WORKSHOP ON RESEARCH DATA

A. Vieten and J. Steinkamp Johannes Gutenberg University (JGU) Mainz
B. Lindstädt, A. Shutsko & J. Vandendorpe, ZB MED - Information Centre for Life Sciences

This work is licensed under a Creative Commons Attribution 2.0 Generic License.
Outline

➔ Introduction
➔ Planning
➔ Data collection: Electronic Lab Notebooks (ELNs)
➔ Data sharing & publishing
➔ Data preservation
➔ Best practice example
➔ Further offer
➔ Q&A
Outline

➔ Introduction
➔ Planning
➔ Data collection: Electronic Lab Notebooks (ELNs)
➔ Data sharing & publishing
➔ Data preservation
➔ Best practice example
➔ Further offer
➔ Q&A
Introduction of the presenters
Team of Experts Research Data at JGU

- **Dr. Anne Vieten** (Department for Research and Technology Transfer) – general RDM aspects, requirements of funders, DMPs, RDMO… -- University Medicine (UM): only for CRC, RTG, FOR…
  Further projects UM please contact: forschung.um@uni-mainz.de
- **Dr. Jörg Steinkamp** (Data Center – technical RDM aspects, GitLab, data archiving…)
- **Karin Eckert** (University library - Metadata, Gutenberg Open Science,…)
- **Esther Reineke** (University library – good research practise, academic integrity)
- **N.N.** (Mainzed – digital methods in humanities and cultural science …)

forschungsdaten@uni-mainz.de
What are we offering?

**Informational service and assistance:**
general RDM aspects, requirements of funders, DMPs incl. RDMO usage, individual RDM-policies, technical RDM aspects, iRODS archive usage, metadata, good scientific practise

**Training:**
RDM-Introduction, DMPs incl. RDMO, GitLab

**Informational events:**
Research Data Management Day at JGU, FORSCHUNGSDATEN@RMU both 1 per year

**RDM-Infrastructures:**
iRODS Archive, Gutenberg Open Science repository, RDMO (DMP-tool), GitLab, Seafile sync and share (cloud storage)
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 and NFDI4Microbiota
- **Aliaksandra Shutsko**
  - Background in Computational Linguistics & Information Science
  - Member of the RDM team
  - Data steward within NFDI4Health
- **Justine Vandendorpe**
  - Background in Organism Biology & Ecology and Computational Evolutionary Biology
  - Member of the RDM team
  - Data steward within NFDI4Microbiota
Research data & Research Data Management (RDM)
A definition of research data

- There is **no consensus** on the definition, 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]
Examples of research data

- Documents (e.g. text, Word), 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

Source: NC STATE University Libraries
Examples of research data in medicine

- 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))
Research data management

A definition of research data management:

‘series of steps and methods that aim to make research data usable over the long term’

Steps:

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

Source: ZB MED
ZB MED’s services

PUBLISSO, the ZB MED publication portal for life sciences, offers dedicated services and tools along the research data life cycle:

- RDMO4Life
- ELN guide
- DOI service
- Repository for Life Science
- Digital long-term preservation
- LIVIVO - The Search portal for Life Sciences
Good scientific practice
Defining good scientific practice

Principles, values and standards of behavior and practice that must be achieved and maintained in the delivery of work activities, the provision of care and personal conduct [Academy for Healthcare Science (AHCS)].

Photo by Scott Graham on Unsplash
Examples of good scientific practice

- Core principles: honesty, respect and accountability
- Documenting results
- Safeguarding and storing primary data
- Observing ethical standards

Source: Bosch, 2010, Guidelines for Safeguarding Good Scientific Practice at the Friedrich Schiller University Jena, O’Grady 2021

Photo by frank mckenna on Unsplash
Examples of scientific misconduct

- Giving false information
- Infringement of intellectual property
- Self-plagiarism
- Compromising research activity of others
- Financial conflicts of interest
- Manipulating authorship

Source: Bosch, 2010, Guidelines for Safeguarding Good Scientific Practice at the Friedrich Schiller University Jena, O’Grady 2021
Ways of securing research integrity

- Establishing harmonize codes of good scientific practice such as:
  - DFG’s Guidelines for Safeguarding Good Research Practice
  - The European Code of Conduct for Research Integrity
- Regulating procedures for handling allegations of research misconduct.

Source: Bosch, 2010, Guidelines for Safeguarding Good Scientific Practice at the Friedrich Schiller University Jena, O’Grady 2021
Implementation at JGU Mainz

- Will be implemented until summer 2022
- Draft is currently being reviewed by the DFG
- More RDM content than before
- **FAIR** principles to be fulfilled!
Policies & guidelines
General guidelines

DFG Guidelines for Safeguarding Good Research Practice. RDM-relevant guidelines:

- 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
Discipline-specific guidelines

- Guidelines 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 guidelines: ZB MED’s Research Data Policy (German only)
Metadata & Metadata standards
Definition of metadata

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

- To make data:
  - Findable, Accessible, Interoperable, Reusable (FAIR)
  - Understandable
  - Citable
- To facilitate the long-term archiving of data
- To make the context for how your data was created, analysed, stored reproducible
- To uphold research integrity
## Types of metadata

<table>
<thead>
<tr>
<th>Descriptive</th>
<th>vs.</th>
<th>Technical</th>
</tr>
</thead>
<tbody>
<tr>
<td>e.g. resource identifier, title, author, date of publication, subject, publisher, description</td>
<td>vs.</td>
<td>e.g. physical database tables, access permissions, data models, backup rule</td>
</tr>
<tr>
<td><strong>Project-specific / dataset-level</strong></td>
<td>vs.</td>
<td><strong>Data-specific</strong></td>
</tr>
<tr>
<td>e.g. project title, description, subject, coverage, creators, publisher, contributors, identifiers</td>
<td>e.g. unique ID, file path, unique project ID, date the file was created</td>
<td></td>
</tr>
<tr>
<td><strong>General</strong></td>
<td>vs.</td>
<td>** Discipline-specific**</td>
</tr>
<tr>
<td>e.g. file name and format, software and hardware used to create the files</td>
<td>e.g. reagent, technical, experimental, analytical, dataset-level</td>
<td></td>
</tr>
</tbody>
</table>
Example of general metadata

**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

<table>
<thead>
<tr>
<th>nr.</th>
<th>Dublin Core element</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Title</td>
</tr>
<tr>
<td>2</td>
<td>Subject</td>
</tr>
<tr>
<td>3</td>
<td>Description</td>
</tr>
<tr>
<td>4</td>
<td>Type</td>
</tr>
<tr>
<td>5</td>
<td>Source</td>
</tr>
<tr>
<td>6</td>
<td>Relation</td>
</tr>
<tr>
<td>7</td>
<td>Coverage</td>
</tr>
<tr>
<td>8</td>
<td>Creator</td>
</tr>
<tr>
<td>9</td>
<td>Publisher</td>
</tr>
<tr>
<td>10</td>
<td>Contributor</td>
</tr>
<tr>
<td>11</td>
<td>Rights</td>
</tr>
<tr>
<td>12</td>
<td>Date</td>
</tr>
<tr>
<td>13</td>
<td>Format</td>
</tr>
<tr>
<td>14</td>
<td>Identifier</td>
</tr>
<tr>
<td>15</td>
<td>Language</td>
</tr>
</tbody>
</table>
## Examples of discipline-specific metadata

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

Source: [Harvard University](http://www.harvard.edu)
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
Technical metadata

<table>
<thead>
<tr>
<th>Metadata element</th>
<th>Metadata value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scanner model</td>
<td>Siemens 3T Prisma</td>
</tr>
<tr>
<td>Head coil</td>
<td>24-channels</td>
</tr>
<tr>
<td>Sequence</td>
<td>T1-weighted MPRAGE</td>
</tr>
<tr>
<td>TR</td>
<td>2300 ms</td>
</tr>
<tr>
<td>TE</td>
<td>2.98 ms</td>
</tr>
<tr>
<td>Flip angle</td>
<td>9°</td>
</tr>
<tr>
<td>Voxel size</td>
<td>1 x 1 x 1 mm³</td>
</tr>
<tr>
<td>FOV</td>
<td>256 x 256 mm²</td>
</tr>
<tr>
<td>Number of slices</td>
<td>176</td>
</tr>
<tr>
<td>Slice thickness</td>
<td>1 mm</td>
</tr>
</tbody>
</table>
Examples of metadata standards in the life sciences

- **To report:**
  - Clinical data: [SNOMED CT](https://www.snomed.org)
  - Diseases and health conditions: [ICD](https://icd.who.int)
  - Data derived by relevant methods in biosciences: [MIBBI](https://www.mibbi.org)

- **To index** journal articles and books in the life sciences: [MeSH](https://www.nlm.nih.gov/mesh/)

- **To exchange:**
  - Clinical and translational research data: [CDISC](https://www.cdisc.org) - [ODM-XML](https://www.cdisc.org/standards/odm/
  - Healthcare information electronically: [HL7 FHIR](https://www.hl7.org/fhir)

- **Formats:**
  - For neutron, x-ray, and muon science: [NeXus](https://nexus.nist.gov)
  - For storing microscopy information: [OME-XML](https://fiji.sc/OME-JAVA)
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.

[Core data set of the MII]
FAIR data principles
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 [Wilkinson et al. 2016]:
  - Findability
  - Accessibility
  - Interoperability
  - Reusability

- **Aims** [Wilkinson et al. 2016]:
  - 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

- **Example of FAIR health data**
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]

\[\text{FAIR} \neq \text{FOIR} \ (O = \text{Open})\]
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]
Outline

- Introduction
- Planning
- Data collection: Electronic Lab Notebooks (ELNs)
- Data sharing & publishing
- Data preservation
- Best practice example
- Further offer
- Q&A
Data Management Plans (DMPs)
A definition of DMPs

Formal and living document to describe the data, their generation and processing during the project, as well as how the data and research results will be archived afterwards to remain available, usable and comprehensible.
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**
Importance of DMPs

- Giving the project’s team an overview about the **data**, their **storage** and **usage**
  → Easing **coordination** and **common handling** of research data
- Supporting **research integrity**
- Prevent **data loss** and **security holes**
- Facilitating **data reuse**
- **Saving** yourself **time** in the future
- Increasing **data citation**
- **Requirement** from funding organizations (sometimes)
Software tool to set up DMPs
Research Data Management Organiser (RDMO) at JGU Mainz

https://rdmo.zdv.uni-mainz.de
Research Data Management Organiser (RDMO) at JGU Mainz

Choose „Create new Project“

Title and describe your project

Choose a catalog and create the project
Research Data Management Organiser (RDMO) at JGU Mainz

Write your DMP by answering questions

Possibility to share your DMP with colleagues

Export options

Possibility to save intermediate versions → several export options (PDF, Microsoft Office, Open Office, LaTeX etc.)
Research Data Management Organiser (RDMO) at JGU Mainz
Outline

➔ Introduction
➔ Planning
➔ **Data collection: Electronic Lab Notebooks (ELNs)**
➔ Data sharing & publishing
➔ Data preservation
➔ Best practice example
➔ Further offer
➔ Q&A
A definition of ELNs

- **Software**
- **Documentation** of experiments and research data
- Replacement of the **paper format** in the context of the digital transformation
- Examples of **features**: protocol templates, collaboration tools, support for electronic signatures and the ability to manage the lab inventory
Benefits of using an ELN

● Boosting **efficiency** of everyday task thanks to, for instance:
  ○ Protocol templates
  ○ Search features
  ○ Access to the ELN anywhere, anytime

● Increased **verifiability** and **traceability** of research findings, preventing:
  ○ Media discontinuities between handwritten and digital entries
  ○ Information loss due to illegibility

● **Networked digital research environment**
  ○ Seamless interfaces to other programs, such as Application Programming Interfaces (APIs)
  ○ Im- and export functions
  ○ Direct links to research data repositories, data services and publishing platforms
The ELN communicates in a networked research environment

Caption: Integration of the ELN RSpace into the research data management infrastructure of the University of Edinburgh (based on Macdonald and Macneil 2015)

Based on: Krause, E.: Elektronische Laborbücher im Forschungsdatenmanagement - Eine neue Aufgabe für Bibliotheken?, ABI Technik 2016; 36(2): 78-87
# Types of Electronic Lab Notebooks

<table>
<thead>
<tr>
<th>Basic systems</th>
<th>Specialized systems</th>
<th>High end systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Text entries</td>
<td>● Subject-specific functions</td>
<td>● Laboratory Information Management System (LIMS)</td>
</tr>
<tr>
<td>● Attachments</td>
<td>● Templates</td>
<td>● Inventory management</td>
</tr>
<tr>
<td>● Annotation</td>
<td>● Freehand drawing</td>
<td>● Workflows</td>
</tr>
<tr>
<td>● Search function</td>
<td>● Basic inventory management</td>
<td>● Link to lab equipment</td>
</tr>
<tr>
<td>● Sharing via the cloud</td>
<td>● Task assignment</td>
<td>● Analysis</td>
</tr>
<tr>
<td>e.g. Evernote</td>
<td>● Rights management</td>
<td>● Data mining</td>
</tr>
<tr>
<td></td>
<td>● Extensions/API</td>
<td>e.g. Limsophy LIMS</td>
</tr>
<tr>
<td></td>
<td>● Audit trail</td>
<td></td>
</tr>
<tr>
<td>e.g. Labfolder</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The FAIR data principles and using an ELN

- **Findable**
  - Assignment of persistent identifiers (e.g. DOI)
  - Extensive search functions of ELNs (e.g. database, full-text and conditional searches)
  - Support for the assignment of metadata / tags (e.g. extraction from documents)

- **Accessible**
  - Starting point for institutional research data management (e.g. connection to repositories or archives)
  - Storage of data in fixed, accessible locations (e.g. not on researchers’ USB sticks / portable hard drives)

- **Interoperable**
  - Use / connection of controlled vocabularies in metadata
  - Export to standard formats

- **Reusable**
  - Description of data (e.g. facilitated metadata capture)
  - Audit trail
  - Documentation of data generation (e.g. logging methods) and of devices used (e.g. devices provide their output directly to the ELN)
Using an ELN and good scientific practices

- Audit trail
- Version control
- Authentication functions
- No deletion of data possible
- Searchability of entries
- “Freezing” of work statuses
- Marking of entries

Audit Trail, inspired by Johner, C. (2016): „Was sollte ein Audit-Trail enthalten?“
Electronic Lab Notebook (ELN) guide

**ELN guide**

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

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 TU Darmstadt (library)).
- **ELN filter** (in German only): step towards the ELN finder.

<table>
<thead>
<tr>
<th>Name</th>
<th>Land</th>
<th>Referenzen</th>
<th>Preismodell (akademische Nutzung)</th>
<th>Weitere Informationen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arx-span</td>
<td>USA</td>
<td>Unbekannt</td>
<td>Unbekannt</td>
<td>Unbekannt</td>
</tr>
<tr>
<td>Bench-lining</td>
<td>USA</td>
<td>Unbekannt</td>
<td>✅ Kostenlose akademische Version für Personen, Labore und Lehre mit eingeschränktem Funktionsumfang</td>
<td>Unbekannt</td>
</tr>
</tbody>
</table>
Examples of ELNs in molecular biology

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

➔ Introduction
➔ Planning
➔ Data collection: Electronic Lab Notebooks (ELNs)
➔ **Data sharing & publishing**
➔ Data preservation
➔ Best practice example
➔ Further offer
➔ Q&A
Referencing research data
Digital Object Identifier (DOI)

DOI: ‘system for providing persistent and actionable identification and resolution to resource objects in a digital environment’ [Carpenter 2012]

Benefits:
- Providing a reference to the object itself (i.e. the reference is still valid even if the URL changes).
- Making research data accessible and citable over the long term.
Building a culture of data citation
DOI service of UB Mainz

DOI for articles and textual datasets* published in Gutenberg Open Science repository

*soon
Sharing research data: collaboration tools
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** and **Git Large File Storage** providing file managing/versioning systems without checking the file contents into git.
Local services of JGU Mainz

Seafile RLP:

- Sync and Share (Cloud service)
- 100 GB per JGU scientist
- Enables data exchange
- Collaborative Work on documents
- via Only Office

https://seafile.rlp.net
Local services of JGU Mainz

Versioning tool GitLab:

- code, text, vector graphics
- saves and documents changes to files and their source code
- Braching
- Issue tracking
- ...

Additional feature; Mattermost for easy and quick communication

https://gitlab.rlp.net
Publishing research data
Data repositories

Research data can be published as an independent information object in a data repository AND:

- As a data supplement in an enhanced publication
- Documented in a data paper published in a data journal

Data repositories include:

- Interdisciplinary repositories such as Figshare and Zenodo
- Institutional repositories: see example in the next slides
- Discipline-specific repositories: see examples in the next slides
Local repositories

- Institutional Open Access repository
- Pre and post prints, PhD thesis
- Textual research data*
- Searchable by meta data
- Self archiving
- DOI

*soon
https://openscience.ub.uni-mainz.de/?locale=en
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
  - ...
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**
Examples of repository finders

DataCite’s registry of research data repositories (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 conditions.

Last updated: 12/21/2018

<table>
<thead>
<tr>
<th>Name</th>
<th>Subject area focus in the life sciences</th>
<th>Further subject area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 Functional Connectomes Project</td>
<td>Neurosciences</td>
<td></td>
</tr>
<tr>
<td>AceView</td>
<td>Biology</td>
<td></td>
</tr>
</tbody>
</table>
Privacy issues
What are the challenges?

- Obtaining (broad) **informed consent** from participants
  - to respect their right to **health data privacy**
  - because **NOT sharing** clinical trial data comes with drawbacks (e.g. duplication of trials, bias in the body of evidence, slowing scientific progress)
- Properly **pseudo-/anonymising** data
- Finding **alternative approaches** to data pseudo-/anonymisation, which do not satisfy the General Data Protection Regulation (GDPR) [Rocher et al. 2019]
Where can solutions be found?

- **Informed consent**: template text for patient consent forms

- **Pseudo-/anonymisation** tools
  - Amnesia
  - Data privacy tool

- Alternative to pseudo-/anonymisation - **distributed approach**: 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]
  - Personal Health Train (PHT) Approach [Beyan et al. 2020]
  - DataSHIELD [Wilson et al. 2017]
Example of infrastructure for distributed data analysis

**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

➔ Introduction
➔ Planning
➔ Data collection: Electronic Lab Notebooks (ELNs)
➔ Data sharing & publishing
➔ Data preservation
➔ Best practice example
➔ Further offer
➔ Q&A
Storage
## What is storage?

### Difference between backup, archive and publication

<table>
<thead>
<tr>
<th></th>
<th>Backup</th>
<th>Archive</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Access</strong></td>
<td>Possibility of co-working solutions</td>
<td>For the data producer only</td>
<td>Access to others</td>
</tr>
<tr>
<td><strong>Reuse</strong></td>
<td>• To keep data still being worked on safe</td>
<td>Partial reuse</td>
<td>Possibility of reuse</td>
</tr>
<tr>
<td></td>
<td>• To restore the original only</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Persistence</strong></td>
<td>~ 1 year</td>
<td>~ 10 years</td>
<td>Depends on the publishing institution</td>
</tr>
<tr>
<td><strong>Versions</strong></td>
<td>Usually several data versions</td>
<td>Final version, static records</td>
<td>Final version</td>
</tr>
<tr>
<td><strong>Formats</strong></td>
<td>Working formats</td>
<td>Preservation formats</td>
<td>Depends on the publishing institution</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td>Often stored locally</td>
<td></td>
<td>Stored at official institutions</td>
</tr>
</tbody>
</table>

Source: Cozatl et al. 2021
Which formats and media should be used?

Format recommendation

Save files in the original software format and in an additional recommended file format, i.e. a format that is:

- **Open** (not proprietary)
- **Machine-readable**
- **Well-documented**
- **In widespread use**

(e.g. CSV, XML, TXT, PDF/A)

Media: the lifespan of storage data

Source: Crashplan
Benefits of long-term preservation

To researchers

- To make data:
  - Findable, Accessible,
  - Interoperable, Reusable (FAIR)
  - Readable
- Requirement from funding organizations (sometimes)

To societal challenges

- Reproducibility
- Integrity & validity
- Financial implications
- Global responses (e.g. CoViD-19)
- Reduce animal testing
- Cultural & political heritage
- Economic benefits

Source: Venkataraman 2021, Rathmann et al. 2021
What are valuable data and what are not?

Five steps to decide what data to keep:

1. Identify **reuse purposes** that the data could fulfil
2. Identify data that must be kept considering **funder requirements** and **legal** or **policy compliance risks**
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

Source: DCC, Rathmann et al. 2021
iRODS Archive JGU

- Long term archive
- 2 copies at different locations (tape and SSD/HDD)
- Encrypted storage
- Granular access control (permissions)
- Basic metadata
- Command line Linux tool and WebUI*
- 10+ years
- Publication possibility
- Works like virtual file system
Outline

➔ Introduction
➔ Planning
➔ Data collection: Electronic Lab Notebooks (ELNs)
➔ Data sharing & publishing
➔ Data preservation
➔ Best practice example
➔ Further offer
➔ Q&A
National Research Data Infrastructure (NFDI)
The National Research Data Infrastructure (NFDI)

https://www.youtube.com/watch?v=uJ01g9m8uE4&t=78s&ab_channel=DFGbewegt
Natural & Life-Science related NFDI consortia
Natural & Life-Science related NFDI consortia

- Chemistry Consortium in the NFDI (NFDI4Chem)
- German Human Genome-Phenome Archive (GHGA)
- NFDI Consortium Earth System Science (NFDI4Earth)
- NFDI Consortium of Plant Research (DataPLANT)
- NFDI for Biodiversity, Ecology & Environmental Data (NFDI4Biodiversity)
- NFDI for Catalysis-Related Sciences (NFDI4Cat)
- NFDI for Microbiota Research (NFDI4Microbiota)
- NFDI for Personal Health Data (NFDI4Health)
Involvement of JGU Mainz in NFDI consortia

JGU is Co-Applicant in:
- NFDI4Chem
- PUNCH4NFDI
- NFDIxCS (3rd round)

JGU is Participant in:
- NFDI4Culture
- NFDI4Health
- TEXT+
- NFDI4Objects (3rd round)
- NFDI4Memory (3rd round)
Further offer: content

- **Topics**: same as today, but covered in more details
  - Topics covered by **ZB MED**:
    - Discipline-specific content, e.g. ELNs, metadata
    - Best practice examples, such as NFDI4Health and NFDI4Microbiota
  - Topics covered by **JGU Mainz**:
    - Local services and infrastructure
    - Role within NFDI4Chem
- **Target audience**: subject specific, e.g. for (bio-)medical researchers working in the lab, researchers running clinical trials/studies
Further offer: format

- **6-hour** workshop, spread over two or four days
- **Interactive activities**, such as:
  - Polls
  - Exercises
  - Discussions
- **Cost of ZB MED**: ~ 900 Euro (preparation, conducting, follow-up)
Outline

- Introduction
- Planning
- Data collection: Electronic Lab Notebooks (ELNs)
- Data sharing & publishing
- Data preservation
- Best practice example
- Further offer
- Q&A
Q&A

Photo by [Jon Tyson](https://unsplash.com) on [Unsplash](https://unsplash.com)
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

Johannes Gutenberg University Mainz
Saarstr. 21
55122 Mainz
forschungsdaten@uni-mainz.de
https://www.forschungsdaten.uni-mainz.de/