NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

National Centre
for Information Support of Research,
Development, and Innovation

Dataset preparation and publishing

Zenodo and other repositories

Lecturer: Jan Vališ

22. 04. 2026

The project National Centre for Information Support of Research, Development and Innovation with the identification code MS2101 is implemented with the support of the Ministry of Education, Youth and Sports.

How about you?

- What is your academic affiliation?
- What is your level of study?
- What is your main research field?

Learning objectives

Knowledge:

- Differentiate between data that can be published and data that cannot
- Know when to use anonymization/informed consent/controlled access
- Identify key desirable characteristics of a repository

Skills

- Structure and document a dataset for publication
- Select suitable file formats balancing interoperability and reusability
- Deposit the dataset into a repository, e.g. Zenodo

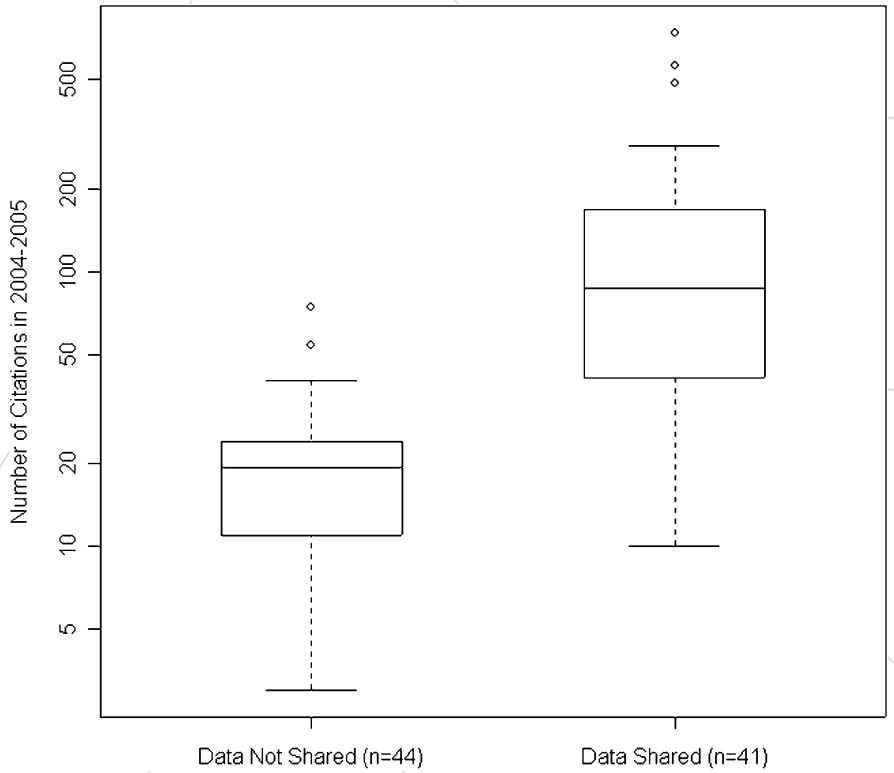
Attitudes:

- Cultivate the habit of planning data publication ahead of time
- Prepare datasets empathetically with the end user in mind

Based on: Petersen, B. *et al.* (2025) Learning Objectives Matrix on the Topic of Research Data Management (RDM) (v 3). Zenodo. 10.5281/zenodo.15846806

Why publish research data

- It may be mandatory (institutional, publisher, or funder policy)
- Enable data reuse → increased efficiency
- Increased visibility → collaboration?
- Increase the impact of research
- Increased transparency
- Support research integrity



2004–2005 citation counts of 85 trials by data availability.
Heather A. et al. 2007. PLOS ONE. License: CC BY 4.0
10.1371/journal.pone.0000308.

What (not) to publish

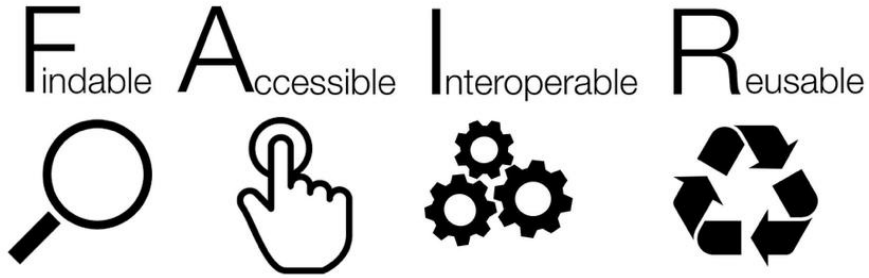
- Personal data
- Sensitive data
- Protected by legitimate interest:
  - Intellectual Property
  - Commercial interests

What to do

- Anonymization?
- Informed consent?
- Embargo?
- Controlled access?

- We can still usually share at least **metadata.**

FAIR principles



Source: SangyaPundir, FAIR data principles, CC BY-SA 4.0

Findable

- Metadata
- Persistent Identifiers (DOI, ORCiD, ROR, IGSN…)
- Registration and indexing in searchable repository

Accessible

- Free and open metadata
- Metadata available even when data are not available

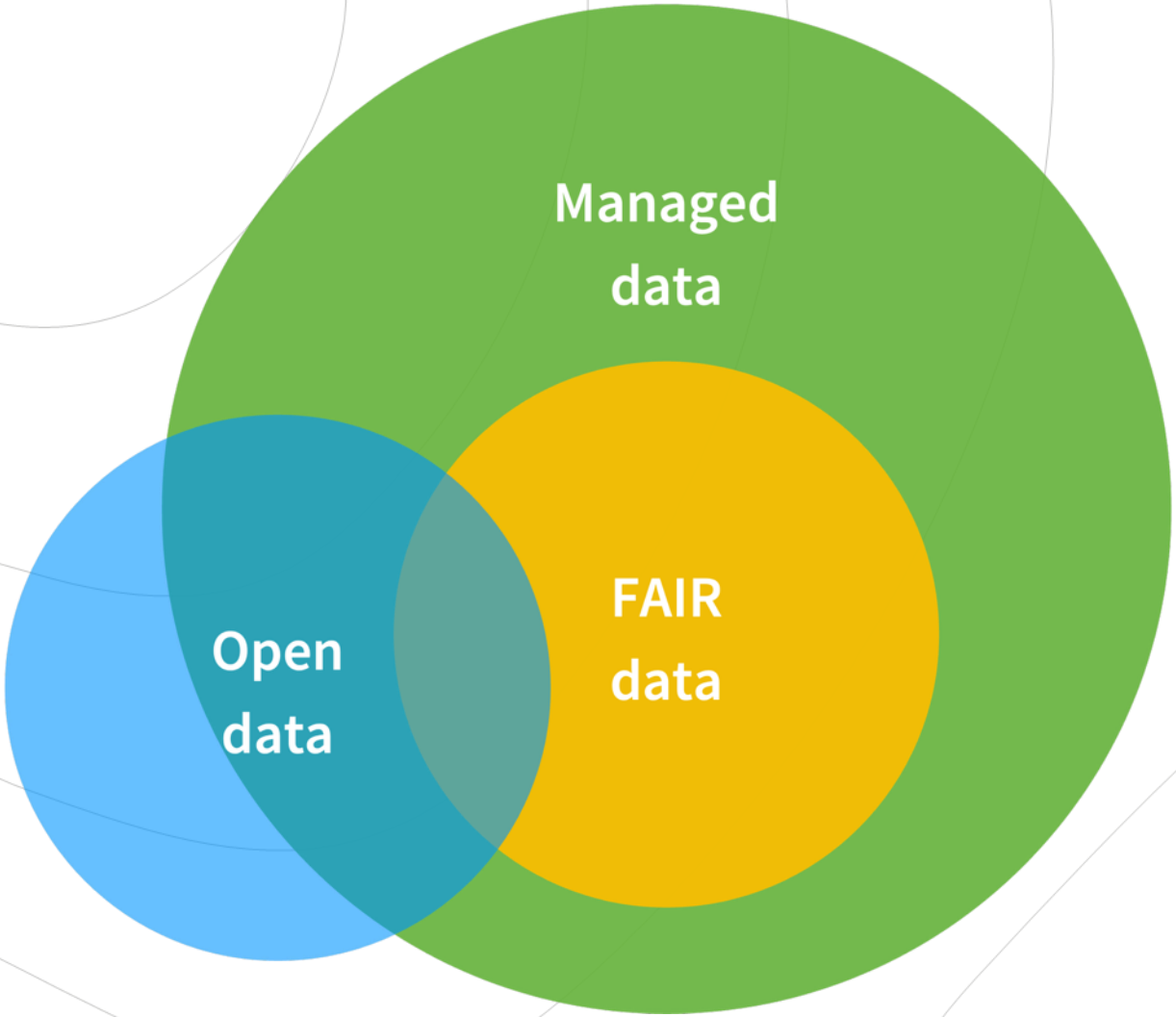
Interoperable

- Widely used language
- Preferred formats
- Vocabularies and ontologies

Reusable

- Rich description (Read Me File)
- License
- Field/Community standards

Open vs. FAIR Data



Source: Valis, Jan (2024). FAIR vs. Open Data (Venn diagram). doi.org/10.6084/m9.figshare.25648548.v1; License: CC BY

Not all publications are created equal

Dataset

- Form: collection of files
- Goal: support paper & reuse
- Where: repository, standalone
- FAIR-compliant
  - dedicated PID & metadata
  - re-use
  - clear licensing
  - citable (incl. RIV output)

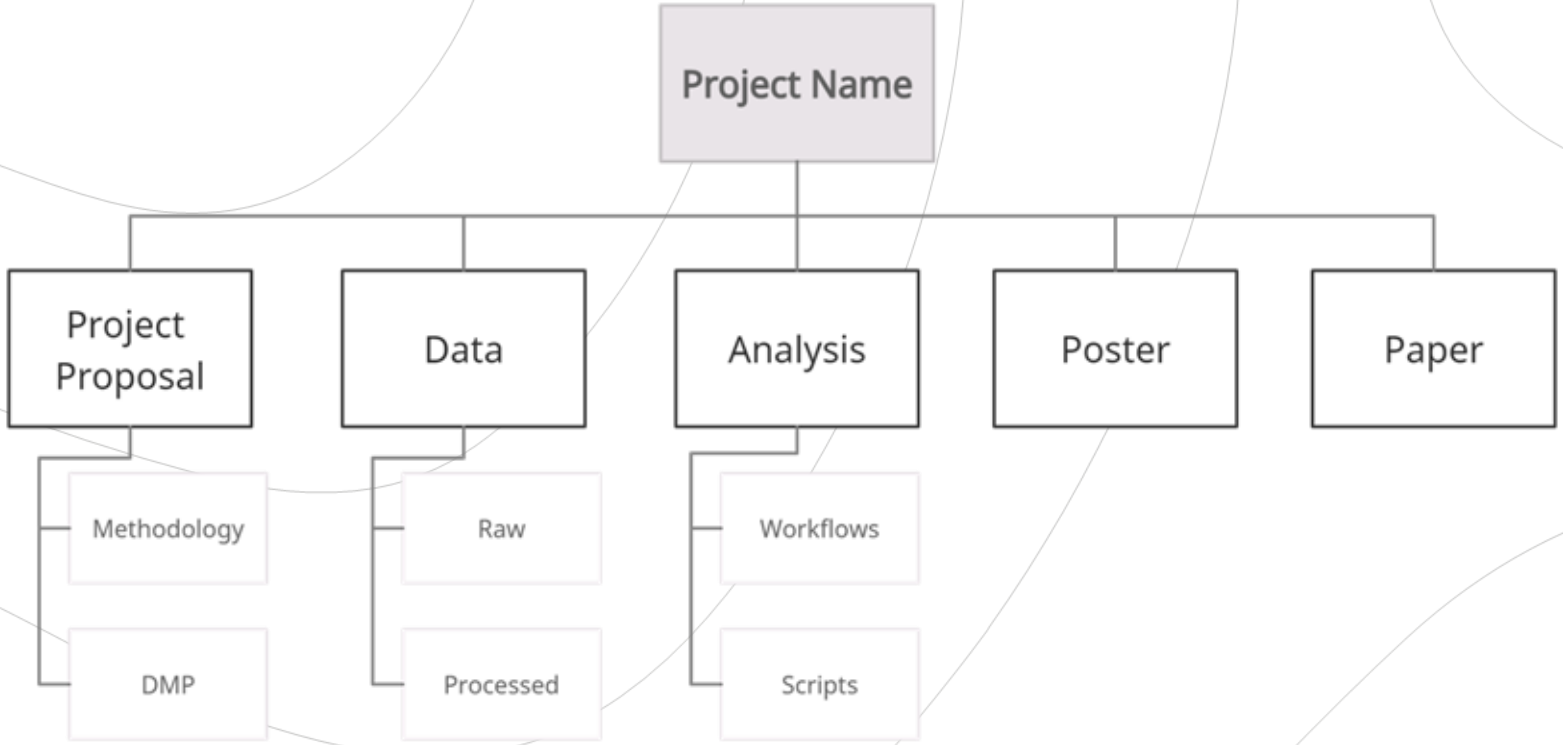
Supplementary Information

- Form: Visualizations, tables etc., (typically in one PDF)
- Goal: support a paper
- Where: publisher, alongside paper
- Can’t be:
  - found separately of the paper
  - re-used
  - cited

Dataset preparation

Organization

- **Structure**



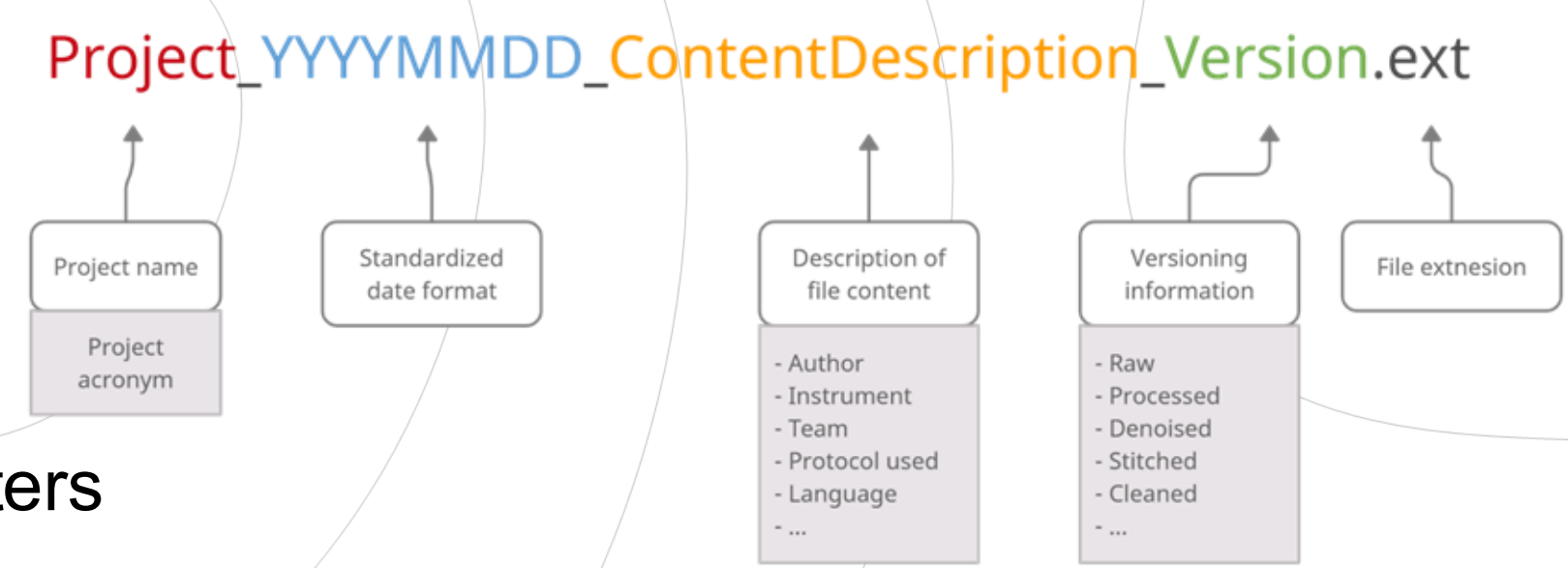
Source: https://biblio.uottawa.ca/en/services/faculty/research-data-management/file-naming-and-organization-data, CC BY

Organization

- Structure
- **Max. 4 levels of folders**
- **Max. 10 subfolders to a folder?**

Source: 10.5281/zenodo.17273968, CC BY

File naming



- Consistent
- No diacritics
- No special characters
- Use
  - hyphens (-) and underscores (_) for separation
  - abbreviations + explanation in a codebook or ReadMe
- Include
  - metadata (language, author, instrument, parameters…)
  - transformation stage (raw, processed, analyzed…)

Edited from: biblio.uottawa.ca/en/services/faculty/research-data-management/file-naming-and-organization-data, CC BY

Documentation

- ReadMe file
- Codebooks
- Guides
- Manuals

Documentation

000_ReadMeFile.txt

```
------------------------------------------------------------------------------------------------------------------------
Raman spectra of rat thymus tissue
------------------------------------------------------------------------------------------------------------------------

* ReadMe version: 1.1 (2026-03-12)
* Dataset version: 1.0 (2025-12-04)
* Dataset DOI: 10.5281/zenodo.18989676

------------------------------------------------------------------------
CONTACT
------------------------------------------------------------------------
* Daniela Janstova
* janstovd@vshct.cz
* ORCID: 0000-0002-0300-3878
* Dept. of . Faculty of Chemical Engineering, University of chemistry and Technology, Prague
* Technicka 5, 166 28, Prague 6, Czech Republic

_____Creators_____
* Matyas Garnol (0009-0006-8397-1364), Dept. of Analytical Chemistry. Faculty of Chemical Engineering, University of chemistry and
Technology, Prague
* Daniela Janstova (0000-0002-0300-3878), Dept. of Mathematics, Cybernetics and Informatics, Faculty of Chemical Engineering, University
of chemistry and Technology, Prague
* Jan Valis (0000-0003-0349-6868), Dept. of Analytical Chemistry. Faculty of Chemical Engineering, University of chemistry and
Technology, Prague
* Marcela Dendisova (0000-0002-5895-243X), Dept. of Physical Chemistry. Faculty of Chemical Engineering, University of chemistry and
Technology, Prague
* Jan Mares (0000-0003-4693-2519), Dept. of Mathematics, Cybernetics and Informatics, Faculty of Chemical Engineering, University of
chemistry and Technology, Prague

_____Contributors_____
* Veronika Spisska (0000-0003-2339-4906), Dept. of Physiology, Faculty of Science, Charles University

------------------------------------------------------------------------
DATA AVAILABILITY AND ACCESS INSTRUCTIONS
------------------------------------------------------------------------
* The dataset is openly accessible under the DOI listed above.

------------------------------------------------------------------------
LICENSE
------------------------------------------------------------------------
_____ReadMe file license_____
* ReadMe by Jan Valis is licensed under CC BY 4.0
* Licence information: https://creativecommons.org/licenses/by/4.0/

_____Dataset license_____
* Dataset Raman spectra of rat thymus tissue by Matyas Garnol, Daniela Janstova, Jan Valis, Marcela Dendisova and Jan Mares is licensed
under CC BY 4.0
* Licence information: https://creativecommons.org/licenses/by/4.0/

------------------------------------------------------------------------
DESCRIPTION AND METHODOLOGY
------------------------------------------------------------------------

_____About the dataset_____
Raman spectroscopy offers a powerful, label-free approach for molecular characterization of biological tissues, with enormous potential
for advancing tissue diagnostics. Here, we present a controlled dataset of rat thymus Raman spectra measured under near-native
conditions (non-fixed, unstained cryosections). Both normal healthy and cancerous tissue samples were measured. This labeled dataset can
be used to train and evaluate pre-processing methods as well as machine-learning methods for classification of Raman spectra of
biological samples.

_____Ethics_____
All experimental procedures were conducted in strict accordance with the Guide for the Care and Use of Laboratory Animals of the
National Institutes of Health and complied with the Animal Protection Law of the Czech Republic. The experimental protocol was approved
by the Ministry of Education, Youth and Sports of the Czech Republic (approval number: MSMT-31592/2019-4, December 2, 2019).

_____Sample provenance and preparation_____
Thymus tissues were obtained from three Wistar rats (Rattus norvegicus domestica, Velaz Ltd., Prague, Czech Republic) as part of the
experimental work conducted at the Faculty of Science, Charles University (Prague, Czech
Republic). Animal euthanasia was followed by rapid dissection and isolation of the thymus. Immediately after excision, tissues were
snap-frozen on dry ice (-78 °C) to prevent enzymatic degradation and preserve native biochemical composition. After complete freezing,
samples were transferred into sterile microcentrifuge tubes and stored at -80 °C until further processing.

Frozen thymus tissues were embedded in Tissue-Tek OCT Cryomount compound (Sakura Finetek, Japan) and mounted onto a metal specimen
holder. Cryosectioning was performed using a Leica CM1950 cryostat (Leica Microsystems, Germany) at approximately -20 °C. 60 µm thick
sections provided the highest signal-to-noise ration without compromising section integrity. Prepared cryosections were immediately
stored at -80 °C until Raman measurements. Prior to spectral acquisition, sections were allowed to equilibrate to room temperature under
ambient laboratory conditions.

_____Methods of data collection_____
Raman spectra of thymus tissue were acquired using an inVia Raman microscope (Renishaw, UK) equipped with a 785 nm diode laser (max
```

Metadata

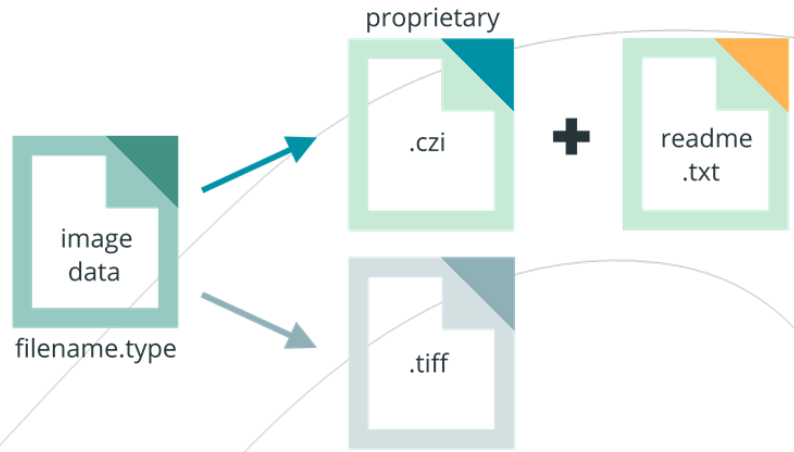
- Dataset level – generic (e.g., DataCite) & domain-specific
- File level
  - Electronic Laboratory Notebook (ELN)
    - Keeps provenance of sample and measurements
    - Keeps metadata structured
    - Helps to keep data FAIR and lowers the threshold for publication
    - Many options (LabFolder, KADI4Mat, eLabFTW, RSpace, CheMotion …)
    - Find your ELN: ELN Finder, Research.com
  - If no ELN, then:
    - Thorough ReadMe
    - Some tabular description of files, samples, conditions etc.

File formats

- Preferred (CSV) vs. popular (XLSX)
- Open (TIFF) vs. proprietary (PSD)
- Lossless (TIFF) vs. lossy compression (JPEG)
- If proprietary necessary, include info in ReadMe on how to open/use the file (SW, workflow etc.)

Preferred format

- archiving-friendly
- open
- well-documented
- human & machine readable
- lossless compression
- not dependent on a specific SW



source: howtofair.dk/how-to-fair/file-formats/

File formats

Type	Preferred	Non-preferred
Text documents	PDF/A ODT	DOC, DOCX RTF
Tables	CSV ODS	XLS, XLSX PDF/A
Raster image	TIFF, JPEG 2000	
Vector image	SVG	EPS, AI,
Audio	FLAC, OPUS, BWF, MXF	WAV, MP3, AAC
Video	MKV, MXF	AVI, MOV, M4A, MPEG

source: dans.knaw.nl/en/file-formats/

Data Processing & Analysis

- Algorithmic workflows prefererd:
  - Reproducibility
  - Efficiency
  - Saves space – storing only *raw data* + *code* (vs. *raw* + *processed* + *analyzed*)
- Scripts can be also use to help with dataset preparation, e.g.:
  - Rename files
  - Split files by file-type
  - Print out tree structure
- Open languages (e.g., Python, R) preferred to proprietary

Choosing a repository

Types I

- **Institutional** (e.g., CERN, ESO, ASEP)
- **General purpose** (e.g., Zenodo, National Repository)
- **Domain-specific/Disciplinary**
  - Instrument/analysis/data type (e.g., MS, FT-IR, XRD)
  - Topic (e.g., drug discovery)

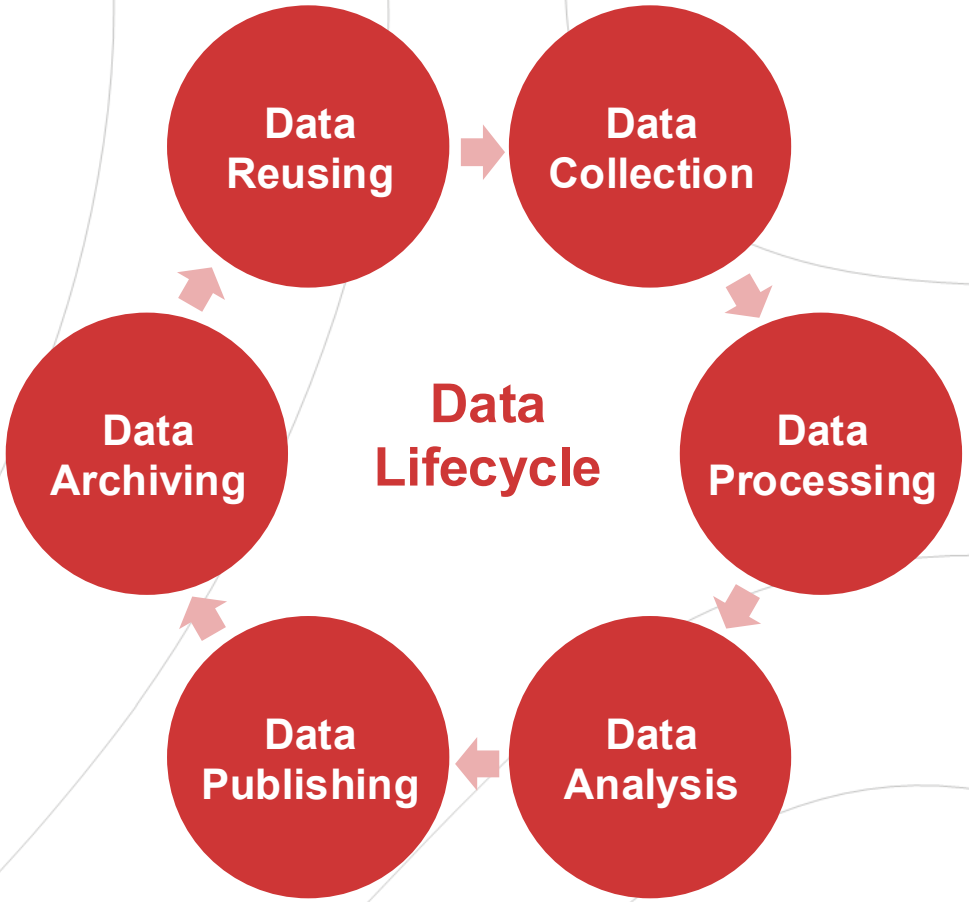
Types II

Position in lifecycle

- Raw primary data repository
- Processed data repository
- Repository for supporting data

By curation

- Open
- Curated
- Peer-reviewed



What to Look For in a Repository

- Motivation? (community, publication or grant compliance?)
- Metadata & ontologies?
- File format compatibility?
- File size/number limitation?
- Access control/embargo?
- Licenses?
- Reputation, persistence, sustainability?
- Price?

Ideal characteristics

- Relevant metadata and ontologies
- PIDs (DOI, ORCiD, ROR, IGSN …)
- Supports versioning
- FAIR
- Desirable licenses
- Grant dedications
- Stable funding and reputation
- Free to contribute to

Ideal characteristics

- **Relevant metadata and ontologies**
- PIDs (DOI, ORCiD, ROR, IGSN …)
- Supports versioning
- FAIR
- Desirable licenses
- Grant dedications
- Stable funding and reputation
- Free to contribute to

Metadata Models

Universal

- Dublin Core
- DataCite
- …

Field-specific

- Open Geospatial Metadata Standard
- Chemical Research Object Framework

Ideal characteristics

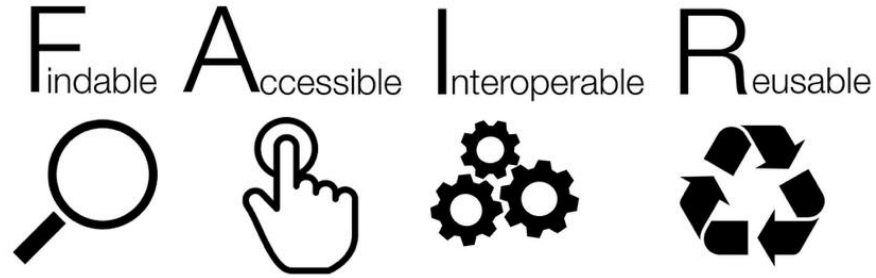
- Relevant metadata and ontologies
- **PIDs (DOI, ORCiD, ROR, IGSN …)**
- Supports versioning
- FAIR
- Desirable licenses
- Grant dedications
- Stable funding and reputation
- Free to contribute to

Ideal characteristics

- Relevant metadata and ontologies
- PIDs (DOI, ORCiD, ROR, IGSN …)
- **Supports versioning**
- FAIR
- Desirable licenses
- Grant dedications
- Stable funding and reputation
- Free to contribute to

Ideal characteristics

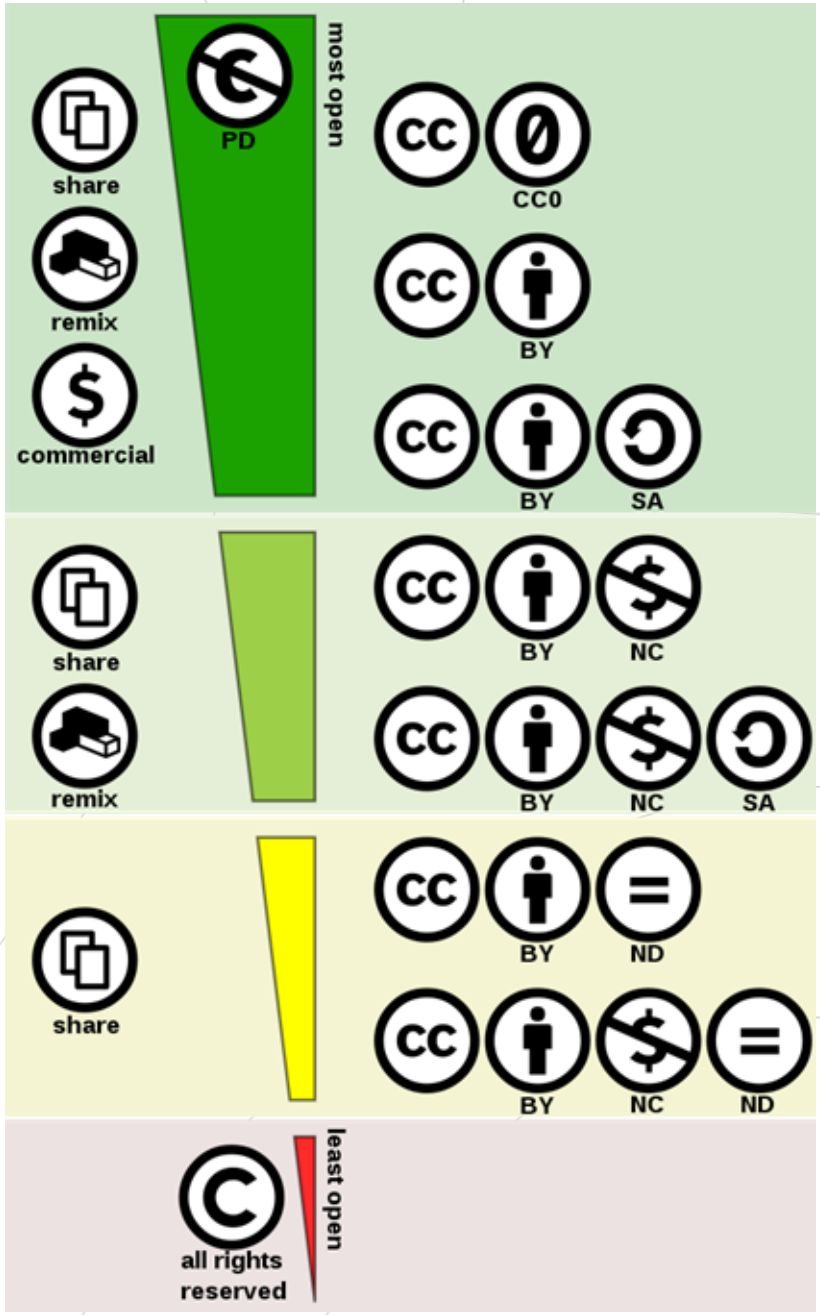
- Relevant metadata and ontologies
- PIDs (DOI, ORCiD, ROR, IGSN …)
- Supports versioning
- **FAIR**
- Desirable licenses
- Grant dedications
- Stable funding and reputation
- Free to contribute to



Source: SangyaPundir, FAIR data principles, CC BY-SA 4.0

Ideal characteristics

- Relevant metadata and ontologies
- PIDs (DOI, ORCiD, ROR, IGSN …)
- Supports versioning
- FAIR
- **Desirable licenses** (CC, SW-specific)
- Grant dedications
- Stable funding and reputation
- Free to contribute to



Source: creativecommons.org/

Ideal characteristics

- Relevant metadata and ontologies
- PIDs (DOI, ORCiD, ROR, IGSN …)
- Supports versioning
- FAIR
- Desirable licenses
- **Grant dedications**
- Stable funding and reputation
- Free to contribute to

Source: 10.5281/zenodo.17273968, CC BY

Ideal characteristics

- Relevant metadata and ontologies
- PIDs (DOI, ORCiD, ROR, IGSN …)
- Supports versioning
- FAIR
- Desirable licenses
- Grant dedications
- **Stable funding and reputation**
- Free to contribute to

Source: Zenodo

Ideal characteristics

- Relevant metadata and ontologies
- PIDs (DOI, ORCiD, ROR, IGSN …)
- Supports versioning
- FAIR
- Desirable licenses
- Grant dedications
- Stable funding and reputation
- **Free to contribute to**

Red flags and pitfalls

- HTTP, expired certificates
- No clear sustainable business plan/funding
- Low number of records
- Publisher-owned
- No grant dedication
- No PIDs
- Irrelevant licenses

Where to look for repositories?

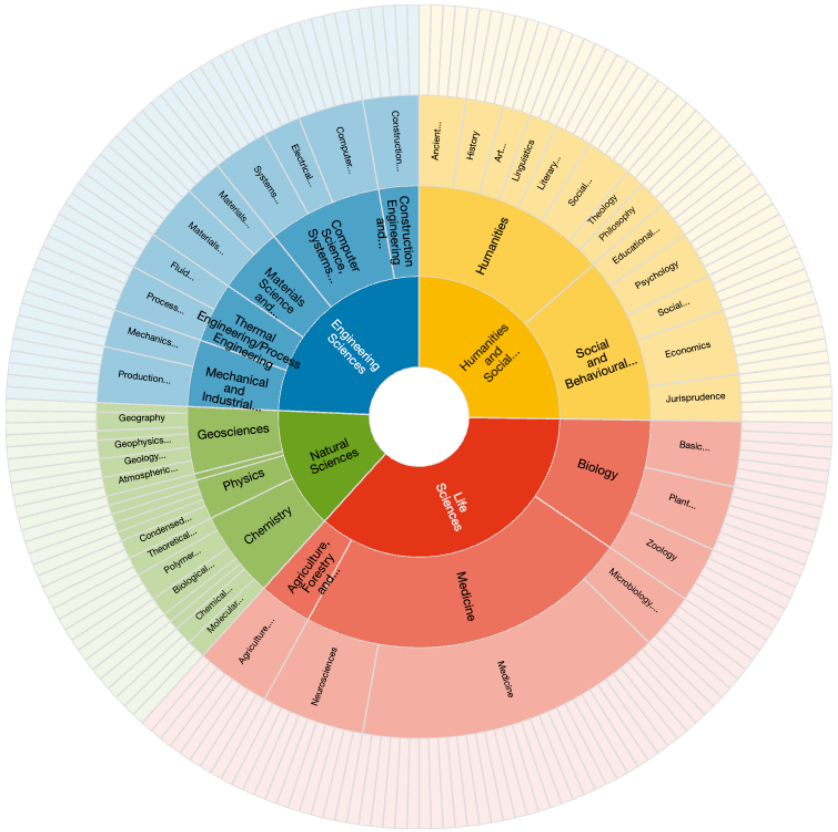
- All-purpose aggregators (e.g., re3data, FAIRsharing)
- Societies and grant agencies (e.g., NFDI4Chem, ERC)
- Publishers?

re3data

Browse by subject

Graphical | Text

click to zoom into subjects or to select a bottommost subject in the hierachy as filter for the re3data search page
shift + click on a top subject to select it as filter



re3data.org — Search | Browse | Suggest | Resources | Contact

Filter (Reset all)

Subjects | Content Types | Countries | AID systems | API | Certificates | Citation References | Data access
closed (17)
embargoed (53)
open (89)
restricted (50)
Data access restrictions | Database access
open (89)
Database licenses | Data licenses
Apache License 2.0 (13)
BSD (11)
CC (89)
CC0 (34)
Copyrights (15)
ODC (16)
OGL (1)
Public Domain (4)
other (30)
Data upload | Data upload restrictions | Enhanced publication | Institution responsibility type | Institution type
commercial (3)
non-profit (89)
Keywords | Metadata standards
ABCD (Access to Biological Collection Data) (1)
CF (Climate and Forecast) Metadata Conventions (2)
CMDI (Component Metadata Infrastructure) (1)
DCAT (Data Catalog Vocabulary) (1)
DCAT-AP (1)
DDI (Data Documentation Initiative) (41)
Darwin Core (1)
DataCite Metadata Schema (89)
Dublin Core (70)
EngMeta (1)
ISA-Tab (2)
ISO 19115 (1)
International Virtual Observatory Alliance Technical Specifications (1)
JSON-LD (JavaScript Object Notation for Linked Data) (12)
MODS (Metadata Object Description Schema) (1)
OAI-ORE (Open Archives Initiative Object Reuse and Exchange) (32)
OECD Dataset Metadata (1)
OpenAIRE Guidelines (12)
RDF Data Cube Vocabulary (1)
Repository-Developed Metadata Schemas (6)
schema.org (13)
Metrics | PID systems
ARK (2)
DOI (89)
PURL (1)
URN (3)
hdl (14)
other (4)
Profiles | Provider types | Quality management
yes (89)
Repository languages
Catalan; Valencian (2)
German (11)
English (89)
French (16)
Galician (1)
Italian (1)
Korean (1)
Polish (6)
Portuguese (5)
Spanish; Castilian (10)
Chinese (1)
Software | Syndications | Repository types | Versioning
yes (89)

Search... | Search | Search help

← Previous | 1 | 2 | 3 | 4 | Next → | Sort by

Found 89 result(s)

Illinois Data Bank
Subject(s): Humanities and Social Sciences | Life Sciences | Natural Sciences | Engineering Sciences
Repository type(s): institutional
Provider type(s): dataProvider
Country: United States
The Illinois Data Bank is a public access data repository that collects, disseminates, and provides persistent and reliable access to the research data of faculty, staff, and students at the University of Illinois at Urbana-Champaign. Faculty, staff, graduate students can deposit their research data directly into the Illinois Data Bank and receive a DOI for citation purposes.

Research Data at Essex
University of Essex Research Data Repository
Subject(s): Humanities and Social Sciences | Life Sciences | Natural Sciences | Engineering Sciences
Repository type(s): institutional
Provider type(s): dataProvider
Country: United Kingdom
The Research Data Repository is the University of Essex's online data repository where research data resulting from research taking place within the university can be deposited, published and made accessible to the research community.

RWTH Publications Research Data
RWTH Publications Forschungsdaten
Subject(s): Humanities and Social Sciences | Life Sciences | Natural Sciences | Engineering Sciences
Repository type(s): institutional
Provider type(s): dataProvider | serviceProvider
Country: Germany
RWTH Publications Research Data offers all RWTH Aachen University affiliates the organizational and technical means to electronically document and publish research data at this institutional repository. Certainly, researchers are encouraged to prefer a subject specific repository whenever appropriate and available. RWTH Aachen University is the largest technical university in Germany and one of nine 'German Universities of Excellence' (elite university). The University library Aachen operates the repository as a member of the join community.

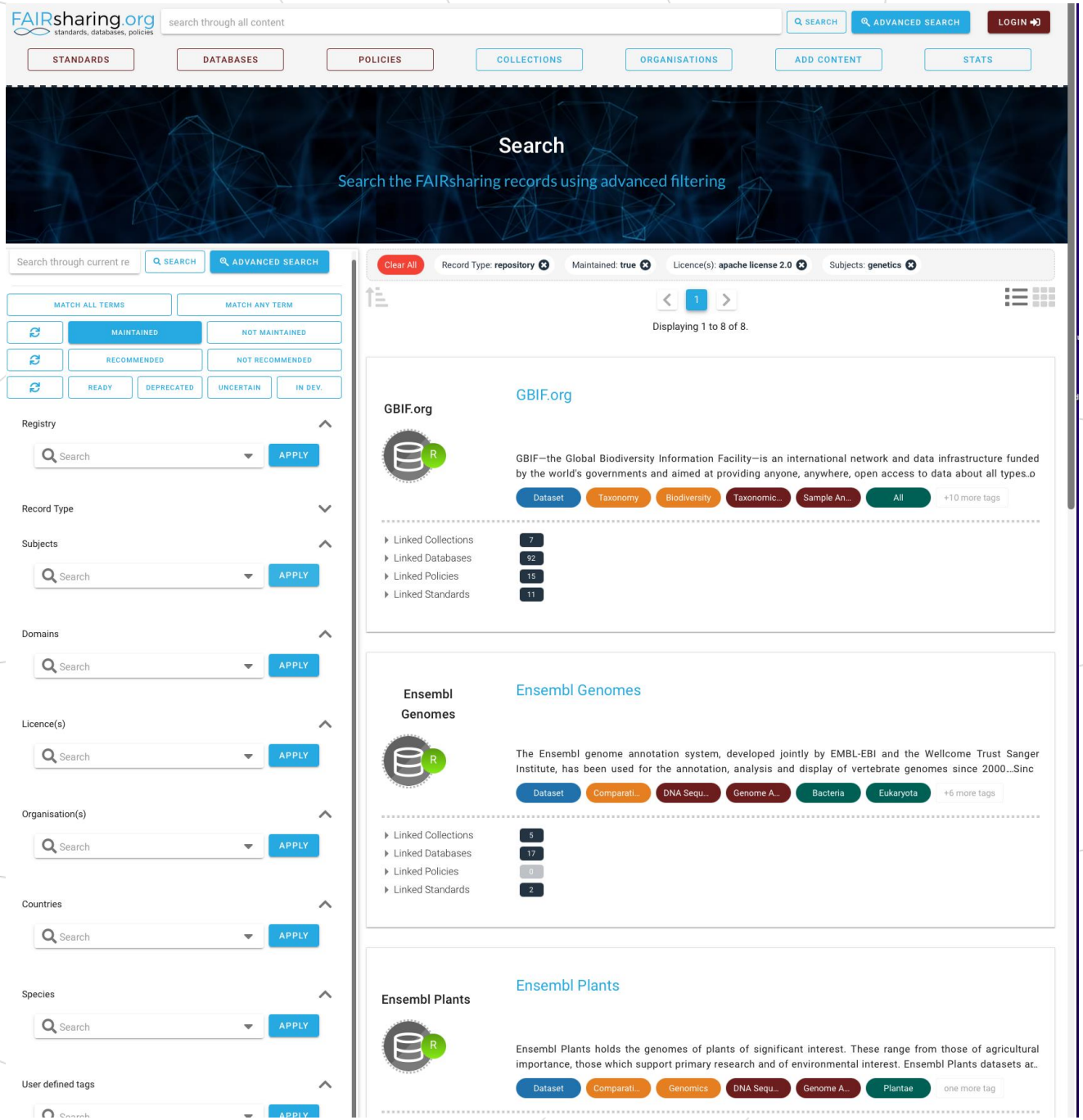
Discover
UB Discover
Subject(s): Humanities and Social Sciences | Life Sciences | Natural Sciences | Engineering Sciences
Repository type(s): institutional
Provider type(s): serviceProvider | dataProvider
Country: Germany
Discover provides access to research data from projects related to LMU Munich, which are stored and preserved in repositories of the University Library LMU.

PubData
Research Data Repository Leuphana University Lüneburg
Subject(s): Humanities and Social Sciences | Social and Behavioural Sciences | Educational Research | Economics | Life Sciences | Natural Sciences | Engineering Sciences
Repository type(s): institutional
Provider type(s): dataProvider | serviceProvider
Country: Germany
PubData is Leuphana's institutional research data repository for the long-term preservation, documentation and publication of research data from scientific projects. PubData is maintained by Leuphana's Media and Information Centre (MIZ) and is free of charge. The service is primarily aimed at Leuphana employees and additionally at researchers from cooperation partners contractually associated with Leuphana.

Research Data Repository FDAT
Research Data Repository of the University of Tuebingen
Subject(s): Humanities and Social Sciences | Life Sciences | Natural Sciences | Engineering Sciences
Repository type(s): institutional
Provider type(s): dataProvider
Country: Germany
FDAT is a research data repository hosted by the University of Tübingen, designed to facilitate long-term archiving and publication of research data. Managed by the Information, Communication and Media Center (IKM), it primarily caters to the humanities and social sciences, while welcoming researchers from all scientific disciplines at the university. Committed to high-quality data management, FDAT emphasizes the importance of adhering to the FAIR Data Principles, promoting findability, accessibility, interoperability, and reusability of the research data it contains.

UltraViolet
Subject(s): Humanities and Social Sciences | Life Sciences | Natural Sciences | Engineering Sciences
Repository type(s): institutional
Provider type(s): dataProvider
Country: United States
UltraViolet is part of a suite of repositories at New York University that provide a home for research materials, operated as a partnership of the Division of Libraries and NYU IT's Research and Instruction Technology. UltraViolet provides faculty, students, and researchers within our university community with a place to deposit scholarly materials for open access and long-term preservation. UltraViolet also houses some NYU Libraries collections, including proprietary data collections.

FAIRsharing



Zenodo

About Zenodo

- **Free-of-charge**
- **Metadata**
  (Dublin Core, Data Cite)
- **PIDs** (DOI, ORCiD, ROR)
- **Embargo, restricted access**
- **Trusted and stable**
  (CERN & OpenAIRE)
- **Open licenses**
  (CC, MIT, custom)
- **Grant dedications**

About Zenodo

- Free-of-charge
- Metadata
  (Dublin Core, Data Cite)
- PIDs (DOI, ORCiD, ROR)
- Embargo, restricted access
- Trusted and stable
  (CERN & OpenAIRE)
- Open licenses
  (CC, MIT, custom)
- Grant dedications

- **Limited**
  (50 GB & 100 files per dataset)
- **Generic repository**
- **Generic metadata only**
- **Limited access control**
- **Limited curating options**

Dataset size exception

- One-time extension up to 150 GB
- Reviewed on case-by-case basis
- The 100 files limit is non-negotiable

When is Zenodo not an option?

- Sensitive and/or personal data requiring access control
- Specialized metadata
- Large datasets
- Domain standards
- Grant provider requirements

Practice time

Before you go

Learn more

NTK

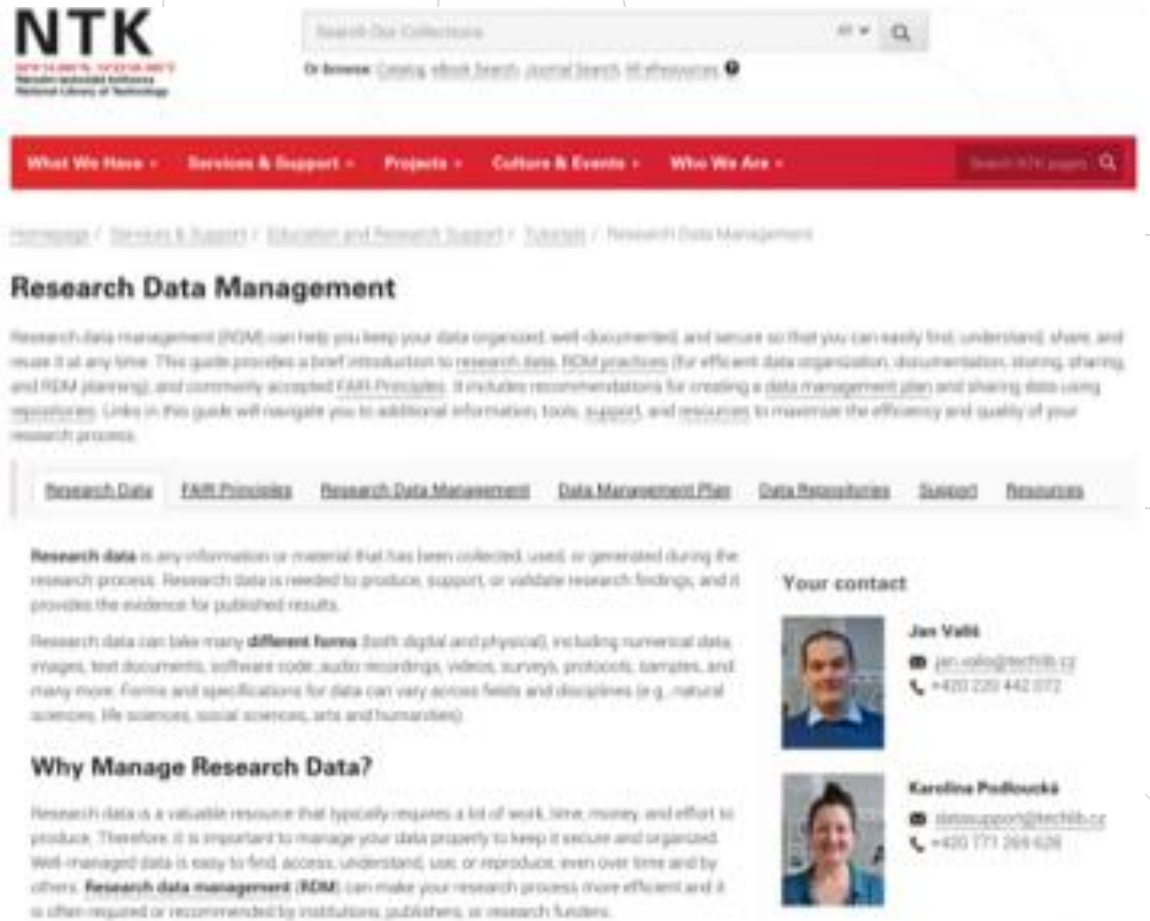
- RDM Guide
- Data Stewardship Course (CZK 0)
  self-guided, finished by a test & certificate

Self-study

- EOSC CZ – webinars (all)
- NPOS – Czech only (all)
- RDMkit (life-sciences)
- OpenAIRE (all)
- UCT Prague's Zenodo manual

Full courses

- UISK, CUNI (CZK 12 600)
- Uni Wien (CZK 72 000)
- UCT Prague (CZK 16 000)



Get help

Your institution

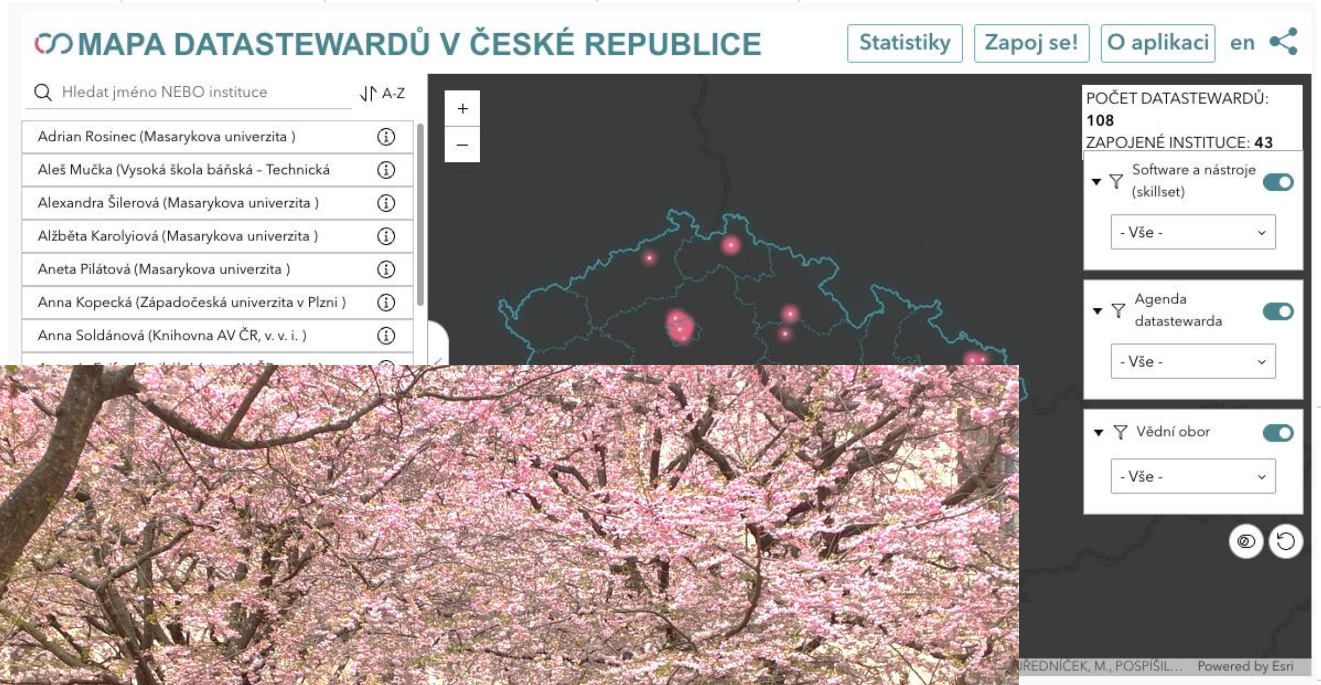
- Library/Open Science Center
- Data Steward

Community

- Map
- Manuals
- Discord
- Meetings

NTK

- Free consultations



Source: EOSC CZ; Erik Dudinský & Lucie Skřičková

NTK Information Support Team

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

1) **Schedule a free consultation with us**

Don't be shy; our team includes doctoral candidates who understand the issues you face.

2) **Attend another webinar**

3) **Explore on your own:** Tutorials, AI tools for research or STEMskiller

4) **Subscribe to our newsletter** for updates on resources, writing support, publishing, research evaluation, and training opportunities.

Any questions? Contact us at info@techlib.cz

What to take home?

- Plan ahead – look for a relevant repository before the project starts
- If possible, select a domain-specific repository
- Decide what (not) to publish
- Include ReadMe file, documentation
- Imagine how you would go over the data as a user
- If in doubt, reach out for support

Contact

Jan Vališ
jan.valis@techlib.cz

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

Questions?