No changes)

# Licences - Creative Commons

- Licence Chooser: https://creativecommons.org/choose/#
- FAQs:
  - English: https://creativecommons.org/faq/
  - German: https://de.creativecommons.net/faqs/#

# Outline

- Data publishing & sharing: Privacy issues
- Data preservation: Storage
- **Data reuse & search**
  - Licences
  - **ZB MED’s services**
- Best practice example
  - The National Research Data Infrastructure (NFDI)
  - The NFDI for Personal Health Data (NFDI4Health)
  - The NFDI for Microbiota Research (NFDI4Microbiota)
  - Other life-science related consortia
- Q&A
- Feedback

# ZB MED's Search Portal for Life Sciences: LIVIVO

- Europe's largest **search engine** for literature and research data in the field of life sciences
- Access to about 50 **sources of data** (incl. MEDLINE)
- Automated linguistic enrichment, semantic linking of search terms, etc.

  → **Quick discovery** of information of interest
- Services:
  - **Link resolver** to allow users to check whether items in their hit list are available locally
  - LIVIVO **search field** on your website to allow users to perform searches directly in LIVIVO
  - LIVIVO **News page**
  - **Online tutorials** on LIVIVO (available soon)

**ZB MED service**

# Outline

- Data publishing & sharing: Privacy issues
- Data preservation: Storage
- Data reuse & search
  - Licences
  - ZB MED’s services
- **Best practice example**
  - **The National Research Data Infrastructure (NFDI)**
  - The NFDI for Personal Health Data (NFDI4Health)
  - The NFDI for Microbiota Research (NFDI4Microbiota)
  - Other life-science related consortia
- Q&A
- Feedback

# The National Research Data Infrastructure (NFDI)

- Central recommendation of the Council for Information Infrastructures (RfII) **2016** in the paper "Performance from Diversity": Establishment of the NFDI.
- **Service portfolios** to make research data accessible and usable: organized along **subject/thematic domains**, strong role of scientific data producers and users (consortia and subject communities)
- **Equally good** provision of research data infrastructures **nationwide** (across disciplines and institutions)

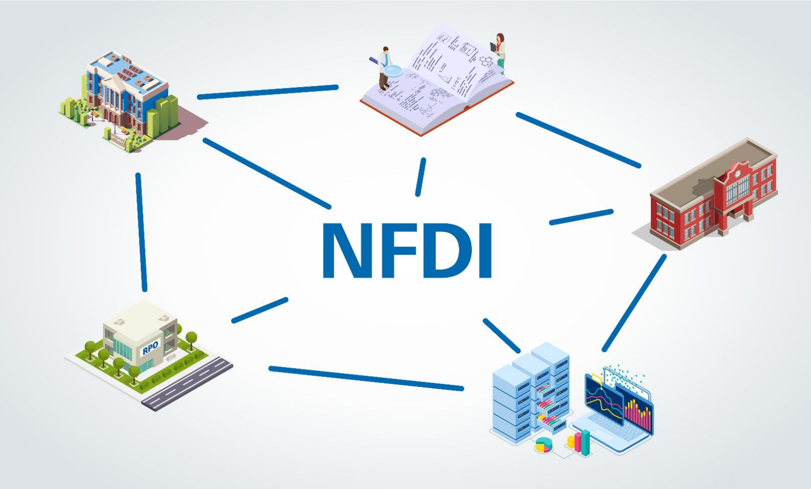


Photo by DFG

# NFDI: central characteristics

- ► **User involvement** - embedding in specialist communities
- ► **Collaboration** - using synergies via networking
- ► **Science-led** process (DFG)
- ► Create **sustainable**, **permanent** infrastructure
- ► **Networking** of NFDI consortia into ONE NFDI

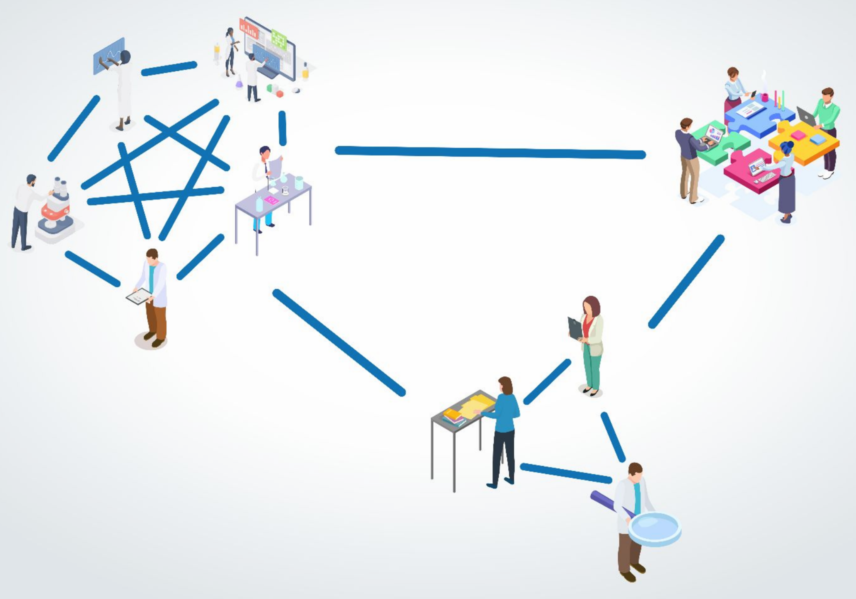


Photo by DFG

# NFDI: consortia formation

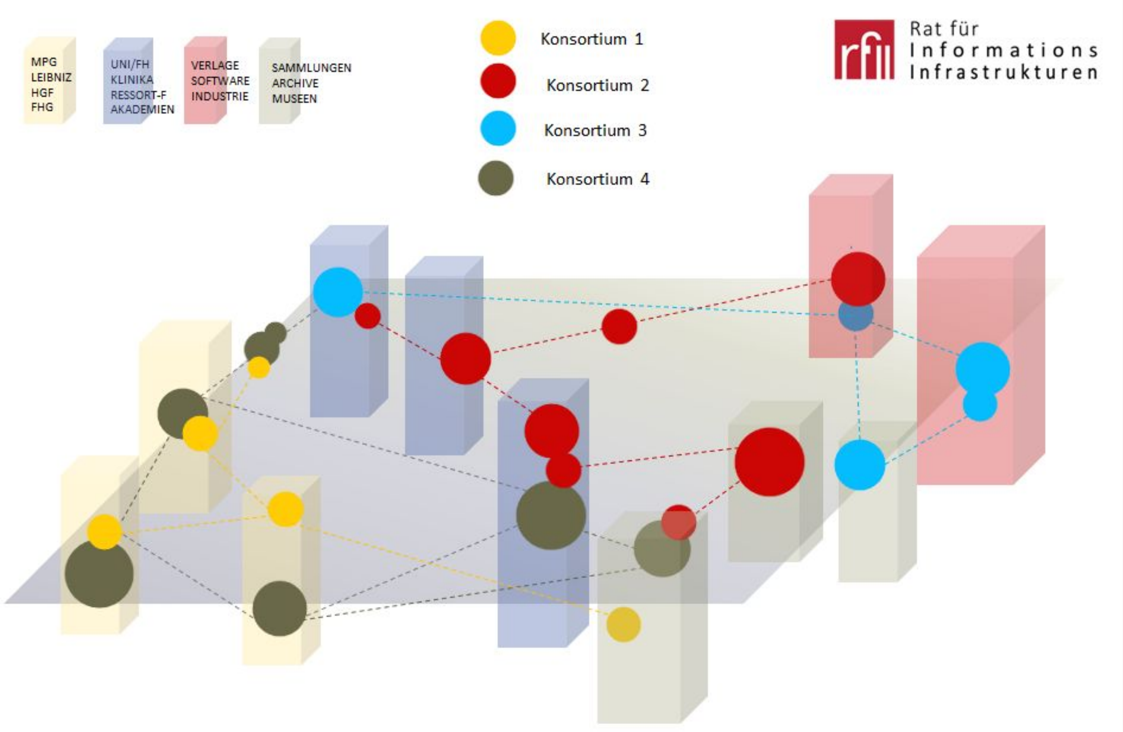


Photo by RfII

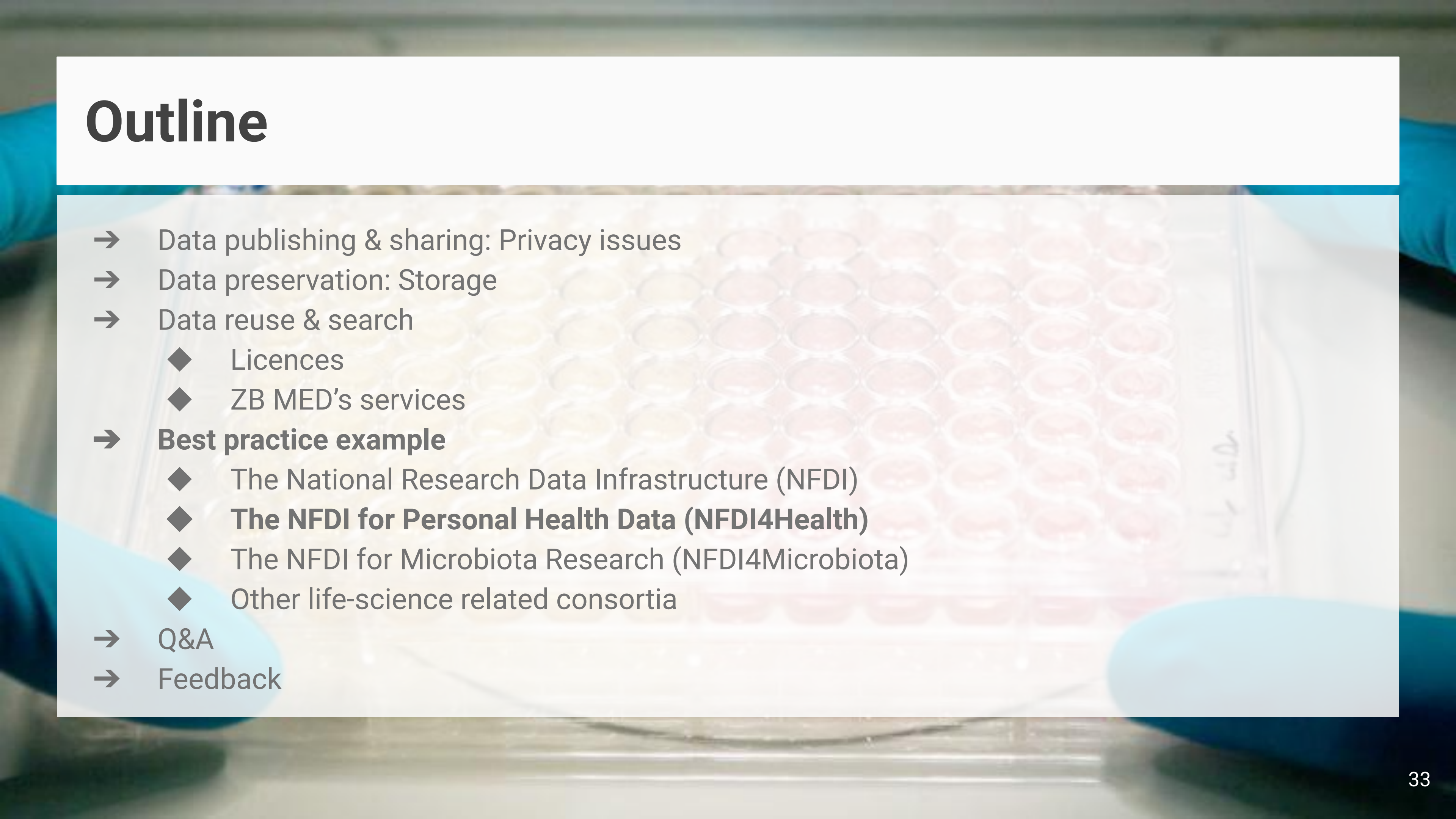
**Networking:**
Consortia formation horizontal to existing pillars in the science system.

# NFDI: overview

Deutsche Forschungsgemeinschaft

The National Research Data Infrastructure in Germany (NFDI)

https://www.dfg.de/en/research_funding/programmes/nfdi/information_material/index.html



# Outline

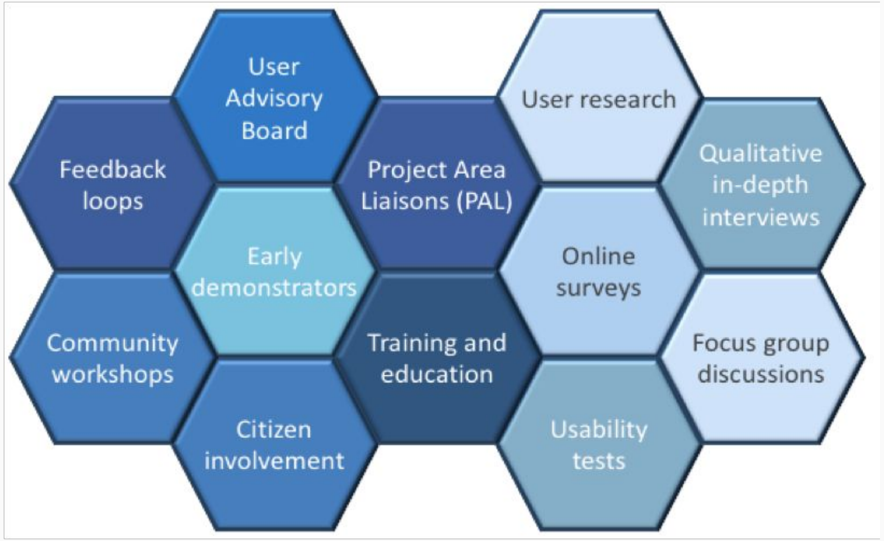
- ➔ Data publishing & sharing: Privacy issues
- ➔ Data preservation: Storage
- ➔ Data reuse & search
  - ◆ Licences
  - ◆ ZB MED’s services
- ➔ **Best practice example**
  - ◆ The National Research Data Infrastructure (NFDI)
  - ◆ **The NFDI for Personal Health Data (NFDI4Health)**
  - ◆ The NFDI for Microbiota Research (NFDI4Microbiota)
  - ◆ Other life-science related consortia
- ➔ Q&A
- ➔ Feedback

# NFDI4Health: user community

## Community

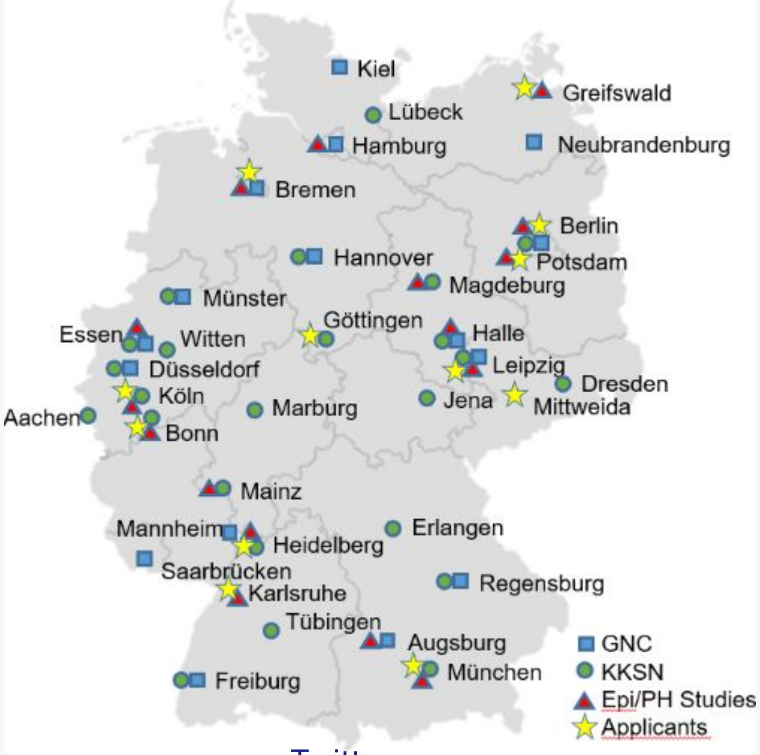
- Epidemiologists, public health and clinical researchers
- Other NFDI consortia
- Citizens and patients

## Community involvement



# NFDI4Health: research data

- Epidemiological and public health studies
  - 26 local studies with > 400,000 participants
  - German National Cohort Study (GNC/NAKO)
- Clinical studies
  - 24 university study centers
- Registries
- Health surveillance systems
- Administrative health data banks

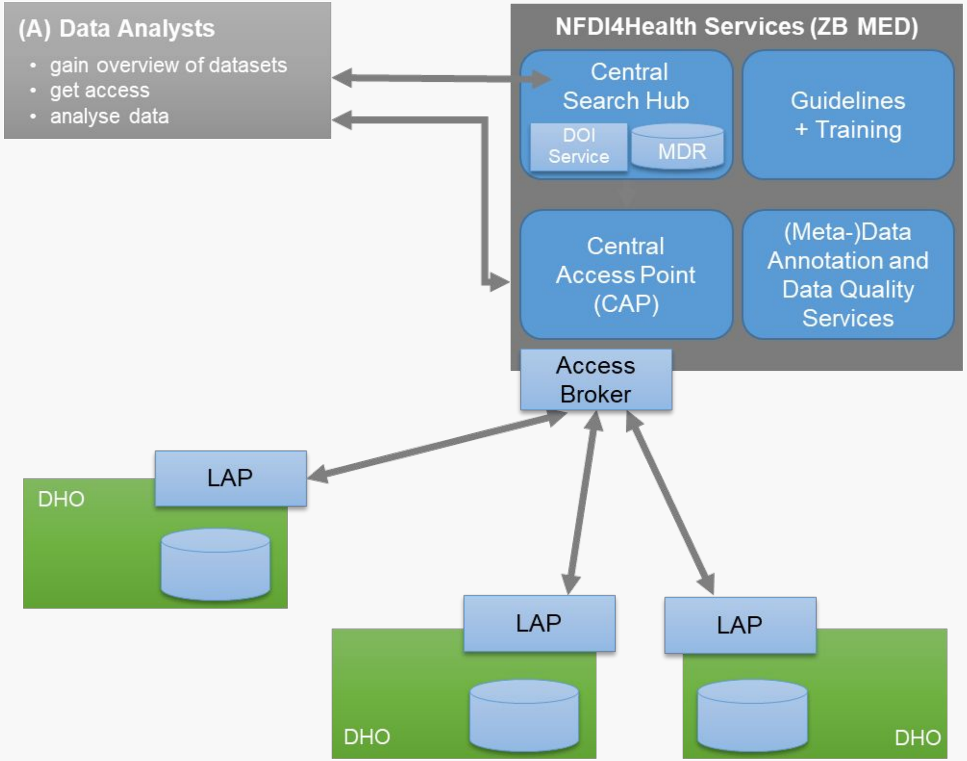


Website Email Twitter

# NFDI4Health: objectives

- Improve discoverability of health data through data publications: **Central Search Hub**
- **Standardization** of **metadata**, improvement of interoperability
- Implementation of a higher-level data access and data use process: **Central Access Point**
- Ensure use only in accordance with **consent and privacy policies**
- **Data analysis**: Further development of services for **controlled access to distributed data** using analysis tools.
- Close **involvement of the community** to achieve sustainability

# NFDI4Health: services



(A) Data Analysts
• gain overview of datasets
• get access
• analyse data
NFDI4Health Services (ZB MED)
Central Search Hub
DOI Service
MDR
Guidelines + Training
Central Access Point (CAP)
(Meta-)Data Annotation and Data Quality Services
Access Broker
LAP
DHO
LAP
DHO
LAP
DHO

**NFDI4Health** is building an **access system**:

- Locating data in the "**Central Search Hub**".
- Request from researchers at the "**Central Access Point**".
- Data is provided by the "**Local Access Points**". They are not held centrally but decentrally by the "**Data Holding Organizations**".

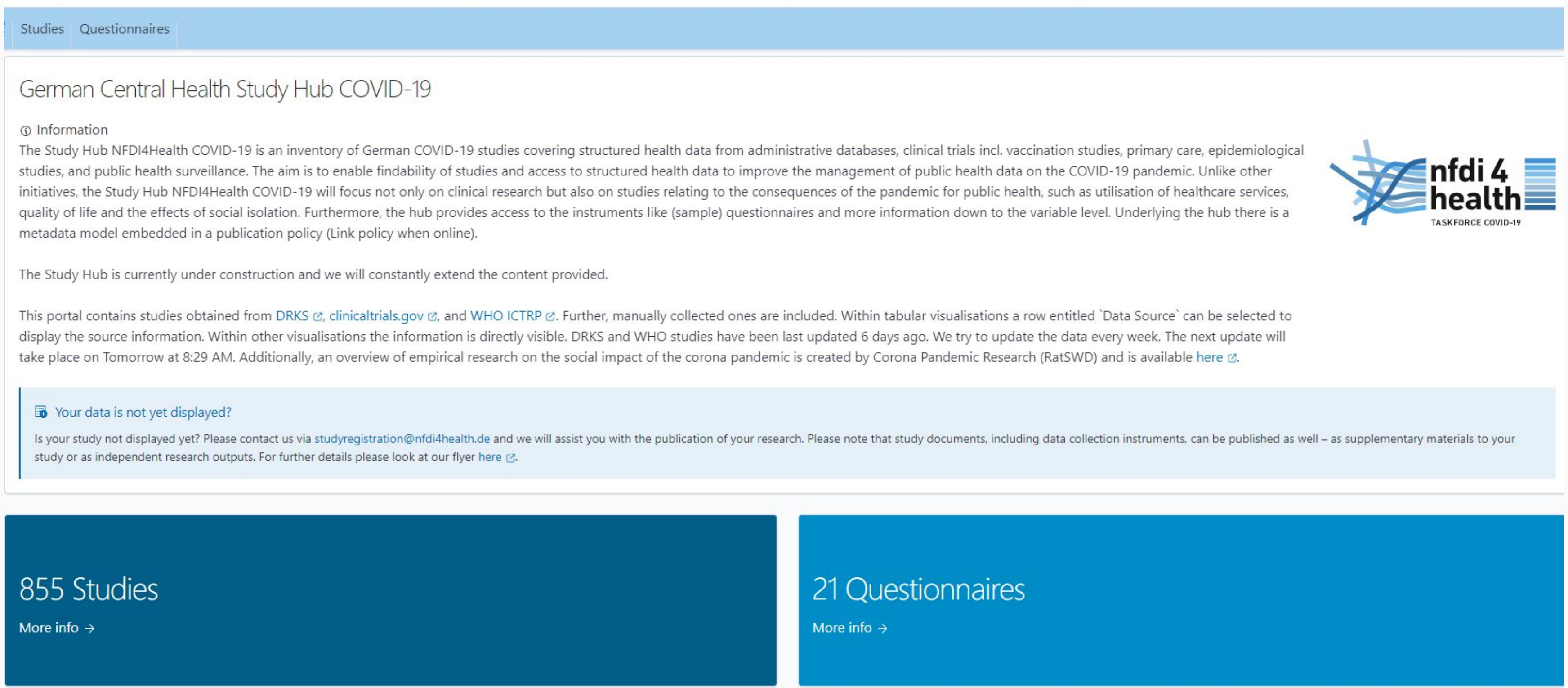
| | | |
|---|---|---|
| DHO | = | Data Holding Organisation |
| LAP | = | Local Access Point |
| MDR | = | Metadata Repository |
| ▬ | = | Researcher Queries |
| ▬ | = | Data Holder Support |
| ▬ | = | Distributed Data Analyses |

# NFDI4Health: first results

- **German Central Health Study Hub COVID-19**
  - https://covid19.studyhub.nfdi4health.de/

- **Publication guidelines** for the Study Hub:
  - **Types of resources** for publication: study metadata, documents/instruments
  - **Metadata schema**: MDS developed on standards (DataCite, HL7 FHIR)
  - **Licensing**: CC-Licences (?)
  - **Assignment of DOIs**: for documents/instruments
  - **File formats** of the study documents: machine-readable, but also human-readable
  - **Language**: English prefered, but German also accepted

# NFDI4Health: German Central Health Study Hub COVID-19



## German Central Health Study Hub COVID-19

ⓘ Information

The Study Hub NFDI4Health COVID-19 is an inventory of German COVID-19 studies covering structured health data from administrative databases, clinical trials incl. vaccination studies, primary care, epidemiological studies, and public health surveillance. The aim is to enable findability of studies and access to structured health data to improve the management of public health data on the COVID-19 pandemic. Unlike other initiatives, the Study Hub NFDI4Health COVID-19 will focus not only on clinical research but also on studies relating to the consequences of the pandemic for public health, such as utilisation of healthcare services, quality of life and the effects of social isolation. Furthermore, the hub provides access to the instruments like (sample) questionnaires and more information down to the variable level. Underlying the hub there is a metadata model embedded in a publication policy (Link policy when online).

The Study Hub is currently under construction and we will constantly extend the content provided.

This portal contains studies obtained from DRKS, clinicaltrials.gov, and WHO ICTRP. Further, manually collected ones are included. Within tabular visualisations a row entitled `Data Source` can be selected to display the source information. Within other visualisations the information is directly visible. DRKS and WHO studies have been last updated 6 days ago. We try to update the data every week. The next update will take place on Tomorrow at 8:29 AM. Additionally, an overview of empirical research on the social impact of the corona pandemic is created by Corona Pandemic Research (RatSWD) and is available here.

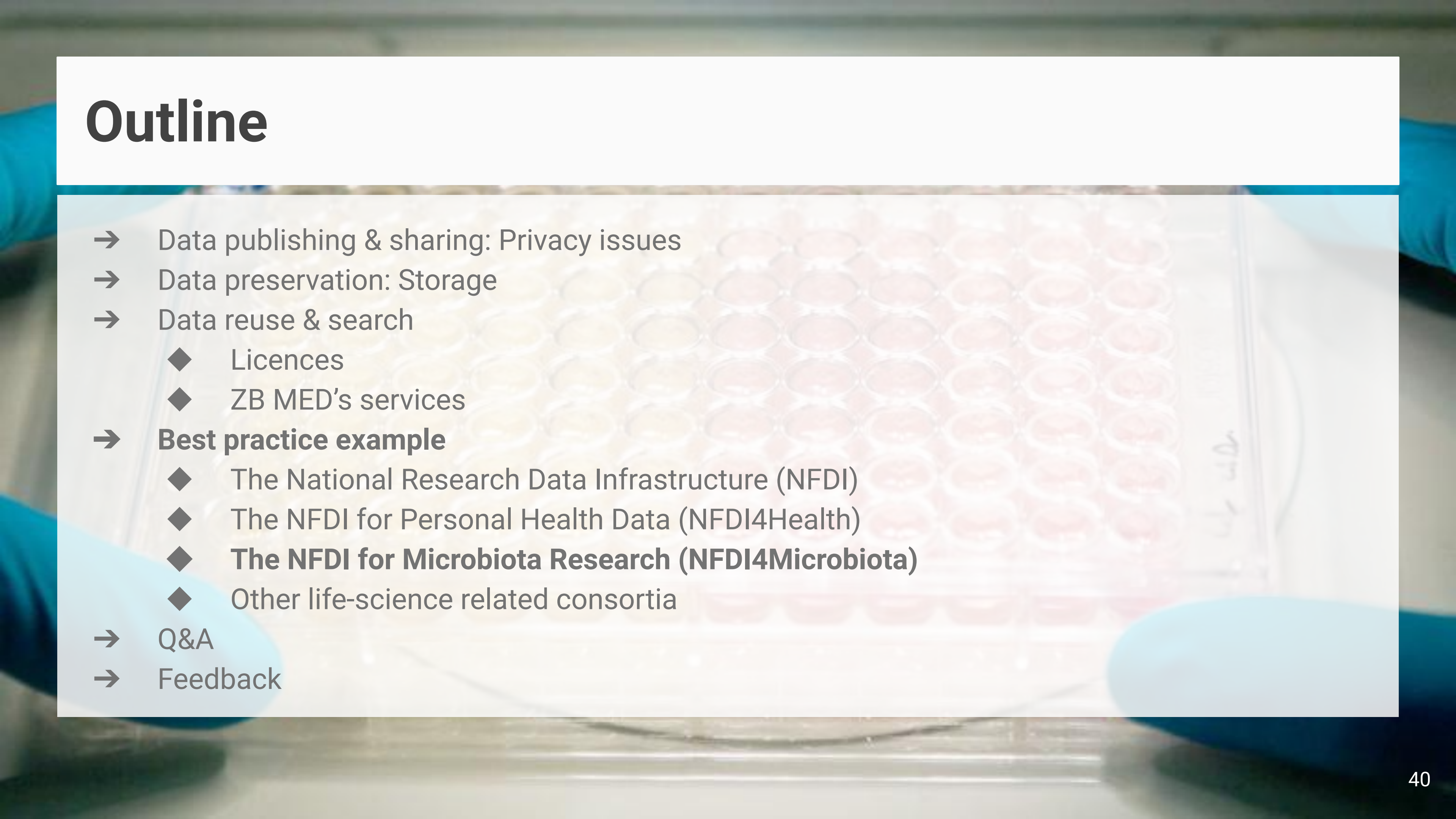
Your data is not yet displayed?

Is your study not displayed yet? Please contact us via studyregistration@nfdi4health.de and we will assist you with the publication of your research. Please note that study documents, including data collection instruments, can be published as well – as supplementary materials to your study or as independent research outputs. For further details please look at our flyer here.

855 Studies

21 Questionnaires

# Outline

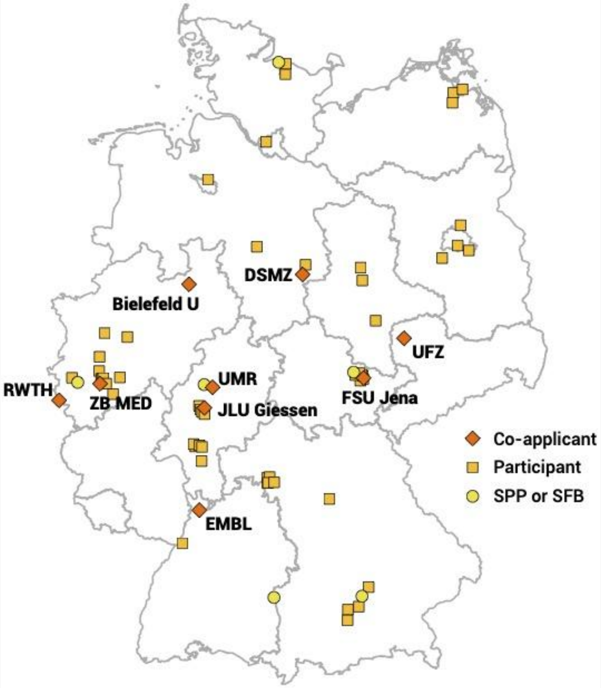


- ➔ Data publishing & sharing: Privacy issues
- ➔ Data preservation: Storage
- ➔ Data reuse & search
  - ◆ Licences
  - ◆ ZB MED’s services
- ➔ **Best practice example**
  - ◆ The National Research Data Infrastructure (NFDI)
  - ◆ The NFDI for Personal Health Data (NFDI4Health)
  - ◆ **The NFDI for Microbiota Research (NFDI4Microbiota)**
  - ◆ Other life-science related consortia
- ➔ Q&A
- ➔ Feedback

# NFDI4Microbiota: user community

**Community:** researchers from the German microbiology research community

- Bacteriologists
- Virologists
- Protistologists
- Mycologists
- Parasitologists



Website • Email • Twitter

# NFDI4Microbiota: research data

Data related to microbial species and diverse microbiomes:

- (Meta-)genomics
- (Meta-)transcriptomics
- (Meta-)proteomics
- Metabolomics
- Image analysis
- Linked open data / semantics

Photo by CDC on Unsplash

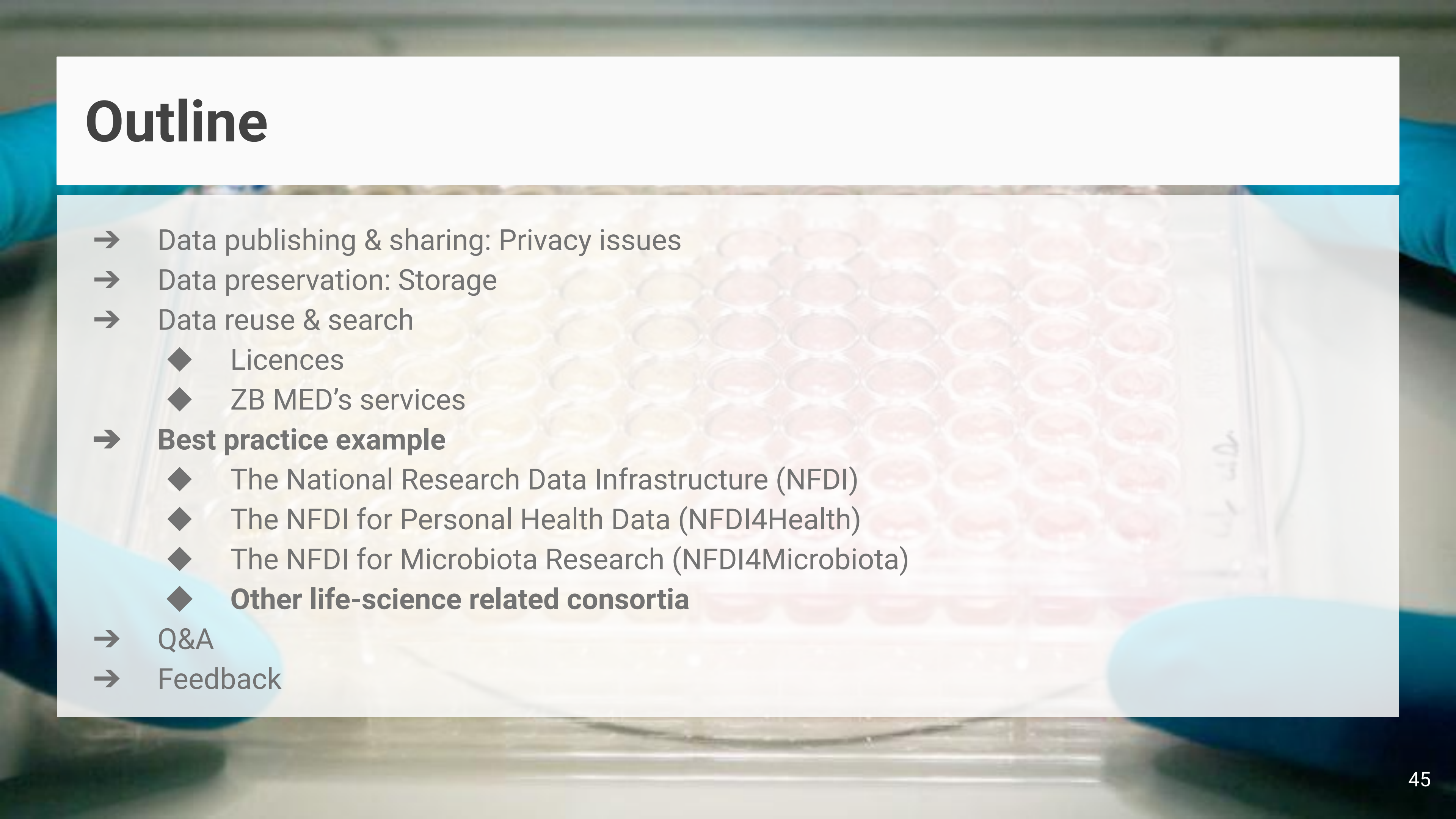
# NFDI4Microbiota: objectives

- Making **data** related to microbial species and diverse microbiomes **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (**FAIR**)
- Making the **analysis** of such data **accessible**, **consistent** and **reproducible**
- Supporting the deep understanding of **microbial species** and their **interactions** on a molecular level
- Offering **automated data processing systems**
- **Use cases** oriented towards researchers‘ needs
- Promoting **data sharing** and **reuse**
- Improving the **quality of research**
- Promoting new **collaborations**

# NFDI4Microbiota: services

NFDI4Microbiota will be the **central hub** in Germany for supporting the microbiology community with:

- **Bioinformatics tools and databases** (e.g., multi-omics analyses, interaction modeling, semantic enrichment, BacDive)
- **Computational infrastructure** (e.g., workflow engine, cloud infrastructure)
- Development of **standards** regarding sampling, processing and metadata
- **Trainings** about Research Data Management, Infrastructure & Software (e.g. bioinformatics, cloud, workflow engines) and omics.



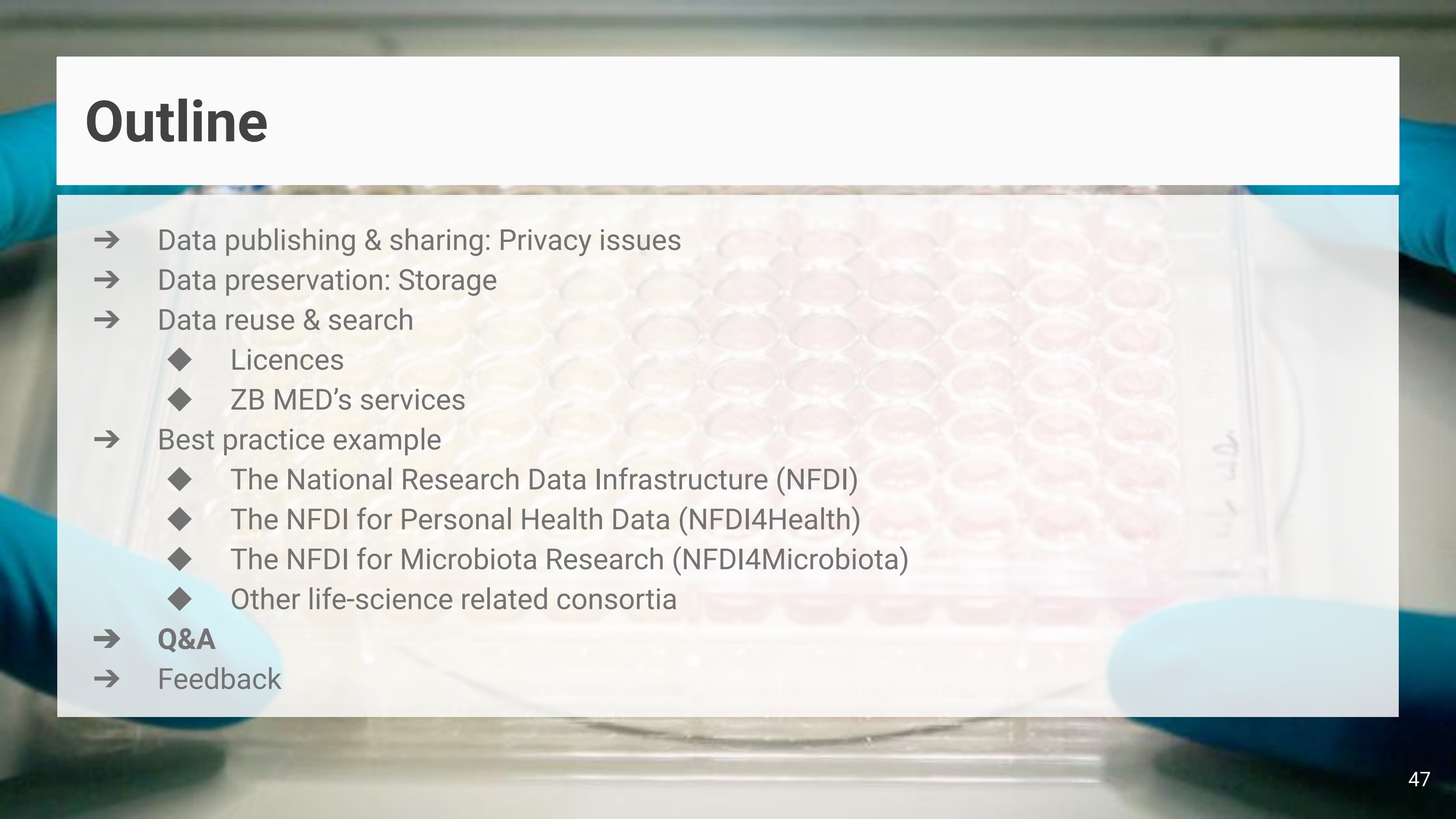
# Outline

- ➔ Data publishing & sharing: Privacy issues
- ➔ Data preservation: Storage
- ➔ Data reuse & search
  - ◆ Licences
  - ◆ ZB MED’s services
- ➔ **Best practice example**
  - ◆ The National Research Data Infrastructure (NFDI)
  - ◆ The NFDI for Personal Health Data (NFDI4Health)
  - ◆ The NFDI for Microbiota Research (NFDI4Microbiota)
  - ◆ **Other life-science related consortia**
- ➔ Q&A
- ➔ Feedback

# Other life-science related consortia

| Round funding | Name | Acronym | Email |
|---|---|---|---|
| **1st** | German Human Genome-Phenome Archive | GHGA | contact@ghga.de |
| **3rd** | National Research Data Infrastructure for Immunology | NFDI4Immuno | christian.busse@dkfz-heidelberg.de |
| | NFDI Neuroscience | NFDI-Neuro | nfo@nfdi-neuro.de |
| | NFDI for Pre-clinical Drug Discovery and Chemical Biology | DeBioData | philip.gribbon@ime.fraunhofer.de |
| | NFDI4 Biological Imaging and Medical Photonics | NFDI4BIOIMAGE | elisa.may@uni-konstanz.de |
| | National Research Data Infrastructure for Digital Pathology | NFDI4Patho | pboor@ukaachen.de |

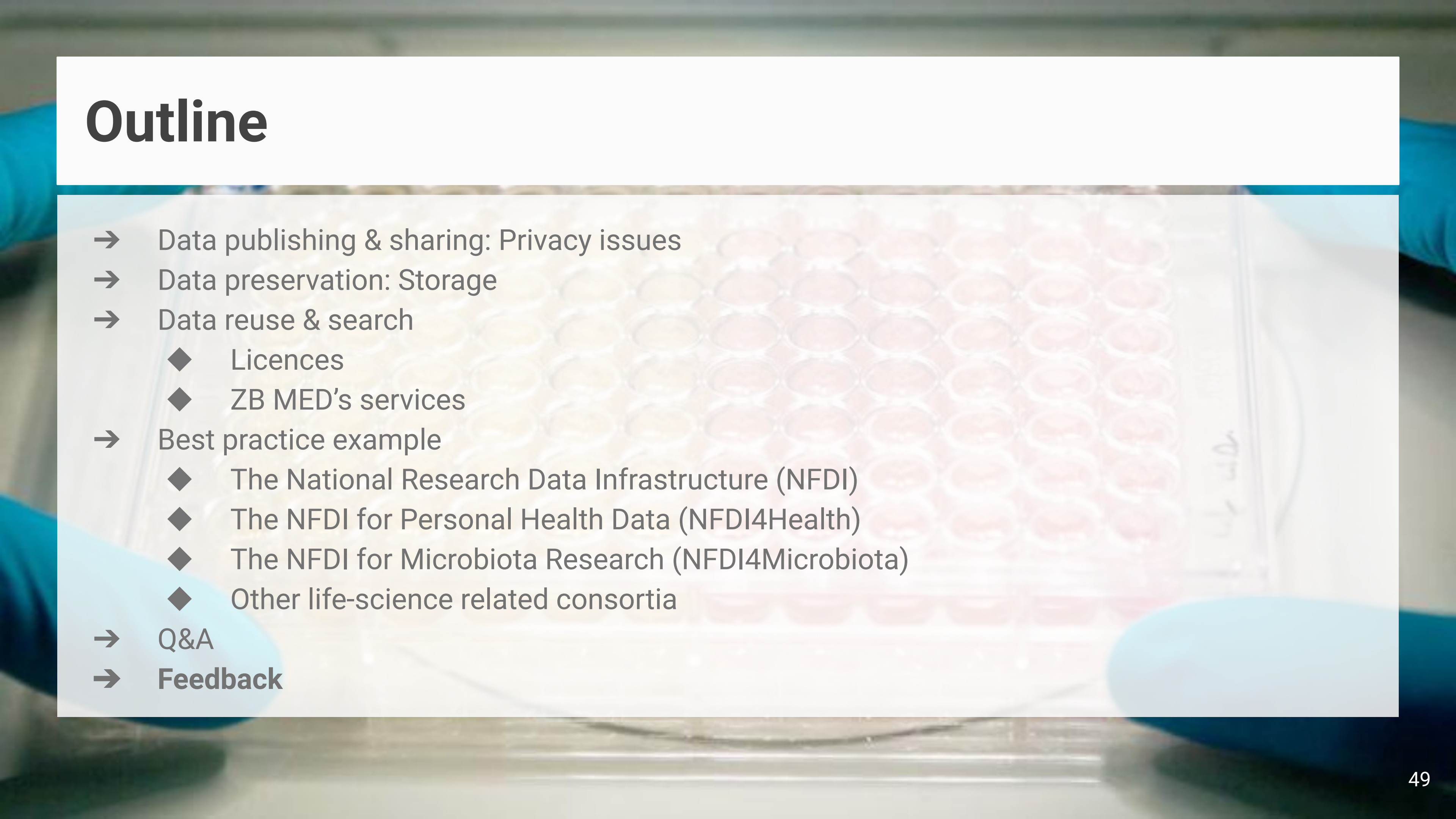
# Outline



- ➔ Data publishing & sharing: Privacy issues
- ➔ Data preservation: Storage
- ➔ Data reuse & search
  - ◆ Licences
  - ◆ ZB MED’s services
- ➔ Best practice example
  - ◆ The National Research Data Infrastructure (NFDI)
  - ◆ The NFDI for Personal Health Data (NFDI4Health)
  - ◆ The NFDI for Microbiota Research (NFDI4Microbiota)
  - ◆ Other life-science related consortia
- ➔ **Q&A**
- ➔ Feedback

# Q&A

Photo by Jon Tyson on Unsplash



# Outline

- ➔ Data publishing & sharing: Privacy issues
- ➔ Data preservation: Storage
- ➔ Data reuse & search
  - ◆ Licences
  - ◆ ZB MED’s services
- ➔ Best practice example
  - ◆ The National Research Data Infrastructure (NFDI)
  - ◆ The NFDI for Personal Health Data (NFDI4Health)
  - ◆ The NFDI for Microbiota Research (NFDI4Microbiota)
  - ◆ Other life-science related consortia
- ➔ Q&A
- ➔ **Feedback**

# Feedback

Thank you for attending our webinar. We now would like to ask you to fill in a 4-question survey to improve our webinars. In advance, thank you for your help.

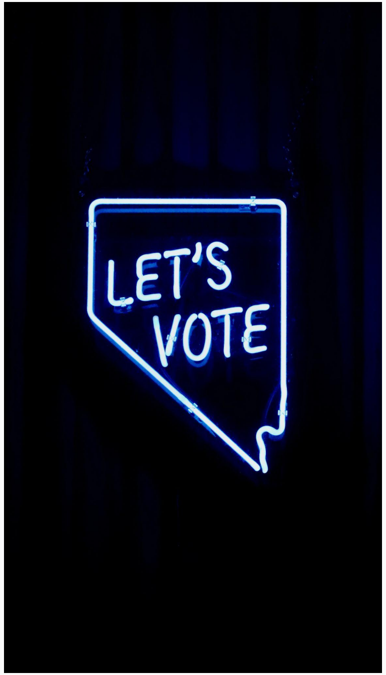


Photo by Manny Becerra on Unsplash

# Thank you!

For further information we are at
your disposal

**ZB MED –**
**Information Centre for**
**Life Sciences**
Gleueler Straße 60
50931 Köln

forschungsdaten@zbmed.de
www.zbmed.de

**HHU Düsseldorf**
Universitätsstraße 1
40225 Düsseldorf

fdm@hhu.de
https://www.fdm.hhu.de

**Bergische Universität**
**Wuppertal –**
**Servicezentrum FDM**
Gaußstraße 20
42119 Wuppertal

fdm@uni-wuppertal.de
fdm.uni-wuppertal.de

**Universität Siegen –**
**e-Science-Service**

e-science-service@uni-siegen.de
https://e-science-service.uni-siegen.de/