

# NFDI4Health

## ONLINE TRAINING WORKSHOP ON RESEARCH DATA MANAGEMENT IN (BIO-)MEDICINE

J. Fürst<sup>1</sup>, B. Lindstädt<sup>1</sup>, I. Perrar<sup>2</sup>, A. Shutsko<sup>1</sup> & J. Vandendorpe<sup>1</sup>

<sup>1</sup> ZB MED - Information Centre for Life Sciences

<sup>2</sup> Institut für Ernährungs- und Lebensmittelwissenschaften, University Bonn



Unless stated otherwise within the sources cited in this work, this work is licensed under the Creative Commons Attribution 4.0 International Public License (CC BY 4.0). To view a copy of the license, visit <https://creativecommons.org/licenses/by/4.0/legalcode>



# What you will find in this slide deck

## Slides containing content provided by ZB MED / NFDI4Health

In your adaptation of the training workshop, these are the slides you as the library or other institution would present

## Slides containing content provided by the IEL Bonn

In your adaptation of the training workshop, these are the slides provided by the discipline-specific research institute you host the training workshop with/at

## Slides containing activities such as exercises and polls

You will find instructions on how to use the activities on the respective slides

# Schedule

## Day 1- Introduction to Research data management (RDM) - Fundamental concepts - Duration: 90mins

- 00:00 - 00:25 Welcome + Introduction of affiliation
- 00:25 - 00:30 Research data
- 00:30 - 00:45 Research data life cycle
- 00:45 - 00:55 Metadata
- 00:55 - 01:05 FAIR data principles
- 01:05 - 01:10 Guidelines on RDM
- 01:10 - 01:25 Q&A
- 01:25 - 01:30 Feedback

## Day 2 - A step forward to making data FAIR with NFDI4Health – Duration: 90mins

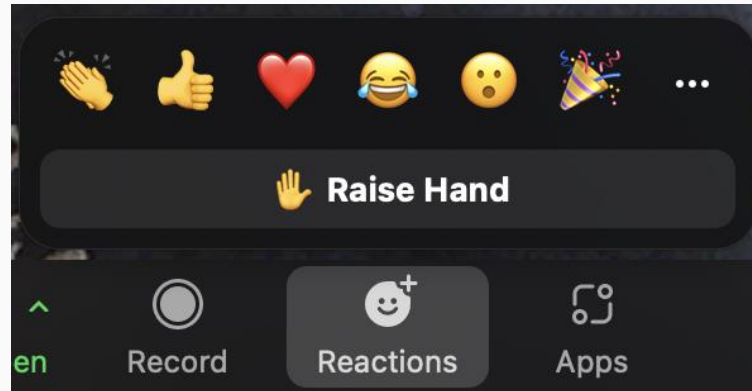
- 00:00 - 00:10 Introduction
- 00:10 - 00:30 Publication policy
- 00:30 - 00:55 Central Health Study Hub
- 00:55 - 01:00 Metadata schema
- 01:00 - 01:05 Recommended standards
- 01:05 - 01:25 Q&A
- 01:25 - 01:30 Feedback

A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains a color gradient of liquid, transitioning from yellow on the left to red on the right. A white rectangular box is overlaid in the center of the plate.

# Practical information

# Practical information

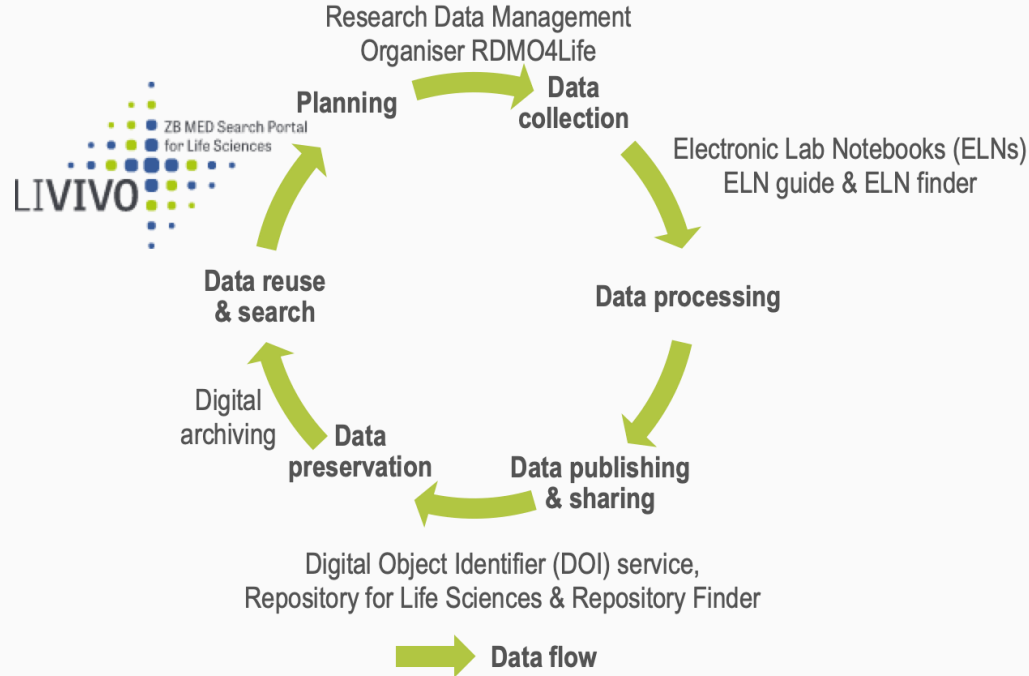
- Slides (with a CC BY license) will be shared after the workshop
- Questions or comments: either in the **chat** or **raise your digital hand**



A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains two rows of yellow liquid and two rows of red liquid. A white text box is overlaid in the center.

# Introduction of our affiliations

# ZB MED - Information Centre for Life Sciences



- **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).

# Research Data Management (RDM) team

- **Birte Lindstädt**
  - Background in Economic Geography and Library & Information Science
  - RDM team leader
  - Involved in NFDI4Health, NFDI4Microbiota and FAIRAgro
- **Julia Fürst**
  - Background in Computational Linguistics & Information Science
  - Member of the RDM team
  - Data steward within NFDI4Health
- **Aliaksandra Shutsko**
  - Background in Computational Linguistics & Information Science
  - Member of the RDM team
  - Data steward within NFDI4Health

# NFDI4Health

**NFDI4Health** = National Research Data Infrastructure for Personal Health Data

- **Who?** A consortium of 17 partners and 48 participants.
- **When?** From October 2020 until October 2025 (might be extended).
- **Why?** Making **data** generated in **clinical trials, epidemiological and public health studies** Findable, Accessible, Interoperable and Accessible (**FAIR**).
- **How?** By providing researchers with **services** to permanently store, semantically enrich and share data in interoperable form and to merge data from different sources.

## Contact details:

- Twitter: [@nfdi4health](https://twitter.com/nfdi4health)
- Email address: [contact@nfdi4health.de](mailto:contact@nfdi4health.de)
- Website: <https://www.nfdi4health.de/en/>



# IEL engagement in RDM / NFDI4Health

## Co-applicant in the NFDI4Health project



1. Task Area 4.3. "Community & Networking - Training and education"
  - Pilot Project on Data Stewardship in nutritional epidemiology
2. Task Area 5.1. - "Use Cases – Nutritional epidemiology" (Measure-Lead: Ute Nöthlings)
  - Main objective: Harmonization and standardization of research data
  - Use of data/metadata/experiences from the DONALD study

A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains two main groups of wells: a 4x8 grid of yellow liquid on the left and a 4x8 grid of red liquid on the right. The plate is transparent, and the liquid levels are visible in each well. A white text box is overlaid in the center of the image.

## Introduction of the participants

# What is your background in RDM?

## Poll

- Do you add **metadata** to your data?
- Have you ever heard of the **FAIR data principles**?
- Have you ever created a **Data Management Plan (DMP)**?
- Have you ever **published** your **data**?

Note for trainers:

The questions of this poll are presented to the participants using the polling feature of Zoom



A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains two rows of yellow liquid and two rows of red liquid. A white rectangular box is overlaid on the center of the plate.

**Research data**

# A definition of research data

**Research data** = “any information that has been collected, observed, generated or created to validate original research findings.”

There is **no consensus** on the definition, the definition varies depending on:

- Disciplines
- Research funders

# General examples of research data

- Documents (e.g. text, Word, PDF)
- Spreadsheets
- Laboratory notebooks, field notebooks, diaries
- Questionnaires, transcripts, codebooks
- Audiotapes, videotapes
- Protein or genetic sequences
- Spectra
- Test responses
- Slides, artifacts, specimens, samples
- Database contents (e.g., video, audio, text, images)
- Models, algorithms, scripts
- Contents of an app (e.g., software)
- Methodologies and workflows
- Standard operating procedures and protocols

# Examples of research data in biomedicine

- Data from **basic research** (e.g. sequencing or -omics data)
- **Electronic Medical Records** (EMRs) and **Electronic Health Records** (EHRs)
- **Patient/disease registries** (e.g. [ENCePP Resources Database](#))
- **Health surveys** (e.g. National Cohort Study ([NAKO](#)))
- **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](#)))

# Examples of documents in biomedicine

- Study protocol (e.g. study plan, study proposal)
- Protocol amendment
- Data dictionary and catalogue
- Participant / patient information sheet (template)
- Informed consent forms (template)
- Manual of operations / Standard operating procedures (SOPs)
- Statistical analysis plan
- Data management plan (DMP)
- Data collection instruments (e.g., Food Frequency Questionnaire)

# Example of metadata in a clinical trials registry

## German Clinical Trials Register (DRKS)

- Open Access
- [Search](#), [register](#) and [share](#) information on clinical trials
- 14,800 studies (+ 2,000/year)
- **Information:** title, short descriptions, inclusion and exclusion criteria, status and outcomes

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 from the chaff – COVID-19 in the Emergency department Score: 0.853   DRKS-ID: DRKS00021675	Recruiting ongoing
2	Title: Retinal Microangiopathy in Patients after Covid-19 Disease Score: 0.851   DRKS-ID: DRKS00022874	Recruiting ongoing
3	Title: COVID-19 Infection Rate in the Construction- and Cleaning Industry Score: 0.826   DRKS-ID: DRKS00023558	Recruiting planned
4	Title: Covid-19 HOspitalized patients RegisTry Score: 0.809   DRKS-ID: DRKS00021575	Recruiting ongoing
5	Title: hospital-based case-control study for efficacy and safety of COVID-19 vaccines Score: 0.809   DRKS-ID: DRKS00025004	Recruiting planned
6	Title: Investigation of COVID-19 associated neurological comorbidities 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 Score: 0.757   DRKS-ID: DRKS00024702	Recruiting ongoing
8	Title: SARS-CoVid-Endothelitis Study via Retinal Vascular Analysis (COVID-19) 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 Score: 0.735   DRKS-ID: DRKS00023913	Recruiting ongoing
10	Title: Safety profile of COVID-19 vaccines and the comparison to COVID-19 symptomatology Score: 0.735   DRKS-ID: DRKS00024800	Recruiting ongoing

# Examples of research data in DONALD



- **Metadata descriptions**
  - Hand books, code plans etc.
  - Registration in Repositories ([DRKS](#), [MDC MICA](#), [Central Health Study Hub](#))
- **SOPs for data collection**
  - For anthropometric measurements, sample collection, urine analyses etc.
- **Participants personal data (not visible to all scientists)**
  - Contact data, signed informed consent, data protection declaration
- **Data sets/Data pool (pseudonymised)**
  - Nutritional data, anthropometrics, health data etc.
- **Bio-samples**
  - Stored urine samples, blood samples
- **Data collection instruments (templates)**
  - Questionnaires, dietary records etc.

A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains two distinct liquid phases: a yellow liquid in the left half and a red liquid in the right half. The wells are arranged in a grid, and the liquid levels are consistent across each column. The background is a light-colored laboratory surface.

## RDM and the research data life cycle

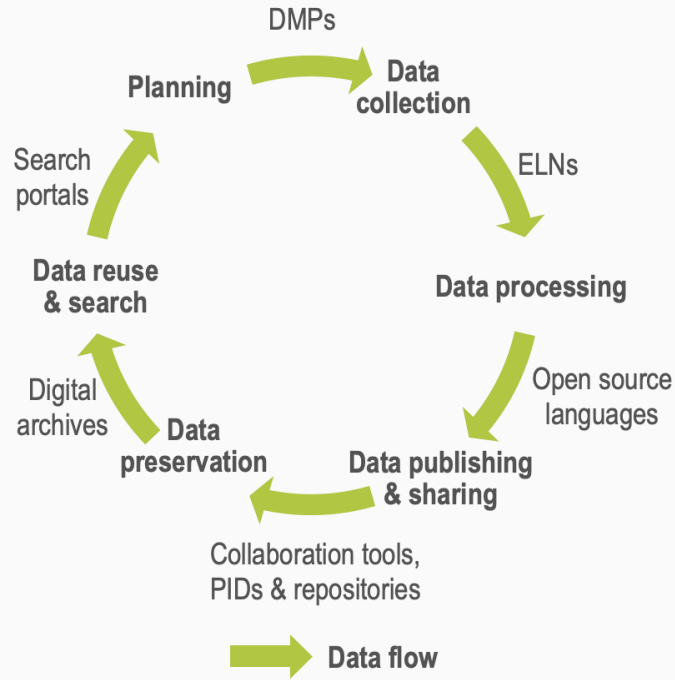
# Definition of RDM

**Research Data Management (RDM)** = “series of steps and methods that aim to make research data usable over the long term”

## Steps:

- Planning
- Data collection
- Data processing (incl. data documentation)
- Data quality control
- Data publishing & safeguarding access to data
- Data preservation (ensuring the long-term interpretability of data)
- Data reuse & search

# Research data life cycle



# Benefits of RDM

## Own interests

- Good scientific practice / GEP
- Knowledge management / transfer
- Prevent data loss
- Saving yourself time in the future

## External interests

- Research funders
- Publishers
- Guidelines of research associations and university management

# Consequences of poor RDM



Science & Justice  
Volume 55, Issue 3, May 2015, Page 218



Retraction notice  
**Retraction notice to A model study into the effects of light and temperature on the degradation of fingerprint constituents**  
[Science and Justice, 54 (2014) 346 - 350]  
Belén González Amorós, M. de Puit  
[Show more](#)  
<https://doi.org/10.1016/j.scjus.2015.04.005>  
**Refers to** Belén González Amorós, M. de Puit  
**RETRACTED: A model study into the effects of light and temperature on the degradatio...**  
Science & Justice, Volume 54, Issue 5, September 2014, Pages 346-350  
[Download PDF](#)

[<](#) Previous article in issue      Next article in issue [>](#)

This article has been retracted: please see Elsevier Policy on Article Withdrawal (<http://www.elsevier.com/locate/withdrawalpolicy>).

This article has been retracted at the request of the authors. The authors identified a inconsistency in the accepted paper and were unable to reproduce the average values that were used for the graphs and tables in the paper, due to the loss of the raw data. This, in turn, means that the authors cannot fulfil the demands of the Association of Dutch Universities and the Royal Dutch Academy of Science in respect to their ethical and research data standards.

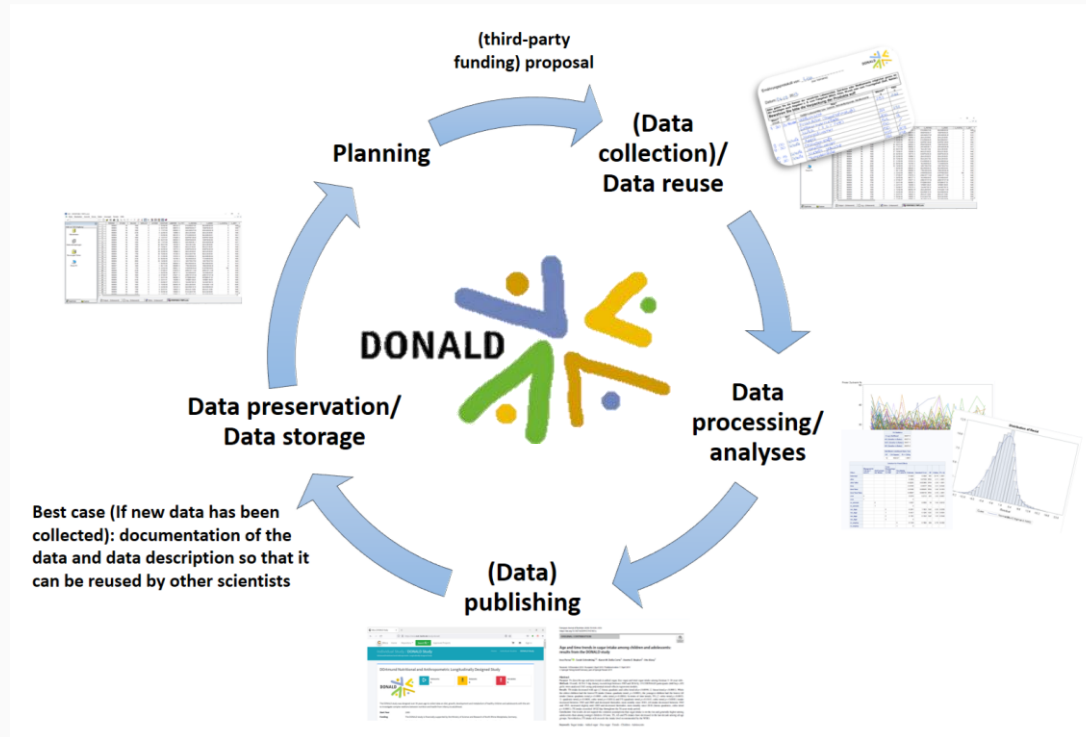
Copyright © 2015 The Chartered Society of Forensic Sciences. Published by Elsevier B.V. All rights reserved.

- Values
    - Inconsistent
    - Non reproducible
  - Loss of raw data
- Authors requested to retract the article

This article has been retracted: please see Elsevier Policy on Article Withdrawal (<http://www.elsevier.com/locate/withdrawalpolicy>).

This article has been retracted at the request of the authors. The authors identified a inconsistency in the accepted paper and were unable to reproduce the average values that were used for the graphs and tables in the paper, due to the loss of the raw data. This, in turn, means that the authors cannot fulfil the demands of the Association of Dutch Universities and the Royal Dutch Academy of Science in respect to their ethical and research data standards.

# Research data life cycle steps adapted to DONALD



A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains two rows of yellow liquid and two rows of red liquid. A white text box is overlaid on the center of the plate.

## Metadata & metadata standards

# Data documentation and metadata

**Metadata** = “highly structured data documentation”

**Metadata** = “data about data”

- **Standardised** information
- **Structured** information
- Subset of **documentation**: describes, explains, locates, makes it easier to retrieve, use, manage an information resource
- **Human-** and **machine-readable**

**Examples of metadata**

- Name/Title
- Topic
- Description of input and output (parameters or format)
- Address/geospatial information
- Licensing information

# Types of metadata

<b>Descriptive</b>	<b>vs.</b>	<b>Technical</b>
e.g. resource identifier, title, author, date of publication, subject, publisher, description		e.g. physical database tables, access permissions, data models, backup rule
<b>Project-level/Study-level</b>	<b>vs.</b>	<b>Data-level</b>
e.g. project title, summary/description, institutions involved, funds, methods, coverage, creators, contributors, identifiers		e.g. unique ID, variable names, labels and descriptions, missing values code, unique project ID, date the file was created, file path
<b>General</b>	<b>vs.</b>	<b>Discipline-specific</b>
e.g. description, file name and format, software and hardware used to create the files		e.g. reagent, technical, experimental, analytical, dataset-level

# Example of general metadata standard

**Dublin Core (DC)** = domain-agnostic, basic, widely used metadata standard.

- **Project level:** applicable to all datasets.
- **Data level:** there are discipline-specific standards to branch into (e.g. Ecological Metadata Language ([EML](#)))
- **DC Simple:** 15 metadata elements
- **DC Qualified Elements:** 15 metadata elements + additional ones

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

# Examples of metadata standards in (bio-)medicine

- **To report:**
  - Clinical data: [SNOMED CT](#) (in [Germany](#))
  - Diseases and health conditions: [ICD](#)
- **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](#)

# Ontologies and controlled vocabularies

- **Controlled vocabulary** – “prescribed list of terms, headings or codes, each representing a concept”
- **Ontology** - “controlled *terminology invoking formal semantic relationships between and among concepts*, manifested as a type of description logic, which is a subset of first-order predicate logic, chosen to accommodate computational tractability” or according to ISO 21526:2019 “a conceptualization of a domain”
  - **Terminology** – A) set of preferred or official terms in a domain. B) A terminology may be a systematic nomenclature supported by a centralized body or as simple as the common usage that arises in a specific community of practice or C) according to ISO/TS 21564:2019 a “structured human and machine-readable representation of concepts”.
- Describe **concepts** and **relationships** within a knowledge domain.
- Serve a **basis** for metadata schemas
- Enable both **humans** and **computers** to understand data
- Make data **more reusable** and **interoperable**
- In praxis, no clear-cut division between the terms

# Importance of metadata standards

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

37 year old male	initial phase male	male fetus	six males mixed
600 yr. old male	m	male plant	stallion
adult male	make	male, 8 weeks old	steer
bull	makle	male, castrated	sterile male
castrated male	mal e	male, pooled	strictly male
cm	male	males	tetraploide male
dioecious male	male (7-2872)	man	type i males
diploid male	male (7-3074)	men	type ii males
drone	male (m-a)	normale male	virgin male
engorged male	male (m-o)	ram	winged and wingless males
fertile male	male caucasian	rooster	young male
four males mixed	male child	s1 male sterile	
individual male	male fertile	sex: male	

male (note: this sample was originally provided as a 'female' sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a male individual)

A selection of the many varied words and phrases scientists submitters to the EBI's ENA resource



# How to find appropriate metadata standards?

- Use **specialised search platforms**:
  - RDA [Metadata Standards Directory](#)
  - [FAIRsharing.org](#) (see “Standards” And “Collections”)
    - [Search results](#) for the subject “*nutritional science*”
  - Basel Register of Thesauri, Ontologies & Classifications ([BARTOC](#))
  - [BioPortal](#) - Repository of biomedical ontologies
- Check **submission/publication requirements of a repository** suitable for your project/data, e.g.
  - [Publication Policy](#) of the [NFDI4Health Study Hub COVID-19](#) refers to the [Metadata Schema](#)
  - DRKS-Study submission requirements

# Practical advice: (Meta-)data documentation

## Tools for documentation

- Electronic Lab Notebooks (**ELNs**)
- Electronic Data Capture (**EDC**) systems
- Laboratory Information Management Systems (**LIMS**)
- Online **platforms for collaborative research** and file sharing services (such as [OSF](#))

## Resulting files

- README files  
(.txt or .xlsx (with [Colectica](#)))
- Codebook
- Data dictionary ([how to?](#))
- Data list

Further examples of relevant tools and resources: [ELIXIR RDMKit](#)

# Metadata handling & metadata standards in DONALD

- **Metadata handling**

- Data handling according to: [Declaration of Helsinki](#), Data protection regulations ([EU](#), [national](#), [regional \(NRW\)](#)), [good epidemiological practices \(GEP\)](#)

- **Metadata standards**

- Internal standards
- Orientation towards official definitions e.g. of nutrients based on the Bundeslebensmittelschlüssel or definitions of professional societies
- Use of ontologies/standards e.g. ICD-Codes (Diseases)



A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains two distinct groups of samples: a 4x8 grid of yellow liquid on the left and a 4x8 grid of red liquid on the right. The background is a light-colored laboratory surface.

## **FAIR data principles**

# Definition of the FAIR data principles

**FAIR data principles** = a concise and measurable set of principles that may act as a guideline for those wishing to enhance the reusability of their data holdings.

**FAIR** stands for:

- **F**indability
- **A**ccessibility
- **I**nteroperability
- **R**eusability

# 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** (e.g. data repository)

# To be **A**ccessible

- (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

FAIR  $\neq$  FOIR (O = Open)

*“As Open as Possible, as Closed as Necessary”*

# To be Interoperable

**Data interoperability** = ability of a dataset to work with other datasets or systems without special effort on the part of the user.

- (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 dataset builds on another one, properly citing all datasets)

# 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**

# Exercise: FAIR or not FAIR?

**Case:** in biomedicine, 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.

Note for trainers:

This question is presented to the participants. They have approx. 5 minutes to think about it. They are then asked to give and explain their answer orally.

N.B.: Not every participant is expected to answer, just collect a few comments.

# Resources

- **Learning resources**

- [FAIR cookbook](#) [ELIXIR]
- [FAIR in \(biological\) practice](#) [The Carpentries Incubator]
- [FAIR sharing and access](#) [MANTRA, The University of Edinburgh]
- [How to be FAIR with your data](#) [Engelhardt et al. 2022]

- **How to make data FAIR?**

- [PARTHENOS Guidelines to FAIRify data management and make data reusable](#)
- [Preparing data for sharing: The FAIR principles](#)
- [Top 10 FAIR Data & Software Things](#)

- **How to assess the FAIRness of your datasets?**

- [How FAIR are your data?](#) (checklist) [Jones and Grootveld 2017]
- Self-Assessment Tool to Improve the FAIRness of Your Dataset ([SATIFYD](#)) [DANS]
- [FAIR data maturity model indicators](#) (spreadsheet) [Bahim et al. 2020, Table 1]

A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains two rows of yellow liquid and two rows of red liquid. The background is a light blue surface.

## Policies & guidelines on managing research data

# Policies and guidelines (for RDM)

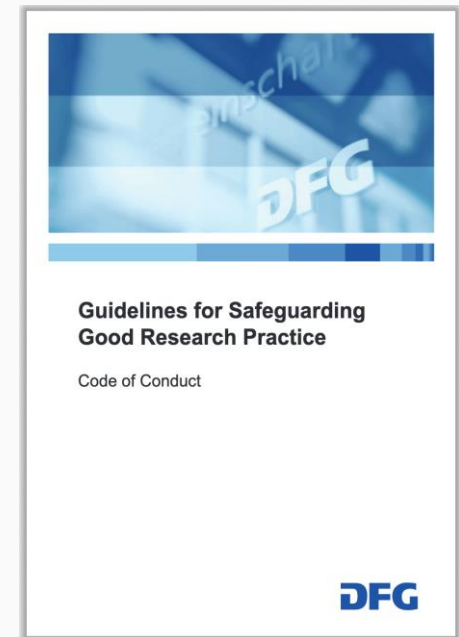
## Types of policies & guidelines:

- ▶ General **vs.** Discipline-specific
- ▶ General **vs.** Institutional **vs.** Research group-specific (e.g. to specify file naming conventions or how to organize research data)

# RDM-relevant DFG's guidelines

RDM-relevant guidelines set out in the DFG's [Guidelines for Safeguarding Good Research Practice](#)

- Guideline 7: Quality assurance across phases
- Guideline 10: Legal and ethical frameworks, rights of use
- Guideline 11: Methods and standards
- Guideline 12: Documentation
- Guideline 13: Establishing public access to research results
- Guideline 14: Authorship
- Guideline 15: Publication organ
- Guideline 17: Archiving



# RDM-relevant DFG's guidelines

What to include in your funding proposal(s)?

- [Checklist](#) for the appropriate **handling of research data** in connection with DFG projects
- [Specification of Requirements Relating to the Handling of Research Data in Funding Proposals](#)

# Discipline-specific policies & guidelines

- **Life sciences**
  - [ELIXIR RDM Toolkit for Life Sciences](#)
- **Health sciences**
  - Guidelines of the [German Society for Epidemiology \(DGEpi\) e.V.](#)
  - [ICH E6 \(R2\) Good clinical practice](#)
  - [Publication policy of the NFDI4Health](#)
  - The German [Medical Informatics Initiative](#) (MII)'s recommendations

# Good Epidemiologic Practice (GEP)

## RDM-relevant guidelines

- Guideline 4: Consent
- Guideline 5.4: Data collection
- Guideline 6: Data management and documentation
- Guideline 7: 10 years long term storing for data
- Guideline 7.3: Storing software in reproducible form
- Guideline 11.2: Accessibility of instruments

# Institutional guidelines

- **Regulations for Safeguarding Good Research Practice at the University of Bonn**
  - <https://www.uni-bonn.de/en/research-and-teaching/quality-assurance-in-research-and-teaching/good-research-practice>
- **Servicestelle Forschungsdaten at the University of Bonn**
  - <https://www.forschungsdaten.uni-bonn.de/de>

# Guidelines within a research group (DONALD study)

- Employees have to sign a data secrecy obligation and a confidentiality declaration
- Working according to...
  - the [Declaration of Helsinki](#)
  - data protection regulations ([EU](#), [national](#), [regional \(NRW\)](#))
  - DONALD data protection concept
  - [good epidemiological practices \(GEP\)](#)



# Q&A



Photo by [Jon Tyson](#) 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](mailto:forschungsdaten@zbmed.de)

[www.zbmed.de](http://www.zbmed.de)

**NFDI4Health –  
National Research Data  
Infrastructure for Personal  
Health Data**

[contact@nfdi4health.de](mailto:contact@nfdi4health.de)  
[studyregistration@zbmed.de](mailto:studyregistration@zbmed.de)

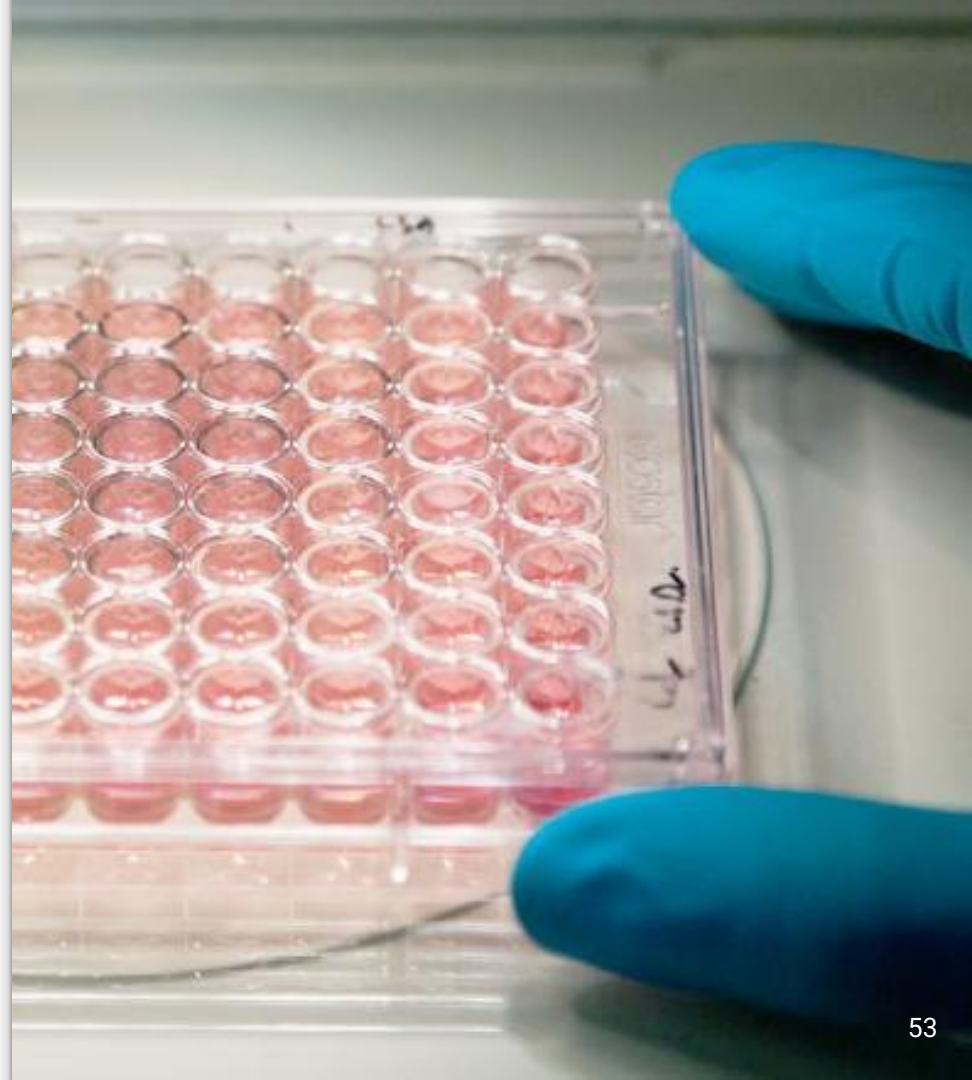
[www.nfdi4health.de](http://www.nfdi4health.de)

**IEL – Nutritional  
Epidemiology  
University of Bonn**

Friedrich-Hirzebruch-Allee 7  
53115 Bonn

[epi@uni-bonn.de](mailto:epi@uni-bonn.de)

<https://www.ernaehrungsepidemiologie.uni-bonn.de/>





**NFDI4Health**

# ONLINE TRAINING WORKSHOP ON RESEARCH DATA MANAGEMENT IN (BIO-)MEDICINE

J. Fürst<sup>1</sup>, B. Lindstädt<sup>1</sup>, I. Perrar<sup>2</sup>, A. Shutsko<sup>1</sup> & J. Vandendorpe<sup>1</sup>

<sup>1</sup> ZB MED - Information Centre for Life Sciences

<sup>2</sup> Institut für Ernährungs- und Lebensmittelwissenschaften, University Bonn



Photo by ZB MED



Unless stated otherwise within the sources cited in this work, this work is licensed under the Creative Commons Attribution 4.0 International Public License (CC BY 4.0). To view a copy of the license, visit <https://creativecommons.org/licenses/by/4.0/legalcode>



# Schedule

## Day 1- Introduction to Research data management (RDM) - Fundamental concepts - Duration: 90mins

- 00:00 - 00:25 Welcome + Introduction of affiliation
- 00:25 - 00:30 Research data
- 00:30 - 00:45 Research data life cycle
- 00:45 - 00:55 Metadata
- 00:55 - 01:05 FAIR data principles
- 01:05 - 01:10 Guidelines on RDM
- 01:10 - 01:25 Q&A
- 01:25 - 01:30 Feedback

## Day 2 - A step forward to making data FAIR with NFDI4Health – Duration: 90mins

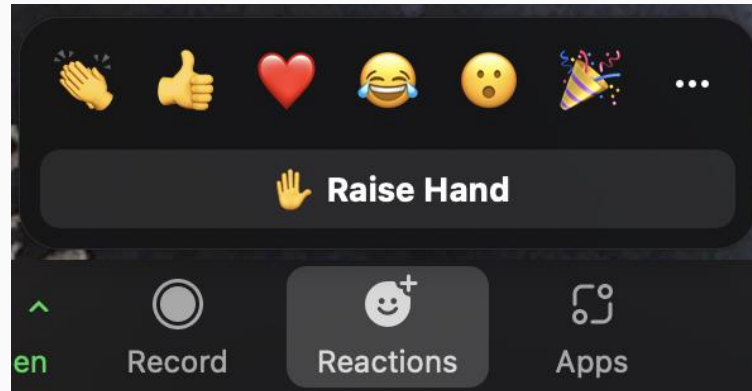
- 00:00 - 00:10 Introduction
- 00:10 - 00:30 Publication policy
- 00:30 - 00:55 Central Health Study Hub
- 00:55 - 01:00 Metadata schema
- 01:00 - 01:05 Recommended standards
- 01:05 - 01:25 Q&A
- 01:25 - 01:30 Feedback

A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains a color gradient of liquid, transitioning from yellow on the left to red on the right. A white rectangular box is superimposed over the center of the plate.

## Practical information

# Practical information

- Slides (with a CC BY license) will be shared after the workshop
- Questions or comments: either in the **chat** or **raise your digital hand**



A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains a color gradient of liquid, transitioning from yellow on the left to red on the right. The text 'NFDI4Health' is overlaid in the center.

**NFDI4Health**

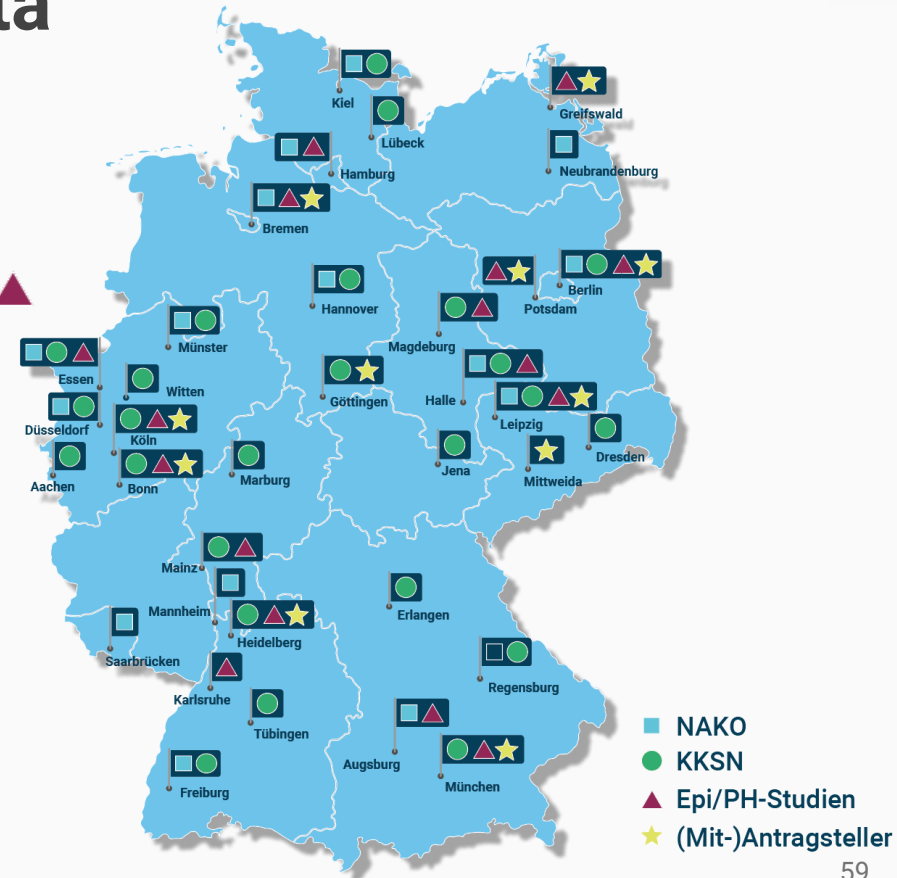
# NFDI4Health: Research data

## Primary focus

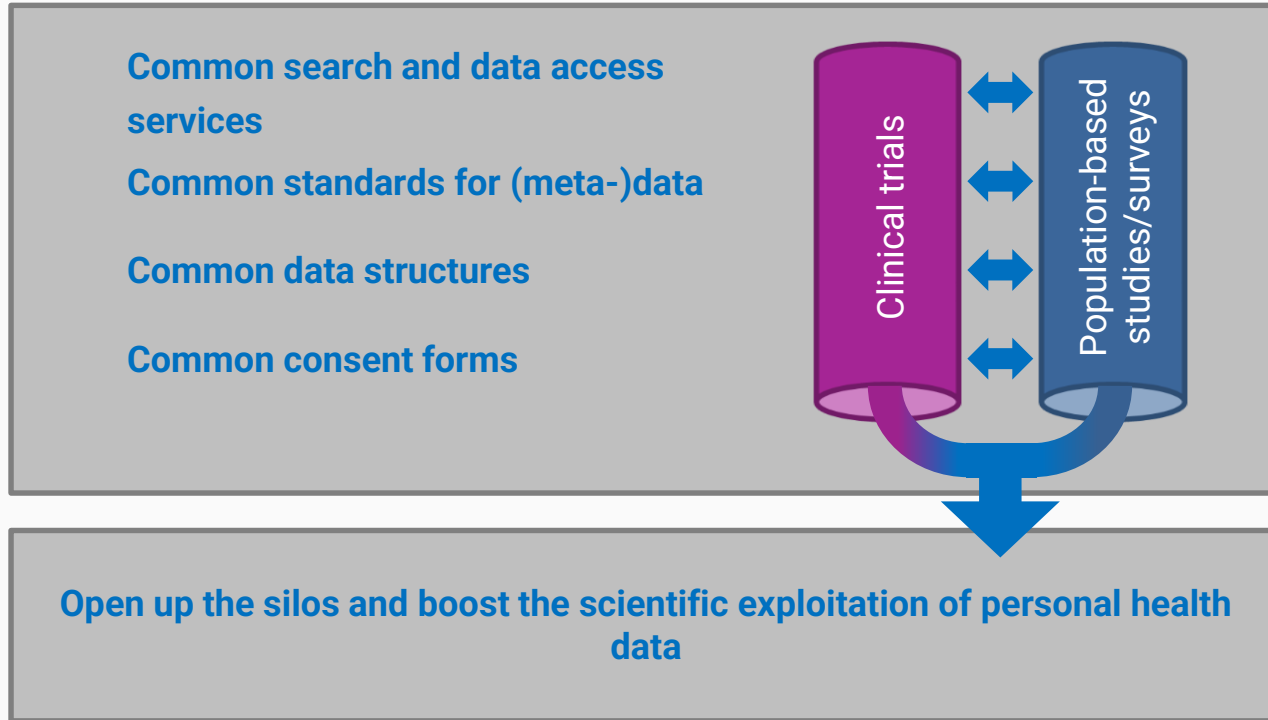
- **Epidemiological** and **public health** studies
  - 26 local studies with > 400,000 participants ▲
  - German National Cohort Study (NAKO) ■
- **Clinical** studies
  - 24 university study centers ●

## Secondary focus

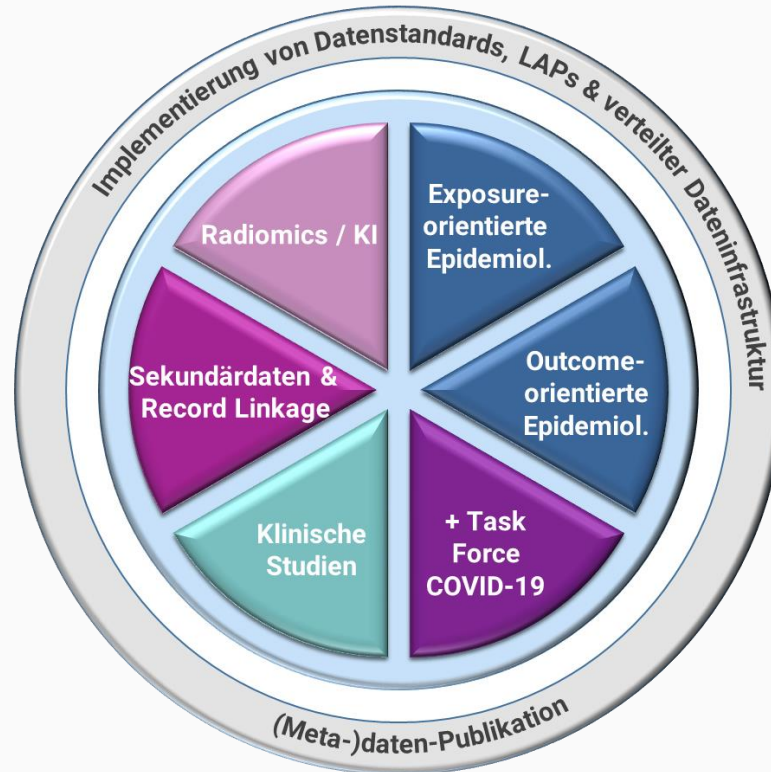
- Health registries
- Administrative **health data banks**



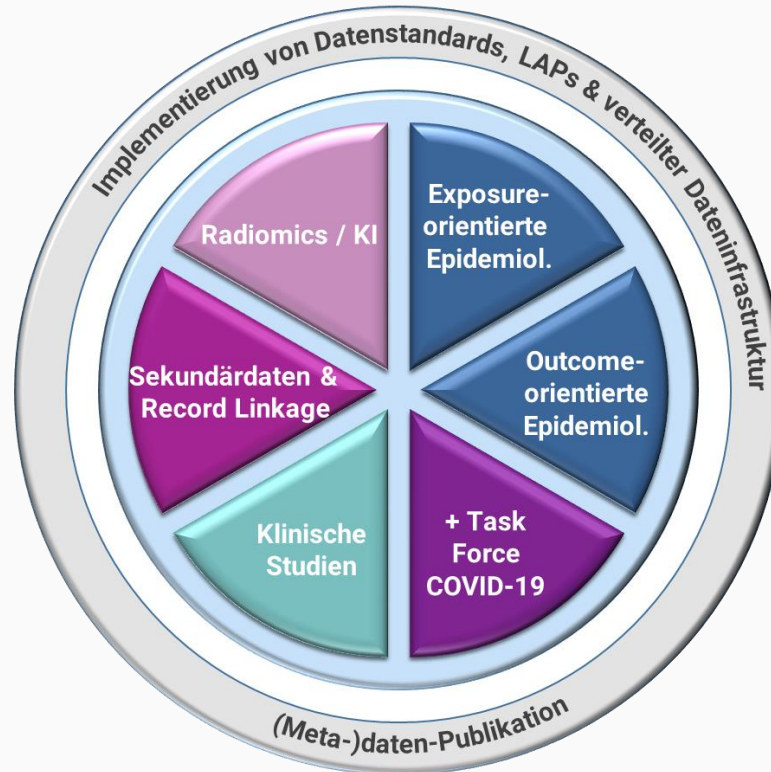
# Amalgamation of epidemiology, public health and clinical research



# NFDI4Health: Use cases



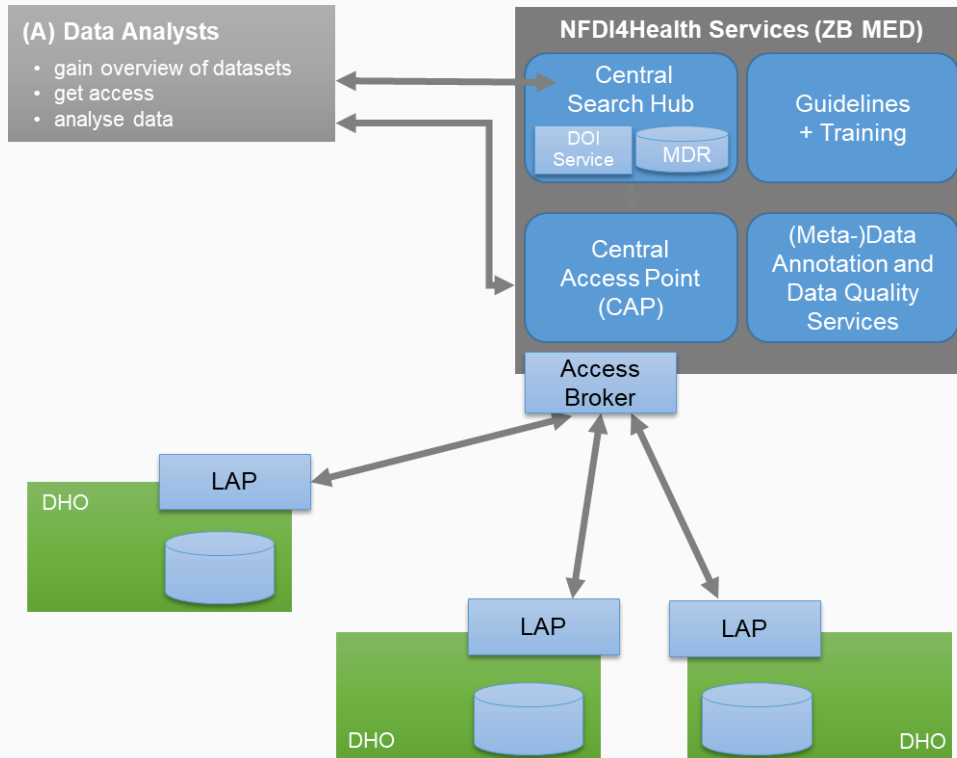
# NFDI4Health: Use cases



# NFDI4Health: Strategy and offer

- **Improved discoverability of personal health data**
    - *Findability of studies via publications of research data on the **Central Search Hub** (NFDI4Health Central Health Study Hub)*
    - *Standardization of metadata → improvement of interoperability*
  - **Federated framework for data access and data analysis (in preparation)**
    - *Implementation of a **higher-level data access and data use process (Central Access Point)***
      - **Access control** remains with research institutions/primary data collectors
    - *Provision of software modules for connection to the NFDI4Health-network for **distributed data analysis** (DataShield and Personal Health Train clients)*
- **Close involvement of the community** to achieve sustainability

# NFDI4Health: Services overview



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

Source: NFDI4Health Proposal,  
doi: [10.4126/frl01-006430386](https://doi.org/10.4126/frl01-006430386)

# NFDI4Health: Findability of studies



Data  
collections



- Metadata
- Protocols
- Data catalogues
- Questionnaires
- ...

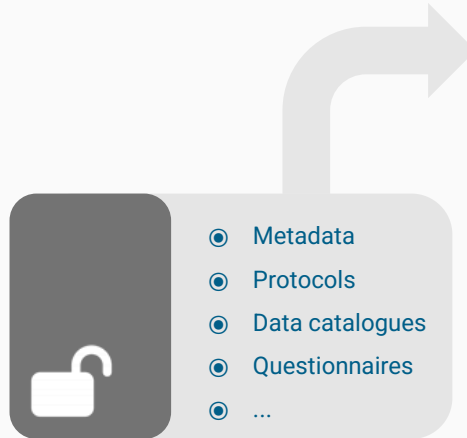
Research institute

involved in collection and analysis of personal  
health data

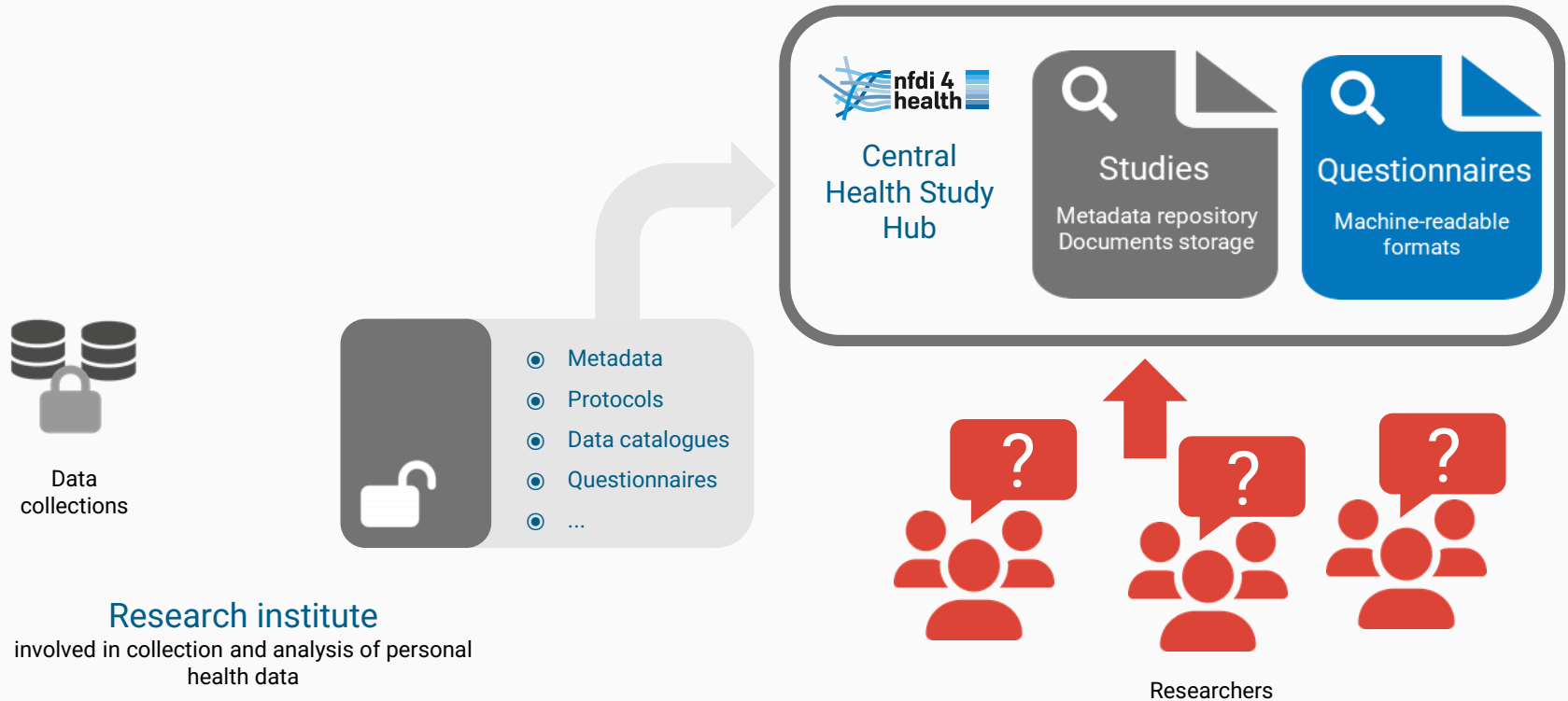
# NFDI4Health: Findability of studies



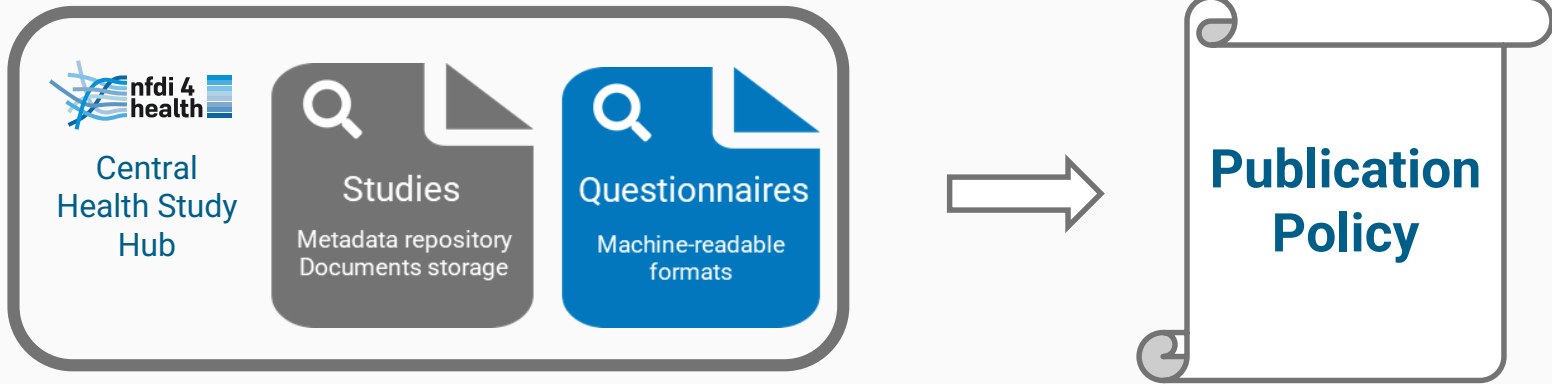
Data  
collections



# NFDI4Health: Findability of studies



# NFDI4Health: Findability of studies



# NFDI4Health: Strategy and offer

- Central Health Study Hub (search portal)
  - Central Access Point
- Distributed Data Analysis Infrastructure (DataSHIELD, PHT)
  - Synthetic data generation (concept)
  - Best practice of record/data linkage

**Data  
reuse &  
search**

**Planning**

- Templates of data management plans (DMPs)
- (Meta-)data standardisation guidelines
- Publication policy

**Data  
collection**

- Terminology and metadata annotation services
- Data quality services
- “Legal map” for processing personal data

**Data  
processing**

- Distributed data analysis infrastructure (DataSHIELD, Personal Health Train (PHT))

**Data  
publishing &  
sharing**

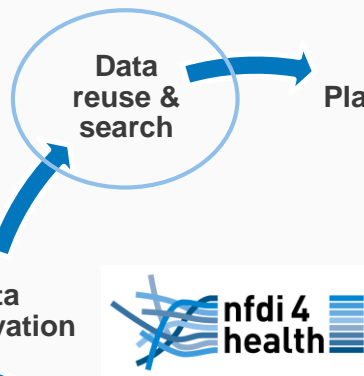
- Central Health Study Hub (repository)
  - DOIs assignment
  - Publication policy
  - Local Data Hub

**Data  
preservation**



# NFDI4Health: Strategy and offer

- Central Health Study Hub (search portal)
  - Central Access Point
- Distributed Data Analysis Infrastructure (DataSHIELD, PHT)
  - Synthetic data generation (concept)
  - Best practice of record/data linkage



**Planning**

- Templates of data management plans (DMPs)
- (Meta-)data standardisation guidelines
- Publication policy

**Data collection**

- Terminology and metadata annotation services
- Data quality services
- "Legal map" for processing personal data

**Data processing**

- Distributed data analysis infrastructure (DataSHIELD, Personal Health Train (PHT))

**Data publishing & sharing**

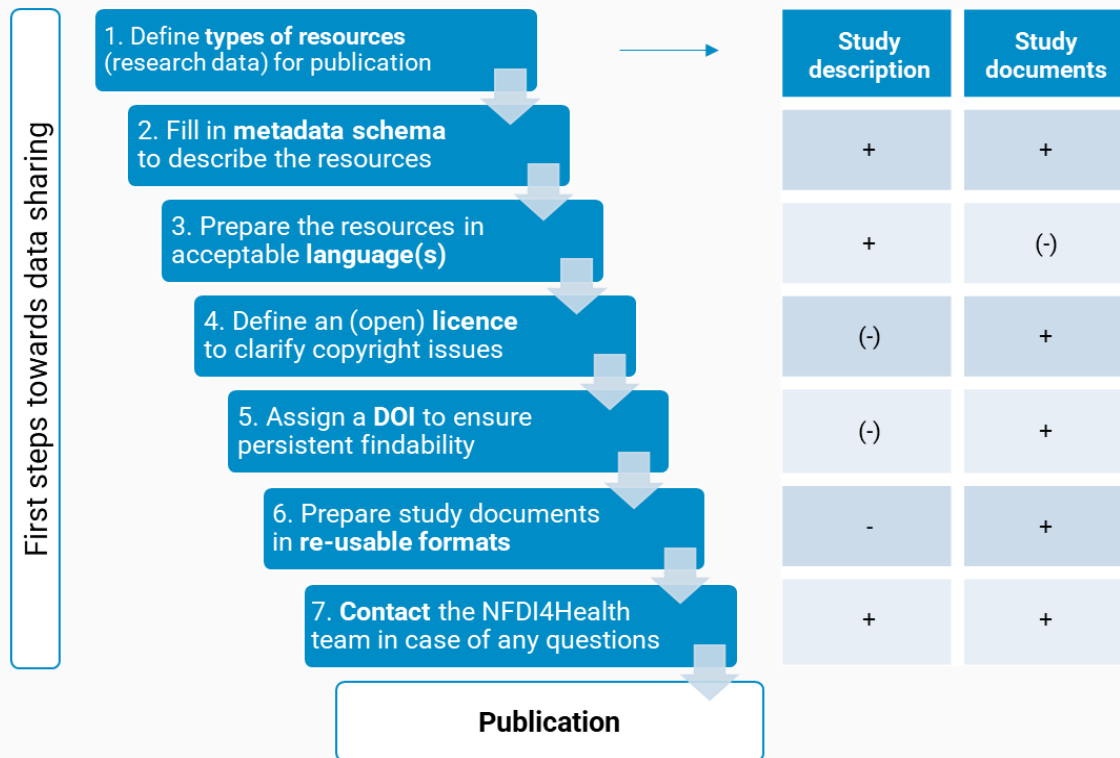
- Central Health Study Hub (repository)
  - DOIs assignment
  - Publication policy
  - Local Data Hub

**Data preservation**

A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains a color gradient of liquid, transitioning from yellow on the left to red on the right. A white rectangular box is superimposed over the center of the plate.

# Publication policy

# NFDI4Health Publication policy



# Publication policy: Types of resources

- What are the constituent parts of a data publication? What should I publish?
  - Study descriptions (= study-level metadata)
  - Study documents + metadata
    - e.g., study protocols, data dictionaries, templates of questionnaires and consent forms, etc.
- **No personal health data** is expected to be published (!)

# Publication policy: Metadata schema

- How should the resources (studies and study documents) be described to become discoverable?
  - **Metadata schema of the NFDI4Health**
    - Based on other international standards
    - Allows provision of:
      - generic information for all types of resources
      - further detailed information about design of interventional and observational studies

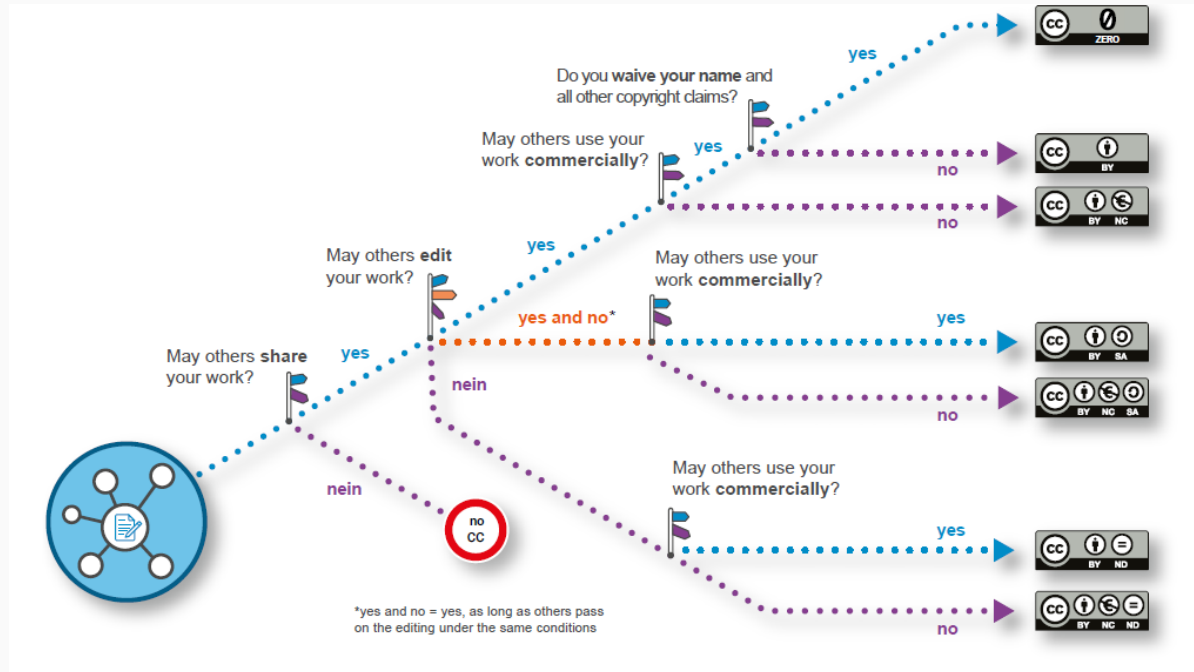
# Publication policy: Languages

- In what language(s) should the resources (studies and study documents) be provided?
  - Study descriptions (= study-level metadata)
    - Preferably in English, but German is also accepted
  - Study documents + metadata
    - Metadata : Preferably in English, but German is also accepted
    - study protocols, data dictionaries, etc.: not limited to English or German

# Publication policy: Licence

- May the resources be reused by others? Under what conditions?
  - **Study descriptions** are “pure” metadata and no subject to copyright → no licence is required, information can be freely used without being licenced
  - **Study documents** are copyrighted works and require a licence to be reused by third parties (i.e. make a copy, redistribute, edit, post online)
    - Open licences are recommended, e.g. **Creative Commons licences**:
      - CC BY 4.0
      - CC BY-NC 4.0
      - CC BY-SA 4.0
      - CC BY-NC-SA 4.0

# Publication policy: How to choose a licence



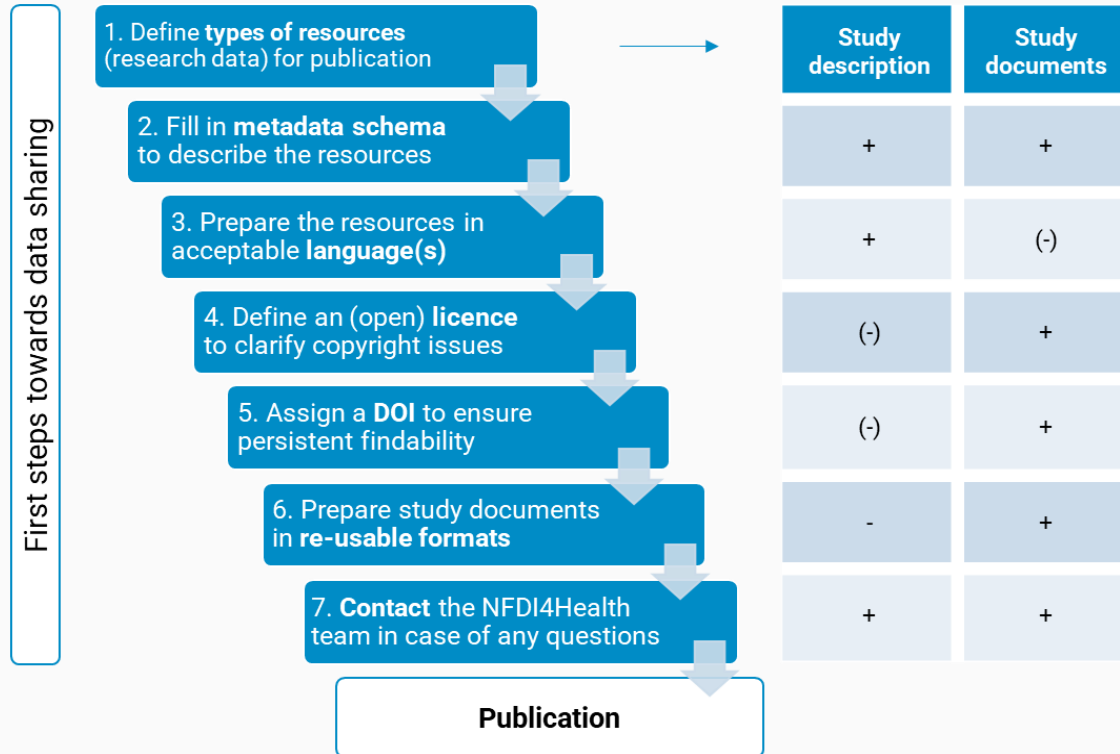
# Publication policy: Persistent identifiers

- Are the resources citable and persistently findable?
  - **Persistent identifier:**
    - Digital Object Identifier (DOI):
      - Used to cite and link physical, digital or abstract objects
      - Identifier refers to object itself
        - Remain valid regardless of URL changes

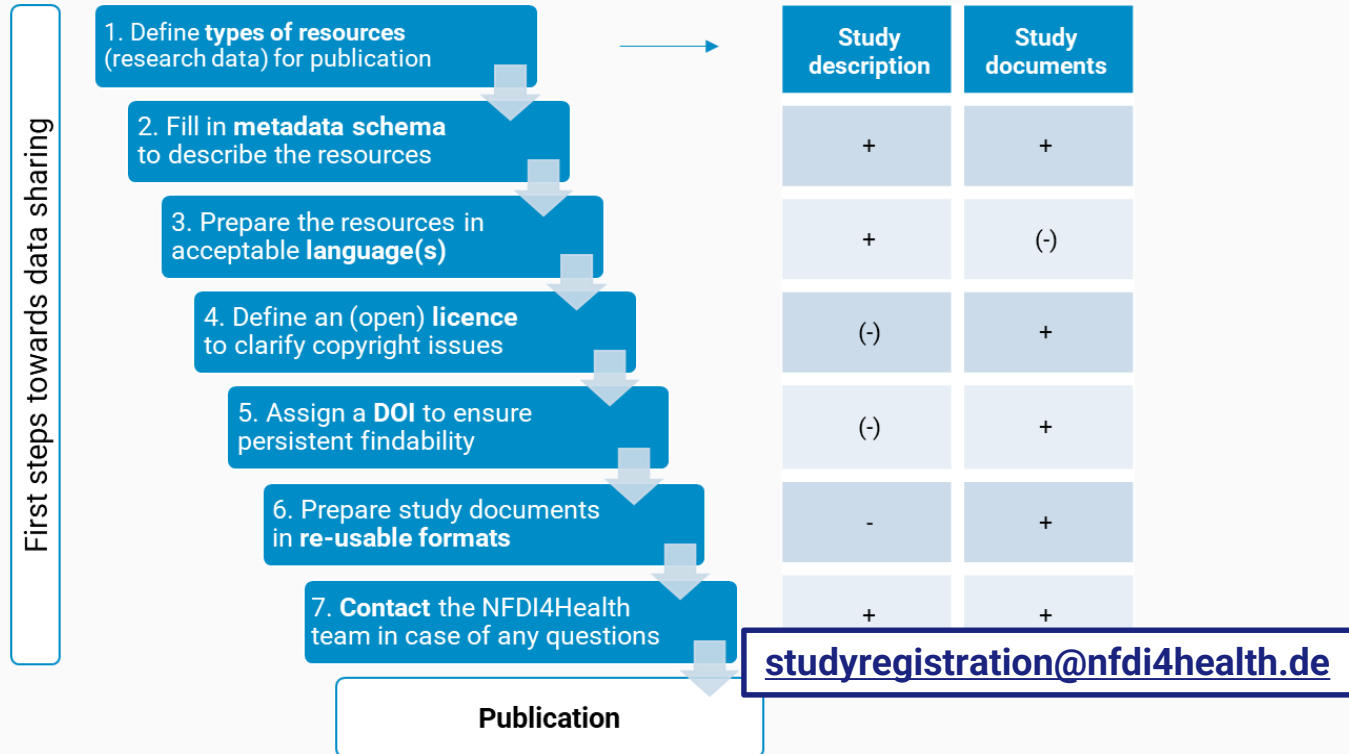
# Publication policy: Reusable formats

- Is the reusability of the documents guaranteed (also in the long term)?
  - **File formats:**
    - **Preferred** formats: open (non-proprietary), machine-readable and machine-actionable –actionable, e.g.:
      - .xml, .txt, .csv, .json, PDF(A), .rdf, etc.
    - Non-machine-readable and proprietary formats are also acceptable

# NFDI4Health: Publication policy



# NFDI4Health: Publication policy



A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains two distinct groups of samples: a 4x8 grid of yellow liquid on the left and a 4x8 grid of red liquid on the right. The background is a light-colored laboratory surface.

# NFDI4Health Central Health Study Hub

# German Central Health Study Hub COVID-19: Prototype of the NFDI4Health Central Health Study Hub

- **Content (types of research data):** *Study metadata & study documents* from clinical, epidemiological and public health studies
- **Target audience:** Researchers who would like to *find and/or share* their research data

The screenshot displays the search interface of the German Central Health Study Hub COVID-19. At the top, a header states: "Begin your search. 1305 studies, instruments and other documents are there to explore". Below this is a search bar with the placeholder text "Type to search" and a magnifying glass icon, followed by a blue "Explore" button. Under the search bar, a section titled "The following syntax is supported:" lists search rules: "whitespace as split characters, each token is combined with a disjunction.", "signifies AND operation", "signifies OR operation", "negates a single token", and "wraps a number of tokens to signify a phrase for searching". Below this is a light blue box with a question mark icon and the text "Your data is not yet displayed?". It contains a message: "Is your study not displayed yet? Please contact us via [studyregistration@nfdi4health.de](mailto: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 [publication policy](#) and flyer [here](#)." At the bottom, there are two blue buttons: "1259 Studies" with a sub-link "Explore Studies →" and "27 Data dictionary" with a sub-link "Explore Data dictionary →".

Begin your search. 1305 studies, instruments and other documents are there to explore

Type to search Explore

The following syntax is supported:

- whitespace as split characters, each token is combined with a disjunction.
- + signifies AND operation
- | signifies OR operation
- - negates a single token
- " wraps a number of tokens to signify a phrase for searching

Your data is not yet displayed?

Is your study not displayed yet? Please contact us via [studyregistration@nfdi4health.de](mailto: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 [publication policy](#) and flyer [here](#).

**1259 Studies**  
Explore Studies →

**27 Data dictionary**  
Explore Data dictionary →

# German Central Health Study Hub COVID-19: Prototype of the NFDI4Health Central Health Study Hub


## Live Demo

**Begin your search. 1305 studies, instruments and other documents are there to explore**

Explore

The following syntax is supported:

- whitespace as split characters, each token is combined with a disjunction.
- + signifies AND operation
- | signifies OR operation
- - negates a single token
- " wraps a number of tokens to signify a phrase for searching

 Your data is not yet displayed?

Is your study not displayed yet? Please contact us via [studyregistration@nfdi4health.de](mailto: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 [publication policy](#) and flyer [here](#).

**1259 Studies**

Explore Studies →

**27 Data dictionary**

Explore Data dictionary →

# NFDI4Health Central Health Study Hub: Metadata entry mask

1

2

3

General Information

Roles

Related Resources

Type of the resource ⓘ (Mandatory)

Select one value from the list.

Language(s) of the Resource ⓘ

Select all that apply.

Web page ⓘ

Title(s)/name(s) of the Resource

Title/name ⓘ (Mandatory)

Language ⓘ (Mandatory)



EN (English)

Please provide at least one title/name of the resource. If the original title is not in English, please also provide an English translation of the original title.

A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains two rows of yellow liquid and two rows of red liquid. The background is a light gray surface.


## Metadata schema of the NFDI4Health


# NFDI4Health Metadata Schema


 Metadata schema of the NFDI4Health and t...  
 Metadata schema of the NFDI4Health and t...



## Metadata schema of the NFDI4Health and the NFDI4Health Task Force COVID-19 (V2\_1)

Shutsko, Aliaksandra | Schmidt, Carsten Oliver | Klopfenstein, Sophie Anne Ines | Darms, Johannes | Golebiewski, Martin | Vorisek, Carina Nina | NFDI4Health Task Force COVID-19 | NFDI4Health

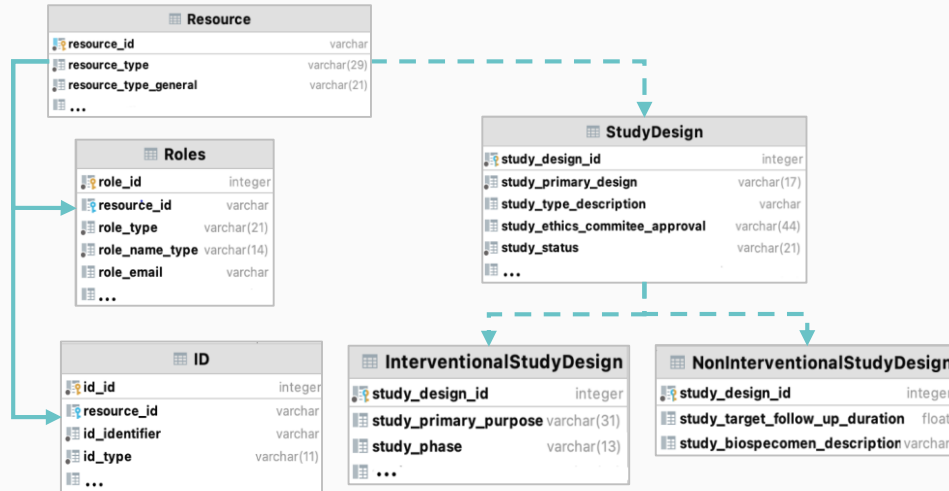


Download 

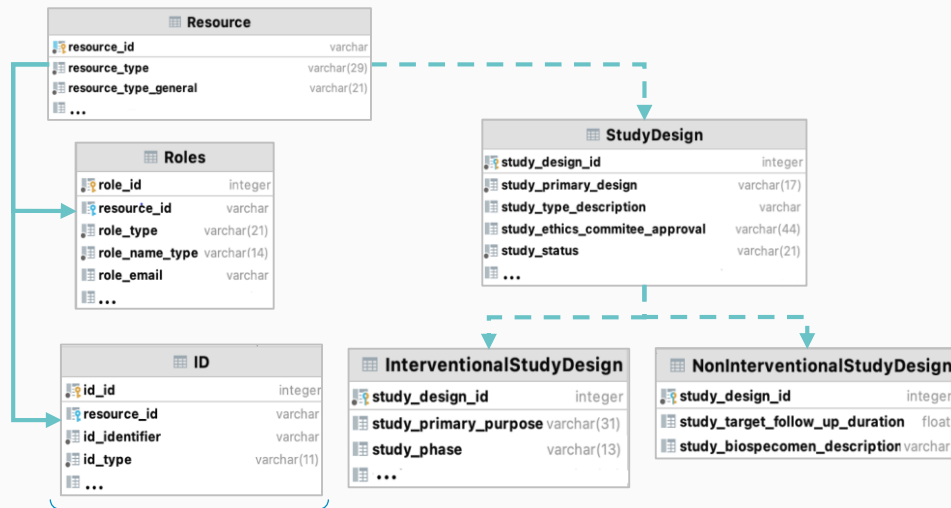
NFDI4Health (TF COVID-19)\_Metadata schema\_V2\_1.xlsx 169,25KB 

Art der Datei	Datensatz
Publikationstyp	Forschungsdaten
FRL-Sammlung	NFDI4Health - Nationale Forschungsdateninfrastruktur für personenbezogene Gesundheitsdaten (Projekt)   
Copyrightjahr	2022
Lizenz	CC BY 4.0   
Sprache der Publikation	Englisch
Sacherschließung	COVID-19 Metadata Metadata model Clinical trials FAIR principles Standard Public Health Epidemiology

# NFDI4Health Metadata schema: Structure and used standards



# NFDI4Health Metadata schema: Structure and used standards

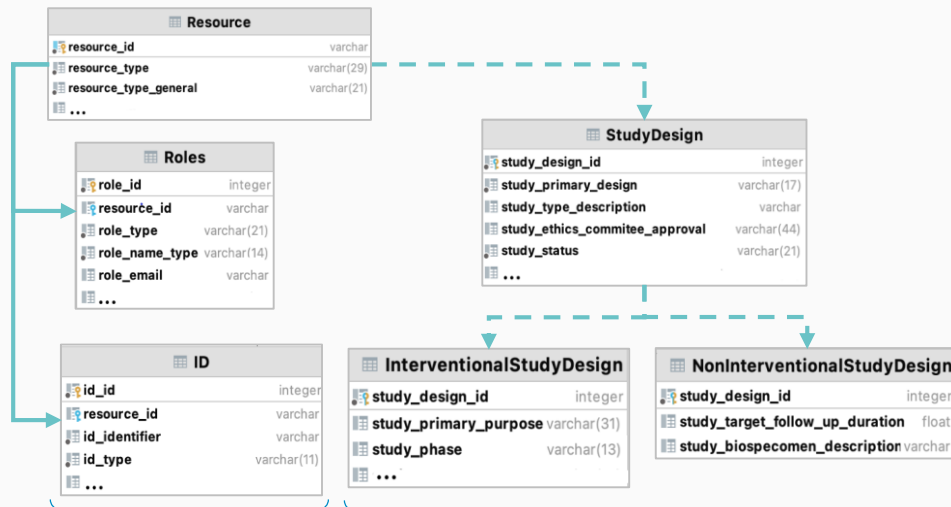


Reference  
data  
models:

*Generic metadata  
standard:*

- [DataCite](#),  
[DublinCore], [DDI](#)

# NFDI4Health Metadata schema: Structure and used standards



Reference  
data  
models:

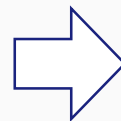
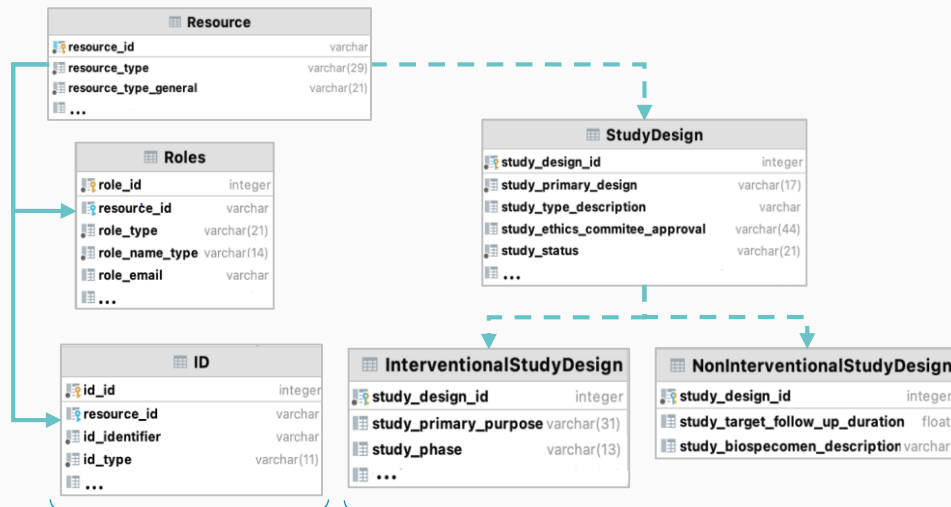
*Generic metadata  
standard:*

- [DataCite](#),  
[DublinCore], [DDI](#)

*Health study specific data models:*

- Clinical trials registries models: [ICTRP](#), [DRKS](#),  
[ClinicalTrials.Gov](#)
- Epidemiological study models: [MIABIS](#), [Maelstrom](#) data  
model, minimal requirements model from [ENPADAS](#)
  - Mapped to: [FHIR](#), [SNOMED CT](#)

# NFDI4Health Metadata schema: Advantages



- Description of **all** types of **resources** (e.g., studies, study documents) **at a general level**
- Provision of **detailed information** depending on resource type
- DOI assignment for all types of resources
- Discoverability & citability of resources

Reference data models:

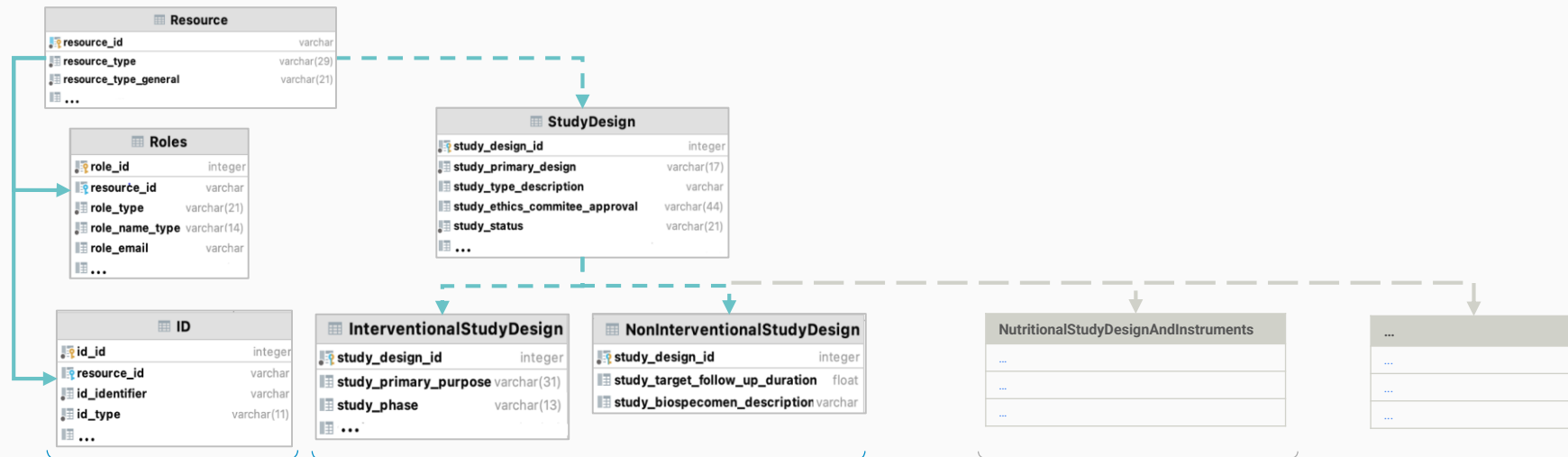
Generic metadata standard:

- [DataCite](#), [DublinCore], [DDI](#)

Health study specific data models:

- Clinical trials registries models: [ICTRP](#), [DRKS](#), [ClinicalTrials.Gov](#)
- Epidemiological study models: [MIABIS](#), [Maelstrom](#) data model, minimal requirements model from [ENPADASI](#)
  - Mapped to: [FHIR](#), [SNOMED CT](#)

# NFDI4Health Metadata schema: Structure and used standards



Reference data models:

Generic metadata standard:

- [DataCite](#), [DublinCore], [DDI](#)

Health study specific data models:

- Clinical trials registries models: [ICTRP](#), [DRKS](#), [ClinicalTrials.Gov](#)
- Epidemiological study models: [MIABIS](#), [Maelstrom](#) data model, minimal requirements model from [ENPADAS](#)
  - Mapped to: [FHIR](#), [SNOMED CT](#), [ECRIN](#)

Metadata models for *nutritional studies and instruments*:

- [ONE](#) (Ontology for Nutritional Epidemiology)
- [ONS](#) (Ontology for Nutritional Studies)
  - Mapped to: [SNOMED CT](#)

A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains two rows of yellow liquid and two rows of red liquid. The background is a light gray laboratory surface.

## Standards for studies in nutritional epidemiology

# NFDI4Health: Use of (meta-)data standards

Recommendations from NFDI4Health on the use of standards in nutritional epidemiology

- Generic metadata standards
  - [DublinCore](#)
  - [DataCite](#)
- Health-study-specific (meta-)data standards
  - [SNOMED CT](#)
  - [HL7® FHIR](#)
  - [OMOP OHDSI](#)
- Standards to describe nutritional studies and dietary assessment instruments
  - [ONE](#)
  - [ONS](#)
  - [NOVA](#)

A 96-well microplate is shown, held by four blue nitrile gloves. The plate contains two rows of yellow liquid and two rows of red liquid. The background is a light gray surface.

## NFDI4Health & FAIR data principles

# NFDI4Health: Link to FAIR data principles

How do the FAIR data principles relate to NFDI4Health?

- **Findability**

- The NFDI4Health assigns a **globally unique and persistent identifier** (DOI) to study documents
- The metadata schema is based on several existing schemas for both generic and health-specific metadata, providing **rich** (i.e. generous and extensive) **metadata**
- The metadata in the German Central Health Study Hub **include the identifier of the data** they describe (e.g. DOI, DRKS ID, ...)
- Metadata are **registered** in the German Central Health Study Hub (= a **searchable resource**)

# NFDI4Health: Link to FAIR data principles

How do the FAIR data principles relate to NFDI4Health?

- **Accessibility**

- The metadata in the Central Health Study Hub are **retrievable** by their identifier **using a standardised communications protocol** (ftp / http(s))
- **The protocol** (http) **is open**, free, and universally implementable
- The protocol (https) allows for an **authentication and authorisation** procedure, where necessary
- The metadata are **accessible** in the German Central Health Study Hub, even when the **data are no longer available**

# NFDI4Health: Link to FAIR data principles

How do the FAIR data principles relate to NFDI4Health?

- **Interoperability**

- The metadata in the German Central Health Study Hub use a **formal, accessible, shared, and broadly applicable language** for knowledge representation (e.g. controlled vocabularies such as MeSH and SNOMED CT)
- The metadata **use vocabularies** that follow FAIR principles, i.e. they are well documented and use unique and persistent identifiers
- The metadata **include qualified references** to other (meta)data, e.g. questionnaires that are part of a study that is registered in the German Central Health Study Hub

# NFDI4Health: Link to FAIR data principles

How do the FAIR data principles relate to NFDI4Health?

- **Reusability**

- The metadata are richly **described with** a plurality of **accurate and relevant attributes** (that describe the context under which data were generated, e.g. (date of) data collection , sampling method, data sources, ... )
- Study documents are **released with** a clear and **accessible data usage license** (= CC-licenses)
- The metadata in the German Central Health Study Hub are **associated with detailed provenance** and state who is responsible for the data /who wants to be acknowledged
- The metadata **meet** domain-relevant **community standards**, e.g. DublinCore, DataCite, DRKS, FHIR, SNOMED CT

# Q&A



Photo by [Jon Tyson](#) on [Unsplash](#)

# Feedback

Thank you for attending the workshop. We would like to ask you to fill in a feedback survey to improve the workshop. In advance, thank you for your help.

Link to survey: *[provide link to survey]*

Note for trainers:

The online survey contains 12 questions on the workshop and was created using a free online tool. You can find the questions on the next slide. The link to the survey was provided to participants on this slide. It may be necessary to send a friendly reminder a week or two after the workshop, asking participants to take part in the survey.



Photo by [Manny Becerra](#)  
on [Unsplash](#)

# Feedback

The online survey contained the following 12 questions:

- The learning objectives of the workshop were clearly stated
- I understand what this workshop was about
- The content of the workshop was easy to follow
- Each topic has been dealt with over a sufficient period of time
- The workshop was a good mix of theory and practice
- I got a good insight into the topic of research data management
- I learned a lot of new and interesting things
- The contents of the workshop are helpful for my scientific work
- All in all, the workshop met my expectations
- How likely is it that you would recommend this workshop to a colleague?
- What did you like the most?
- Please name one thing we could improve



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](mailto:forschungsdaten@zbmed.de)

[www.zbmed.de](http://www.zbmed.de)

**NFDI4Health –  
National Research Data  
Infrastructure for Personal  
Health Data**

[contact@nfdi4health.de](mailto:contact@nfdi4health.de)  
[studyregistration@zbmed.de](mailto:studyregistration@zbmed.de)

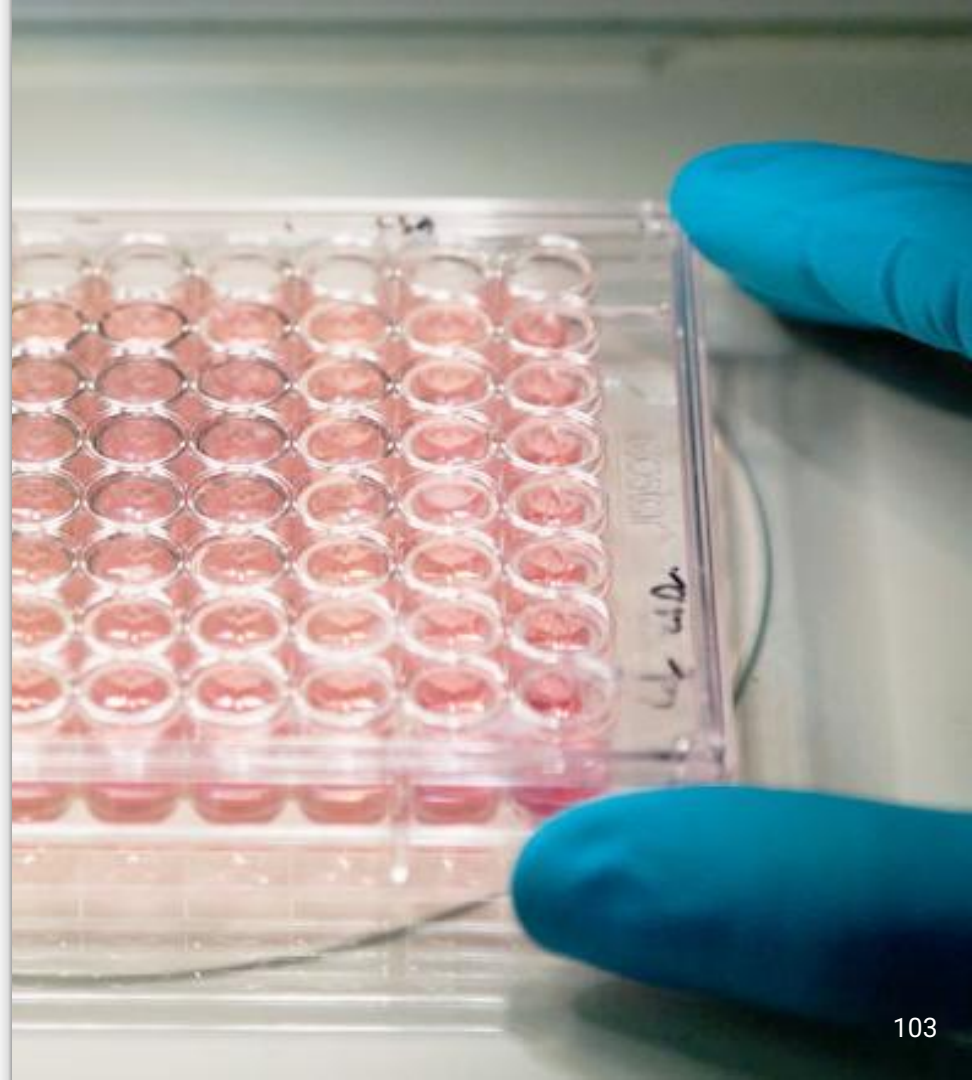
[www.nfdi4health.de](http://www.nfdi4health.de)

**IEL – Nutritional  
Epidemiology  
University of Bonn**

Friedrich-Hirzebruch-Allee 7  
53115 Bonn

[epi@uni-bonn.de](mailto:epi@uni-bonn.de)

<https://www.ernaehrungsepidemiologie.uni-bonn.de/>



# Headline

Content

Note for trainers:

The following two slides can be used as templates to create or add new content to the slidedeck.

An additional slide provides information on the colours and font used throughout the slides.

# Headline

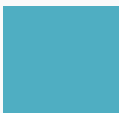
Content

Content

# Graphic colour selection

Font: Roboto

Primary colours:



RGB: 74, 174, 194



RGB: 115, 115, 115



RGB: 0, 176, 80



RGB: 244, 102, 102