



# Research Data Management

Dunja Legat, M.Sc.

The ATHENA Soft & Research Skills Course,

Chania, 19th of May, 2023

# What is Open Science?

- Open Science means **opening all phases of research** to the public.
- The **dissemination of scientific knowledge**, that is as wide as possible, free of charges to all users, and accessible online.
- The research process should be more **transparent**, and **research outputs findable and available** in standardized formats through an interoperable infrastructure.
- Approach to the scientific process based on **cooperative work** and ways of disseminating knowledge, improving accessibility to and re-usability of research outputs using digital technologies and collaborative tools. (Source: EOSC Glossary)

<b>Open Access</b>	<b>Open Science</b>	<b>Open Data</b>
<b>Open Source</b>		<b>Open Education</b>
<b>Open Hardware &amp; Open Software</b>		<b>Citizen Science</b>
<b>Research Infrastructures and the EOSC</b>		<b>Open Science Skills</b>
		<b>Research Integrity</b>

## Who's involved?



researchers



institutions



policymakers



publishers



libraries



funders

# European Union's Policy on Open Science

The EU's open science policy comprises eight aspects:

1. Training and skills for implementing open science in practice,
2. Recognizing, promoting and rewarding open science practices,
3. New-generation metrics and altmetrics,
4. Open publishing and encouraging early sharing of research results,
5. Open data,
6. Research integrity and reproducibility of scientific findings,
7. European Open Science Cloud (EOSC),
8. Citizen science.







# European Open Science Cloud (EOSC)

ATHENA

## Browse EOSC Marketplace Resources



ALL CATALOGS



PUBLICATIONS



DATA



SOFTWARE



SERVICES



DATA SOURCES



TRAININGS



OTHER

The **European Open Science Cloud** is an initiative that aims to develop "a network of FAIR data and services" for European science.

The European Open Science Cloud is set out to solve this problem by establishing digital infrastructure that will ensure the interoperability of databases and **a single entry point**, and easy and open access to data from publicly funded research as a result.

Aspects of this digital infrastructure range from visualization and analytics to long-term data storage and monitoring the adoption of open science practices.



<https://eosc-portal.eu/>

# Horizon Europe

ATHENA



- Horizon Europe is the **ninth key financial program of the European Union** for financing research and innovation, which will last from 2021 to 2027: it follows the Horizon 2020 program, which has already partially **obliged researchers to the practices of open science**.
- With Horizon 2020, **open access publishing became mandatory**.
- With Horizon Europe, **the sharing of research data and the creation of research data management plans have also become mandatory**.
- Research data management must be responsible, planned in advance and compliant **with the FAIR principles** as well as the principle »as open as possible, as closed as necessary«.





# Eligible Exemptions from Openness

**Open access does not necessarily mean that data become open immediately, indefinitely, or unconditionally.**

**Despite the access restrictions, it is necessary to create metadata that prove the existence of the data and describe the access options.**

**Eligible exemptions from openness must be substantiated in the research data management plan.**





# What are Research Data?

ATHENA



Currently, there is no universally accepted definition of research data.



Definition by Springer Nature:  
<https://www.springernature.com/gp/authors/research-data>



Definition by OECD:  
[OECD Principles and Guidelines for Access to Research Data from Public Funding](#)




Definition by CODATA:  
Committee on Data of the International Science Council;  
[www.codata.org](http://www.codata.org)

OECD: Research data are defined as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research and are commonly accepted in scientific community **as necessary to validate research findings.**

# Recommendations

By some definitions, **any number and file you create in your work could be considered as a piece of research data.**



Such a strict definition of research data **is not sustainable from a practical point of view.**



Before starting, you should **check your funder's policy on research data.**

If it does not contain specific provisions, the decision on the scope of open research data **is left to you.** You define it in the **research data management plan.**

# Life Cycle of Research Data: Management of Research Data

1. Collect: experiment, observation, measurements, simulation, survey...
2. Manage: validation, anonymization, transcription, digitization...
3. Analysis: interpretation of data and production of results...
4. Store, Archive: recommended formats, media, backups, metadata and documentation preparation, archiving...
5. Publish, Share: data dissemination, access control, copyright and promotion...
6. Reuse: secondary analyses, upgrading of research, using data for teaching and learning, citing data...



# There are a number of reasons why research data management is important:

- data, like articles and books, are **part of our scientific production**,
- data (especially digital) **is fragile and easily lost**,
- financiers and publishers **require data deposit and data handling planning**,
- managing research data **saves time and resources** for us and the users of our data in the long term,
- good data management **helps prevent errors and increases the quality** of our analyses,
- well-documented and accessible data **enable others to confirm and repeat (reproducibility) our research**,
- research data management **facilitates the sharing of research data**,
- shared, the data **can lead to valuable discoveries by others** who are not part of the original research team.



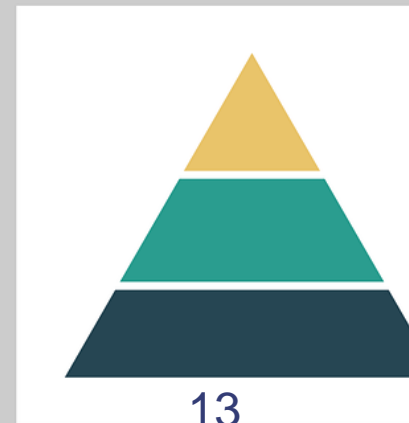
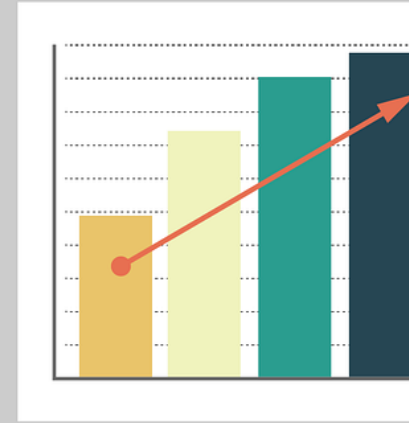
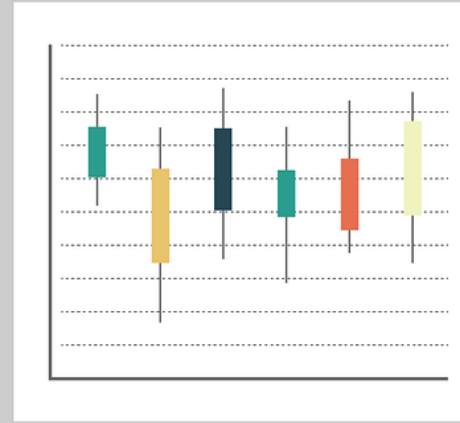
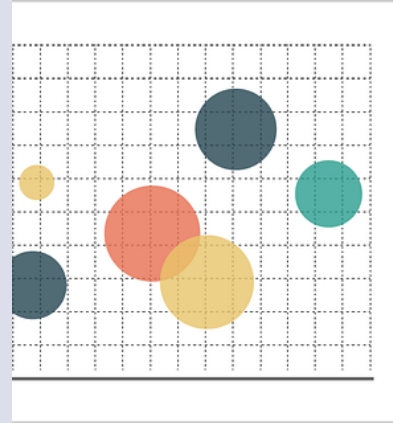
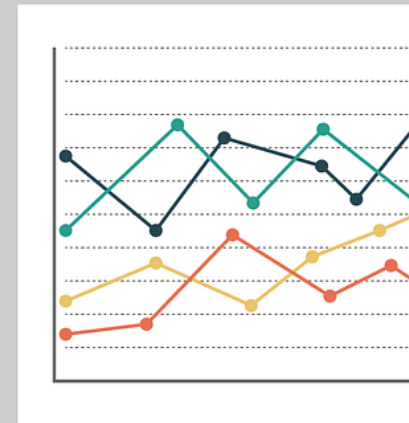
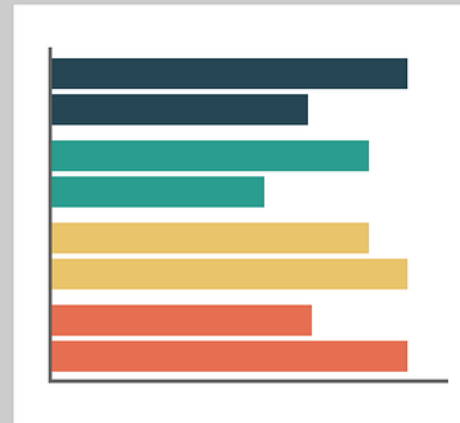
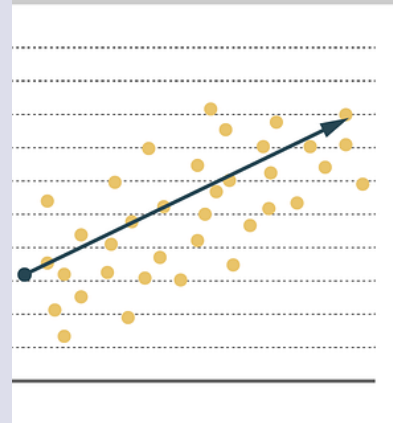
# Types of Research Data

Research data represents everything that was used or created during the research process and supports or confirms the research findings.

Data can appear in a variety of forms, including:

- diaries,
- survey answers,
- software and code,
- measurements from laboratory or field equipment,
- images (such as photographs, films, scanned documents),
- audio recordings
- physical samples, etc.

**Everything that is necessary to validate research findings.**



# What Counts as Open Research Data?

Open, openly accessible or publicly available data are data that meet **the FAIR principles**. In short, this means that they are:

- **deposited in a trusted repository,**
- **described in a formal, generally accessible and widely used language** for the dissemination of knowledge,
- **licensed with an open license** and
- **equipped with all information** (e.g., methods, protocols, software) that enable other researchers to understand and reuse them.



# Research data management (RDM)

**Is a process in the research lifecycle that includes**

- the creation (collection or acquisition) of research data,
- their organisation,
- digital stewardship,
- storage, (long-term) preservation,
- security,
- quality assurance,
- allocation of persistent identifiers,
- providing metadata,
- issuing appropriate licenses and procedures for data exchange, sharing and reuse.

# Stakeholders of RDM



Funder



Ethics review



Legal expert



Researcher



Publisher



Repository  
operator



Infrastructure  
provider



Research  
support staff



Institutional  
administrator

Miksa T, Simms S, Mietchen D, Jones S (2019) [Ten principles for machine-actionable data management plans](#)

## Open Science(RDM) Support Service: DATA STEWARDSHIP

- Teamwork and Collaboration, Infrastructure
- Often in libraries (Data Steward, Data Librarian, etc.)



# What is Data Management Plan (DMP)

Research data management must be accurately **planned in advance** with the help of a research data management plan (RDMP).

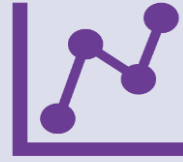
A DMP is a document that **describes how research data is to be managed both during the project and after.**

DMPs are often submitted as a part of grant applications but are useful whenever researchers are creating data.

# What is Data Management Plan (DMP)



**Living  
document**



**What  
information is  
needed to be  
able to read  
and  
understand  
the data in the  
future?**



**We have to  
consider ...**

# Essential content of DMP

ATHENA

A research data management plan must contain the following essential information:

- 1. Description of Data**
- 2. Standards and Metadata**
- 3. Persistent Identifiers**
- 4. Digital Stewardship and Data Protection**
- 5. Data Sharing Terms**
- 6. Management of Other Research Results**
- 7. Data Management Costs**



# Online Tools for Creating DMP

- [DMPOnline](#) (large collection of templates)
- [Data Stewardship Wizard](#) (interactive questionnaire linked to external sources; hints; FAIR metrics; machine readable plans)
- [Argos](#) (machine readable plans, ki upoštevajo načela FAIR; možnost sodelovanja in objave)
- [DMP Template - Horizon Europe](#)





# FAIRification

FAIR research data are those data that are

- Findable,
- Accessible,
- Interoperable and
- Reusable.

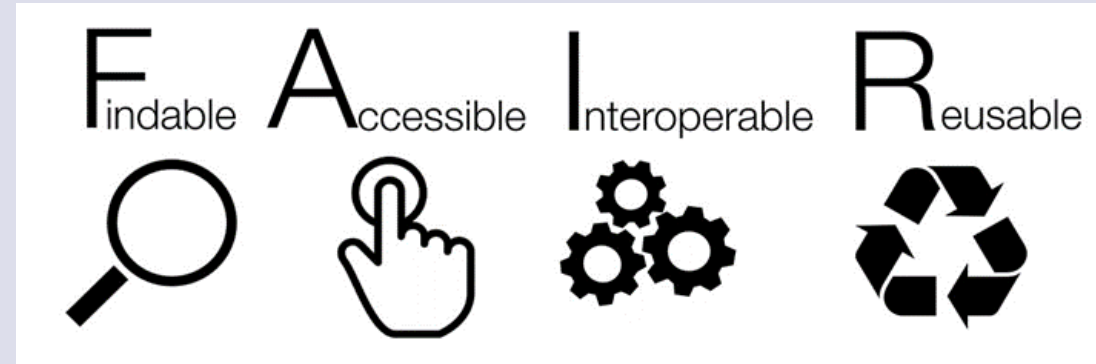


Image CC-BY-SA by Sangya Pundir

Reasons:

- In the current digital ecosystem, **people are increasingly dependent on computer support** to cope with the increasing volume, complexity and speed of data creation.
- FAIR principles: **to improve machine discoverability and data use** (ie, the ability of computer systems to find, access, connect to, and reuse data with little or no human intervention).

## FINDABLE

- data and metadata are assigned a globally **unique and persistent identifier** (e.g., DOI, Handle),
- data are described with **rich metadata**,
- metadata clearly and explicitly include the permanent identifier,
- data and metadata are registered or **indexed in searchable bibliographic indexes** (e.g., in repositories or library catalogues)

## ACCESSIBLE

- data and metadata are retrievable by their identifier using **a standardised communications protocol** (https, ftp, etc.)
- **Metadata are accessible**, even when the data are no longer available.

## INTEROPERABLE

- use a **formal, accessible, shared, and broadly applicable language** for knowledge representation,
- use **vocabularies** that follow FAIR principles,
- metadata contains **qualified references** to other metadata,
- use of **non-proprietary, open formats**.

## REUSABLE

- Data and metadata are richly described** with a plurality of accurate and relevant attributes:
- data and metadata are released with a clear and accessible **data usage license**,
  - data and metadata are associated **with detailed provenance**,
  - data and metadata meet domain-relevant community standards

# Make your software FAIR

- Use a publicly accessible repository with version Control (for example version control with Git)



### Software Heritage

- 1 Prepare your public repository  
README, AUTHORS & LICENSE files
- 2 Save your code  
<http://save.softwareheritage.org/>
- 3 Reference your work  
(full repository, specific version or code fragment)

- Add a license to your code (open software licenses)
- Register your code in a community registry
- Enable citation of the software
- Use a software quality checklist



Recommendation: Open software licensing  
([https://www.youtube.com/watch?v=gvj4dGYiK\\_M](https://www.youtube.com/watch?v=gvj4dGYiK_M))

# Formatting Research Data

Research data **must be properly formatted** before sharing so that other researchers can understand and reuse them. In this way, we satisfy **the FAIR principles of interoperability and reusability**.

In repositories, research data will stand on their own without an accompanying context, which is why it is necessary to pay attention to the appropriate

- **naming of files,**
- **the hierarchy of file folders,**
- **and metadata** (which can be described in ReadMe files or data articles).

# File Naming

Recommendations for naming research data:

- we prepare **an agreed structure for the naming of research data**, consisting of logical elements;
- Relevant information for naming files is: project name or acronym, researcher's name or initials, data type, research method, place and date of research, file version number;
- we use the agreed nomenclature consistently;
- the agreed nomenclature should be **used by all members of the research group**;
- we **avoid using a similar name** for multiple files;
- we avoid vague names;
- file names should not be too long, **maximum 32 characters**;
- we use letters and numbers according to the **ASCII standard** (a - z, A - Z and 0 - 9);
- we avoid using periods or special characters such as &, \*, %, #, :, ( ), !, @, \$, ^, ~, ' , { } , [ ] , ? , < >
- write the **date in the ISO 8601** standard (YYYYMMDD);
- for possible numbering, we **use leading zeros**, e.g.: 001 – 999;
- instead of a space, **use an underscore \_ or a minus -**;
- we avoid **labels such as final version, revision, final**, etc., for major changes we use no. versions, e.g. V2;
- when renaming a bunch of files, we **use the batch renaming tool**;
- the complex structured nomenclature is described in the file README.txt.

# Example of File Naming

An example of good practice: **20190523\_H2020MatChem\_GL\_exp5\_c2\_XRF1**

This name contains information that is important to the author, the research group and other users of the data:

- **The date the file was created**, i.e. 23 May 2019 in YYYYMMDD format,
- **The name of the hypothetical project** titled "Materials Chemistry" (abbreviation MatChem), which was financed within the framework of the Horizon 2020 (H2020) programme,
- **Initials of the hypothetical author**, i.e. G. L.,
- **The title of the experiment**, i.e. exp5 ("Experiment 5")
- **Designation of the compound**, i.e. C2 ("Compound 2")
- **Designation of the analysis**, i.e. XRF1 (first measurement with X-ray fluorescence).

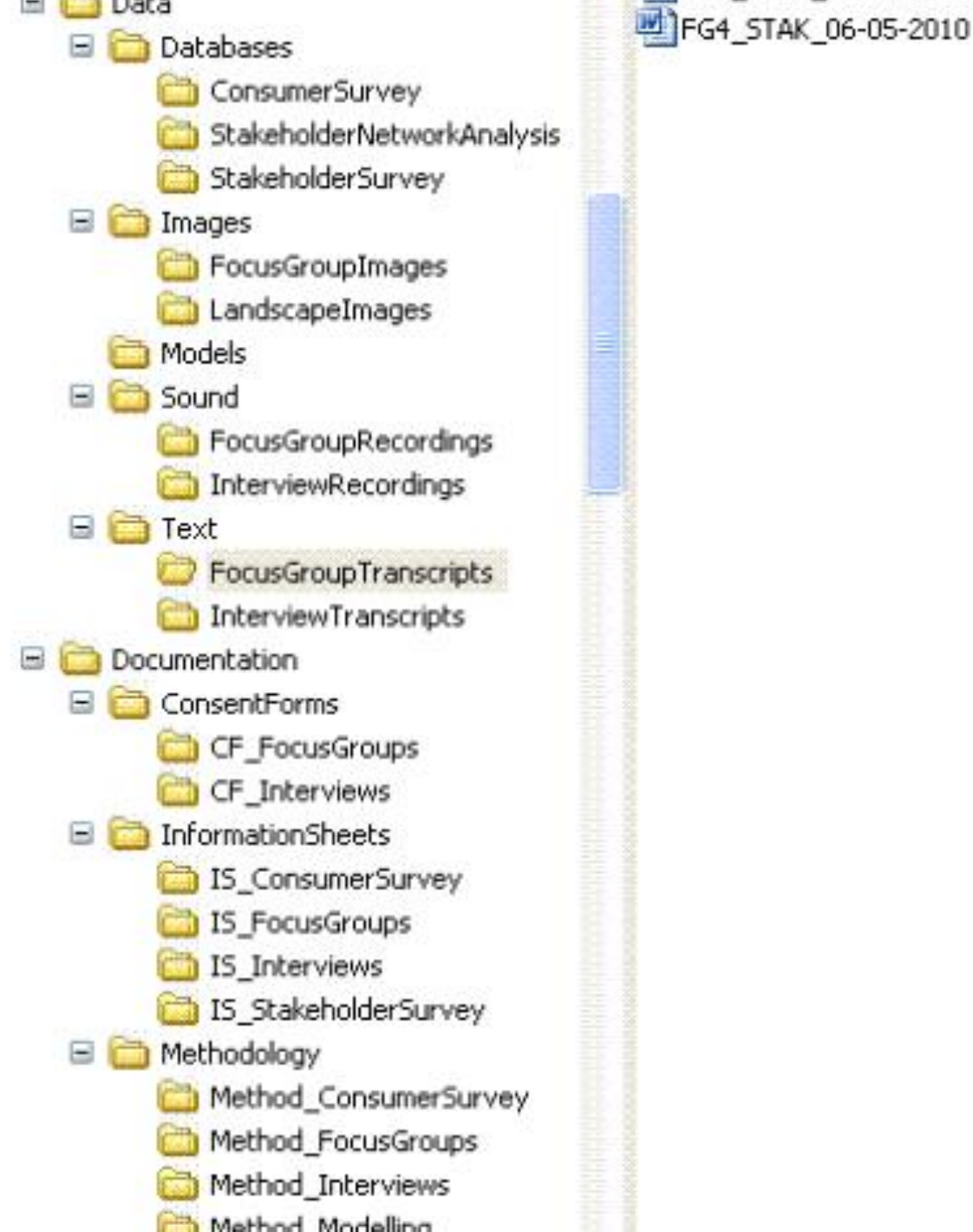
**MIT Libraries Data  
Management  
Services: Batch file  
renaming tools**

Source: DiRROS Data: <https://dirrosdata.ctl.uni-lj.si/en/raziskovalni-podatki/oblikovanje-podatkov-za-deljenje/> )



# Organizing the Research Data – File Folders

1. A folder structure is prepared that will meet the needs of the research project.
2. In doing so, we pay attention **to the type of research data and the way in which both raw and analyzed data, methods, documentation and other supporting files will be organized.**
3. The folders are named **with unique names that correspond to the research project.**
4. Determine **the appropriate depth of the folder hierarchy.**
5. **Tagging files helps us find files** in the research project folders.



# File Formats

For reuse, it is recommended to collect data in formats that are more likely to be accessible in the future:

- **non-proprietary,**
- **open and**
- **documented standards that are uncompressed, exchangeable, widely used by** the research community and use standard character encoding (ASCII, UTF- 8).

The choice of format also depends on the policy of the repository we choose to publish.

# File Formats

Type of Data	Recommended Formats	Acceptable Formats
Quantitative Tabular Data	.csv, .tab, .por, .xml	.txt, xls, .dbf, .ods, .sav, .dta, .mdb
Qualitative data. Textual.	.rtf, .txt, .xml	.html, .doc
Digital image data	.tif	.jpg, .gif, .tif, .tiff, .raw, .psd, .bmp, .png, .pdf
Digital audio data	.flac	.mp3, .aif, .wav
Digital video data	.mp4, .ogv, .ogg, .mj2	.avchd
Documentation and scripts	.rtf, .pdf/A, .xhtml, .htm, .odt	.txt, .doc, .xls, .xml

**An open file format is a file format for storing digital data, which can be used and implemented by anyone.**

**List of open file formats:**

[https://en.wikipedia.org/wiki/List\\_of\\_open\\_file\\_formats](https://en.wikipedia.org/wiki/List_of_open_file_formats)

**More precisely: UK Data Service:** Recommended Formats (Source: <https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/>)

# What information is needed to be able to read and understand the data in the future?

Additional documentation helps users of our data understand and reuse the data. Examples of additional documentation are:

- lab notebooks (e.g. Jupyter Notebook) and protocol descriptions (e.g. protocols.io),
- questionnaires,
- codes, variable definitions, data dictionaries,
- measurement units,
- ontologies, controlled dictionaries,
- programming language syntaxes and software output files,
- information on equipment settings and instrument calibration,
- database schema, file directory structure description, naming structure,
- methodology reports,
- information on analysis and procedures,
- information about the origin of the acquired or digitized data,

If we have developed software code to acquire or process data, we consider whether it is also necessary to store it with data for reproducibility purposes.

## Readme.txt

Example of the University of  
Cornell: Template of  
Readme.txt File

[AUTHOR\\_DATASET\\_ReadmeTemplate.txt](#)

# Secure storage of research data

When choosing a storage location, we consider functionalities such as

- automatic backup storage,
- data sharing, and
- data encryption.



We have to consider guidelines for safe storage of data during research in:

- **physical aspects of data protection** (password protection, access restriction, physical protection),
- **backup storage** (3 completely separate copies of the data),
- **handling of personal and sensitive data** (anonymisation etc., GDPR, encryption),
- **secure data exchange** (SFTP, HTTPS, cloud services).

# Publish, Share and Reuse

Similar to citing scientific articles, sharing research data and allowing other researchers to download, use, and cite it can lead to greater research impact in our field.



Increased transparency and trust in their work



Easier availability of research results for other scientists



Reproducibility and reuse enabled by verifiable results



Increased readability, citation, and impact



Long-term archiving and preservation



New ways of gaining recognition and reputation



New projects and employment opportunities



Increased visibility and impact



## Publish the Research Data

We deposit the data in a suitable repository for our scientific discipline.

The repository can be found in the international **re3data repository registry** or use the **DataCite repository finder**.

1. the data can be uploaded to the institutional repository,
2. or to one of the widely used international general repositories, such as Harvard Dataverse, Dryad, FigShare, Mendeley Data, OSF or Zenodo ([Comparison of general repositories](#) →)

When choosing a repository for permanent storage of research data, it is important that **the repository works in accordance with the FAIR principles** (assigns permanent identifiers, allows license selection and supports rich metadata description).



# Metadata Schemas and Standards

- When preparing metadata, it is best to follow the metadata schemas **prescribed by the repository** where you intend to deposit your data.
- With exceptions to general standards such as **Dublin Core and schema.org, metadata standards** mostly apply only within a specific domain or specialised field.
- **A list of domain-specific metadata schemas** can be found at UK Digital Curation Centre and Research Data Alliance:
  - <https://www.dcc.ac.uk/guidance/standards/metadata/list>
  - <https://rd-alliance.github.io/metadata-directory/>

# Provenance (Origin) of Research Data

Provenance is one of the most important aspects of metadata.

The concept of provenance or the origin of research data refers to

- **all information about the circumstances of data creation**, e.g., about authors, time of creation, research equipment and its calibration, etc.
- Provenance information not only enables interoperability and reusability of the data but also contributes to maintaining research integrity and combating research irreproducibility.

**More:** <https://dirrosdata.ctk.uni-lj.si/en/metapodatki/provenienca/>

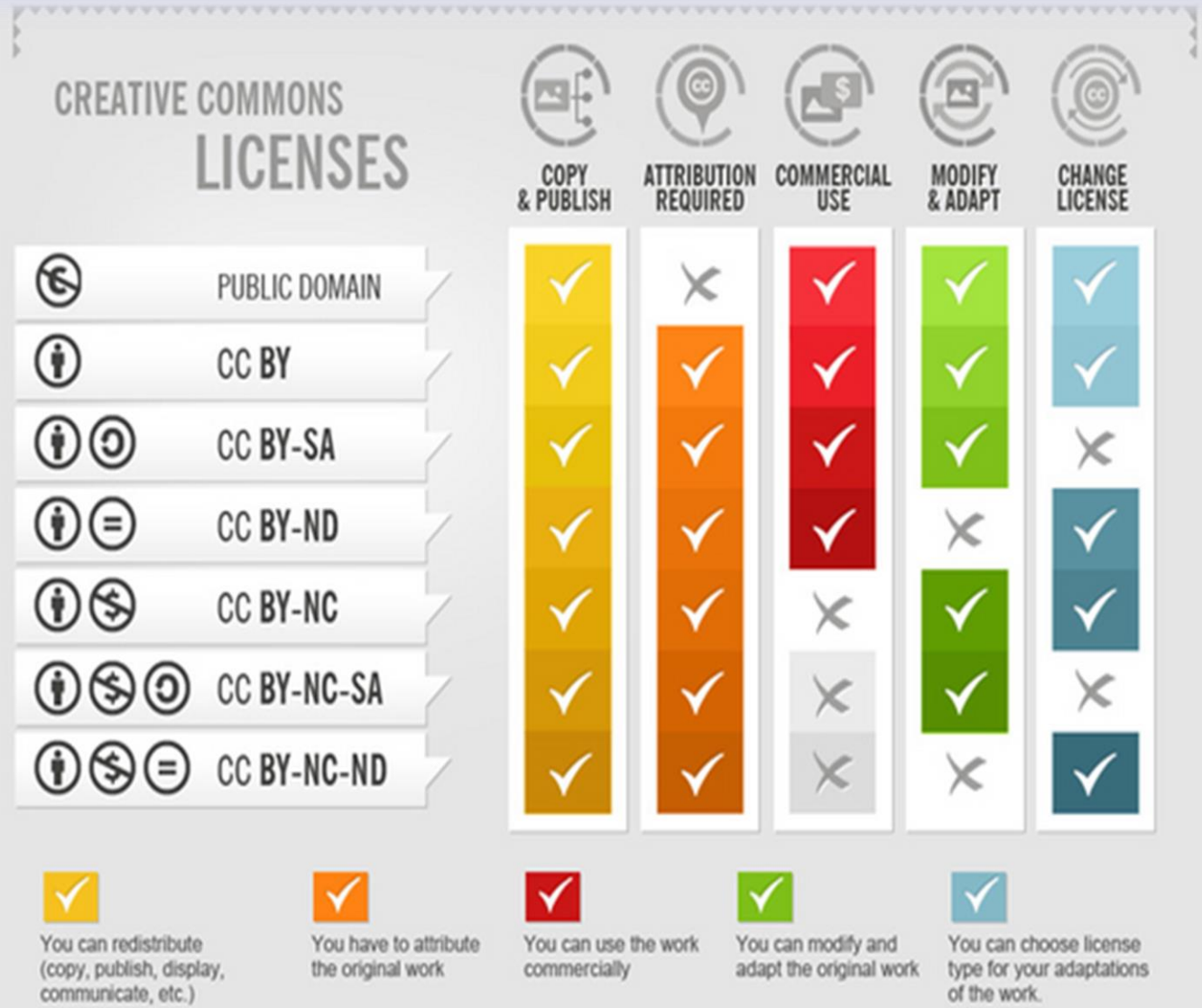
# Definition of rights and limitations of data use

Before publishing data, it is important to define how the research data set will be used.













This is governed by a license, which expresses the permission by which the author defines the conditions for the reuse of the work, in this case the research data set.


By choosing the appropriate license, the author retains the copyright and allows others to use the work under clearly defined conditions.


## Creative Commons Licences





**CREATIVE COMMONS LICENSES**


		 COPY & PUBLISH	 ATTRIBUTION REQUIRED	 COMMERCIAL USE	 MODIFY & ADAPT	 CHANGE LICENSE
 PUBLIC DOMAIN		✓	✗	✓	✓	✓
 CC BY		✓	✓	✓	✓	✓
 CC BY-SA		✓	✓	✓	✓	✗
 CC BY-ND		✓	✓	✓	✗	✓
 CC BY-NC		✓	✓	✗	✓	✓
 CC BY-NC-SA		✓	✓	✗	✓	✗
 CC BY-NC-ND		✓	✓	✗	✗	✓

 You can redistribute (copy, publish, display, communicate, etc.)

 You have to attribute the original work

 You can use the work commercially

 You can modify and adapt the original work

 You can choose license type for your adaptations of the work.

# Research Data Set Citations:

**In-text citation;  
Data citations must also be  
included in the bibliography**

## **Dryad:**

Li, Yanjun; Gao, Yingzhi; van Kleunen, Mark; Liu, Yanjie (2022), Data from: Interactive effects of herbivory and the level and fluctuations of nutrient availability on dominance of alien plants in synthetic native communities, **Dryad, Dataset**, <https://doi.org/10.5061/dryad.fj6q573vn>

## **Social Sciences Archive:**

Petravić, L., Arh, R., Gabrovec, T., Jazbec, L., Rupčić, N., Starešinič, N., ... Slavec, A. (2021). Odnos do cepljenja proti SARS-CoV-2, 2020: Priložnostni vzorec [Podatkovna datoteka]. Ljubljana: Univerza v Ljubljani, Arhiv družboslovnih podatkov. ADP - IDNo: SARSPR20. [https://doi.org/10.17898/ADP\\_SARSPR20\\_V1](https://doi.org/10.17898/ADP_SARSPR20_V1)


# Metadata (example Zenodo)

ATHENA

June 19, 2022

Dataset Open Access

## Basic data visualisations for Figshare State of Open Data 2021 survey

 Horton, Laurence

Data collector(s)



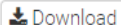

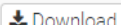

 Nature Research

R markdown files for:


- Downloading and cleaning data from the State of Open Data survey 2021
- Basic visualisations of responses to questions in the State of Open Data survey 2021
- HTML file of those visualisations.

Free text fields are included in the markdown but have been turned off for knitting and in the HTML file.


Files (15.6 MB)

Name	Size	
sod_cleaning.Rmd	18.1 kB	
md5:b96412d3a7c8469a62b6aa6a308c5623 		
sod_descriptives.html	15.4 MB	
md5:92a5210fc9fcb9e2f38e6633536557f2 		
sod_descriptives.Rmd	133.0 kB	
md5:1ccb67718c287224f96c9211806c79f0 		

38

 views

11

 downloads

[See more details...](#)

Indexed in

OpenAIRE

Publication date:

June 19, 2022

DOI:

DOI [10.5281/zenodo.6662740](https://doi.org/10.5281/zenodo.6662740)

Keyword(s):

[author survey](#) [research data](#) [Open Data](#) [R markdown](#)  
[FAIR data](#) [Figshare](#) [RDM](#) [Research Data Management](#)  
[Data sharing](#) [Survey](#) [2021](#)

Related identifiers:

Derived from  
[10.6084/m9.figshare.17081231.v1](https://doi.org/10.6084/m9.figshare.17081231.v1) (Dataset)

License (for files):

 [Creative Commons Zero v1.0 Universal](#)

Versions

Version 1

Jun 19, 2022

[10.5281/zenodo.6662740](https://doi.org/10.5281/zenodo.6662740)

**Cite all versions?** You can cite all versions by using the DOI [10.5281/zenodo.6662739](https://doi.org/10.5281/zenodo.6662739). This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

Share



Cite as

Horton, Laurence. (2022). Basic data visualisations for Figshare State of Open Data 2021 survey (Version 1) [Data set]. Zenodo.  
<https://doi.org/10.5281/zenodo.6662740>

Start typing a citation style...



# National Initiatives for Open Science in Europe – NI4OS promo video



**University of Maribor  
Open Science Summer  
School 2023**

**11.9. – 15.9. 2023**

**[open.um.si](https://open.um.si)**

**INVITATION**



**University of Maribor Library  
Open Science Team**

**Odprimo:UM  
(e.g. Open:Mind)**

**odprimo@um.si**

**THANK YOU**

