Národní technická knihovna National Library of Technolog National Centre for Information Support of Research, **Development**, and Innovation

### Introduction to **Research Data Management**

... and how not to get overwhelmed by data

Workshop lecturer: Jan Vališ

Authors of the presentation: Adéla Jílková D, Martin Schätz D, Jan Vališ D

November 20, 2024

IOCB PhD Skills Day: Essential researcher competencies Content of this presentation is licensed via CC BY 4.0,

except where otherwise noted for content created by third-parties.



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.

### Agenda

### 1. What is research data and why manage it?

- Motivation and benefits of Research Data Management (RDM)
- Research data and RDM overview

### 2. How to approach Research Data Management?

- RDM frameworks (Open Science and FAIR principles)
- RDM strategies and techniques
- RDM plan

## What is research data and why manage it?

### **Research data and Research data management**

#### **Research data**

• Any information **collected**, **observed**, **generated**, or **created** during the research process to produce and support research findings

#### Research data management

- A set of practices, strategies, and activities, including data organization, documentation, storage, and sharing
- Covers all stages of the research process
- Ensures the effectiveness, reproducibility, and reuse of research data

### Why manage research data?

Join at menti.com | use code 1468 0837

# leader bold creative fast 30 transpiration

Mer 🖌 Account  $\mathcal{O}$ Content Ĥ Desian  $\odot$ Settings  $\bigcirc$ 

5

Help & Feedback

### Why manage research data?

### It can help:

#### Keep the research process organized, secure, and smooth

- Increase efficiency, save time and resources
- Share data with colleagues
- Reduce risk of data loss and improve data security

Enhance global data sharing (Open Science and FAIR principles)

- Enable data reuse and enhance collaboration
- Increase the visibility and impact of research
- Increase transparency and improve trust in research findings
- Support research integrity and validation of research results

It may be mandatory (institutional, publisher, or research funder requirements)

### **Research data**

#### **Different fields and disciplines**

- Natural and life sciences
- Medical and health sciences
- Engineering and technology
- Social sciences
- Arts and humanities

#### Different stages of research data lifecycle



Adapted from "<u>Research Data Lifecycle</u>" by <u>UK Data Service</u>, used under <u>Creative Commons Attribution license</u> (CC BY)

### **Research data management strategies**



**Organizing** Directory structure Formats, names, versions

Documentation Data description Experimental details Decisions made Metadata

Storage Backup Long-term preservation

Data access Access rights (open, restricted) Licenses

Adapted from "<u>Research Data Lifecycle</u>" by <u>UK Data Service</u>, used under <u>Creative Commons Attribution license</u> (CC BY)

### **Examples of research data requirements and policies**



Adapted from "<u>Research Data Lifecycle</u>" by <u>UK Data Service</u>, used under Creative Commons Attribution license (CC BY)

#### Funding agency policies

Open Access policy Data management plan

#### Legal and ethical requirements

National and European legislation Ethical framework for researchers Personal data protection Intellectual property rights Commercial use of data

#### Institutional policies

RDM policy Codes of conduct and ethics Data protection Partnership agreement (for collaboration)

#### **Journal & Publisher policies** Data sharing policy

## How to approach Research Data Management

## What is data?

Anything containing information Some might be self-explanatory:

- Text
- Tables

Other might not:

- Measurement results
- Images

Some might not be shared:

- Personal information
- Medical diagnoses

But there is always **metadata:** information (data) about data:

- Date of creation
- Author
- License
- Measurement device



#### Indicators of responsible research practices

#### **Responsible Research Practice**

- For knowledge to benefit research and society, it must be trustworthy.
- Trustworthy research is robust, rigorous, and transparent at all stages of design, execution, and reporting.
- Assessment of researchers still rarely includes considerations related to trustworthiness, rigor, and transparency.



**Example Indicators** 

DOI: <u>10.1371/journal.pbio.3000737.g001;</u> Available via license: <u>CC BY</u>

numerical indicators

### We need to plan in advance

- Instruments
  - Can we properly document what we are doing, and how?
- Size
  - Do we have enough storage?
- Software
  - Do we have workflow for processing of data?
  - Do we have access to proper software?
  - Can we use open file formats?
- Ethics
  - Are there any set procedures for data processing?
  - Collaboration and services!

### We need to plan in advance

#### Backup

- How and where?
- Do we need encryption and access control?
- Copyright License
  - How are we legally bound?
  - How do we want to license our results?
- Publishing
  - Can we publish data?
  - Is there any domain-specific repository?
- Archiving
  - What data to archive?
  - How long?





### **Open Science**

### **Revolution or evolution?**

source: Vers une science ouverte. Gabriela Montors, MA, PhD. Scientific Infographics, April 2021

Creating more ways to improve inclusion and access to research and higher education

Equity

Research and education are transparent for validation, and all contributions are recognised

Integrity

## Open Science

### Collaboration

Exchanging knowledge and perspectives sooner and in every step, from ideation to communication

#### Impact

Open work is more visible and can be reused and adapted to build new research and educational materials



Source: https://www.tudelft.nl/en/open-science; Available via license: CC BY

16

### What we will focus on next:

- FAIR principles
- Data naming conventions
- File formats
- Metadata
- Licensing
- Repositories
- Electronic Laboratory notebook

### FAIR - the ultimate goal







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

## **Organizing your data**

- Restrict level of folders to three or four deep
- Consider limiting the number of folders within each folder, to ten
- Include a folder within the folder structure for "documentation". This might include:
  - Project proposals/protocols
  - Consent and approval forms
  - Methodology documents
  - Data management plan
  - Code used for recodes, analysis, and outputs
  - Readme files with transformation information
  - Readme files with the full names or titles for any abbreviations used in file names
  - Codebooks or guides

### **Setup naming convention**

### Project\_YYYYMMDD\_ContentDescription\_Version.ext

Project name

Project acronym Standardized date format

lized mat Description of file content

- Author

- Instrument - Team

- Protocol used

- Language

- ...

Versioning information

- Raw

- Processed

- Denoised

- Stitched

- Cleaned

- ...

File extnesion

## **Setup naming convention**

- Avoid using spaces, dots and special characters (& or ? or !)
- Use hyphens (-), underscores (\_), or capitalization (C) to separate elements in a file name
- Include an abbreviation in the file name to identify
  - The instrument used
  - The phase (if research has multiple phases)
  - The transformation phase (i.e., original, raw, compressed, digitized, recoded, restructured, cleaned)
  - The source of third-party data (data provider or principal investigator)
  - The team (if working with multiple teams)
  - The language (if working with multiple languages)
- Include versioning within file names as appropriate

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



- Preferred vs. popular
- Open vs. proprietary

When necessary to use a proprietary format, consider including a readme.txt file in your directory that documents the name and version of the software used to generate the file, as well as the company that made the software. This could help you down the road, if you need to figure out how to open these files again.

## **Specific file types**

Here are some examples of preferred FAIR file formats for preservation:

- Images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP, SVG
- Tabular data: CSV, TXT
- Text: XML, PDF/A, HTML, JSON, TXT, RTF
- Containers: TAR, GZIP, ZIP
- Databases: XML, CSV, JSON
- Geospatial: SHP, DBF, GeoTIFF, NetCDF
- Video: MPEG, AVI, MXF, MKV
- Sounds: WAVE, AIFF, MP3, MXF, FLAC
- Statistics: DTA, POR, SAS, SAV

### Sooo... what are the metadata?

### Metadata is documentation that describes data.

Properly describing and documenting data allows you to understand and track important details of the work.

Having metadata about the data also facilitates search and retrieval of the data when deposited in a data repository.

Metadata: the who, what, when, where, why, how of your research.





### Dublin Core (1999, Dublin, Ohio)

A set of 15 metada	ata tags:
Creator	Contributo
Publisher	Title
Date	Language
Format	Subject
Description	Identifier
Relation	Source
Туре	Coverage
Rights	

Element	Definition	
Title	A name given to a resource	
Creator	An entity primarily responsible for making the content of a resource	
Subject	A topic of the content of a resource	
Description	An account of the content of the resource	
Publisher	An entity responsible for making the resource available	
Contributor	An entity responsible for making contributions to the content of a resource	
Date	A data of an event in the lifecycle of a resource	
Туре	The nature or genre of the content of a resource	
Format	The physical or digital format of a resource	
Identifier	An unambiguous reference to the resource within a given context	
Source	A reference to an another resource from which a resource is derived	
Language	A language of the content of a resource	
Relation	A reference to a related resource	
Coverage	The extent or scope of the content of a resource	
Rights	Information about rights held in and over a re- source	

## **Creative Commons licence**

Easy to understand/easy to use

### Meaning of **CC** suffix:

0 - Public domain
BY - By Attribution
ND - No Derivatives
NC - Non-Commercial
SA - Share Alike





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

#### re3data.org

Reset all Subjects 🗆 Humanities and Social Sciences (1) Life Sciences (3) Biology (3) Basic Research in Biology and Medicine (2 Biochemistry (1) Biophysics (1) Agriculture, Forestry and Veterinary Medicine Natural Sciences (10) Chemistry (10) Molecular Chemistry (10) Inorganic Molecular Chemistry - Synthes (3) **Organic Molecular Chemistry - Synthe** Characterisation (10) Chemical Solid State and Surface Research Solid State and Surface Chemistry, Mate Physical Chemistry of Solids and Surface Characterisation (2) Physical Chemistry (2) Physical Chemistry of Molecules, Liquids **Biophysical Chemistry (2)** Analytical Chemistry (2) Analytical Chemistry (2) Biological Chemistry and Food Chemistry ( Biological and Biomimetic Chemistry (1) Food Chemistry (1) Polymer Research (2) Preparatory and Physical Chemistry of P Experimental and Theoretical Physics of Polymer Materials (1) Theoretical Chemistry (2) Theoretical Chemistry: Electron Structure (1) Physics (2) Condensed Matter Physics (1) Statistical Physics, Nonlinear Dynamics, Co and Fluid Matter, Biological Physics (1) Particles, Nuclei and Fields (1) Engineering Sciences (1) Materials Science and Engineering (1) Materials Science (1) Countries 🕀 API 🕀

Filter



## **Pricing estimation**

### DSW Storage Costs Evaluator



Detailed storage properties 🗙

### **Electronic laboratory notebook**



Výzkum Studium a kariéra IOCB Boston Tech transfer Novinky

O nás

EN

Kontakt

ΛΙΤΚ

Národní technická knihovn: National Library of Technology

Q

Kontakt > Adresář > Matúš Drexler Domů



### **Matúš Drexler**

matus.drexler@uochb.cas.cz

ID

+420 220 183 498

## How it looks in practice: VMCF

### Many limitations:

- RDM
- Length of experiments
- Data ownership
- Documenting
- Ethics
- Access planning
- Cost management



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

### What to take home?

- Open Science is an evolution.
- Managing data is a good scientific practice.
- Managing and sharing data can save time, money, and create impact.
- Communities of researchers worldwide define standards, usually they are open to others joining their efforts. The same is happening at the national level.
- Research data management is a helpful tool, not just an administrative task.
- Funding agency will, in time demand (or already are demanding) Data Management Plans, and support RDM tasks financially.
- There are already many resources on-line to learn from.

### **Research data management guide**





Or browse: Catalog, eBook Search, Journal Search, All eResources, 🔮

What We Have + Services & Support + Projects + Culture & Events + Who We Are +

Search NTK pages... Q

Homepage / Services & Support / Education and Research Support / Tutorials / Research Data Management

#### **Research Data Management**

Research data management (RDM) can help you keep your data organized, well-documented, and secure so that you can easily find, understand, share, and reuse it at any time. This guide provides a brief introduction to research data, <u>RDM practices</u> (for efficient data organization, documentation, storing, sharing, and RDM planning), and commonly accepted <u>FAIR Principles</u>. It includes recommendations for creating a <u>data management plan</u> and sharing data using <u>repositories</u>. Links in this guide will navigate you to additional information, tools, <u>support</u>, and <u>resources</u> to maximize the efficiency and quality of your research process.

Research Data FAIR Principles Research Data Management Data Management Plan Data Repositories Support Resources

**Research data** is any information or material that has been collected, used, or generated during the research process. Research data is needed to produce, support, or validate research findings, and it provides the evidence for published results.

Research data can take many **different forms** (both digital and physical), including numerical data, images, text documents, software code, audio recordings, videos, surveys, protocols, samples, and many more. Forms and specifications for data can vary across fields and disciplines (e.g., natural sciences, life sciences, social sciences, arts and humanities).

#### Why Manage Research Data?

Research data is a valuable resource that typically requires a lot of work, time, money, and effort to produce. Therefore, it is important to manage your data properly to keep it secure and organized. Well-managed data is easy to find, access, understand, use, or reproduce, even over time and by others. **Research data management (RDM)** can make your research process more efficient and it is often required or recommended by institutions, publishers, or research funders.

#### Your contact

All V Q



▲ datasupport@techlib.cz ▲ 232 002 562

Adéla Jílková



#### Karolina Podloucká

datasupport@techlib.cz
 771 269 628



### Contacts

# Jan Vališ 🕩 jan.valis@vscht.cz

### Martin Schätz (D) martin.schatz@techlib.cz

### **Questions?**

Source: https://xkcd.com/2582/; Available via license: CC BY NC 2.5





36