

# No (available) data, no paper

Better to start with the end in mind

Flora D'Anna, PhD

Data steward at ELIXIR Belgium

Research Data Management: your ally on the way to your publication

12<sup>th</sup> December 2022, Gent



# Outline

1. What is Research Data Management (RDM)
2. Why do we do RDM
3. Open data VS FAIR data
4. What are the benefits of good RDM
  - During the project
  - After the project
5. How to approach RDM starting with "the end" (FAIR metadata) in mind
6. ELIXIR (Belgium) resources for RDM



# What is Research Data Management (RDM)?

Let's write a definition

*use pens and papers*



# Research Data Management (RDM)

All actions necessary to ensure  
that research data is



Well described



Easy to find  
(for humans and machines)



Secure



Reusable

# Data life cycle to describe RDM

All actions necessary to ensure  
that research data is



Well described



Easy to find  
(for humans and machines)

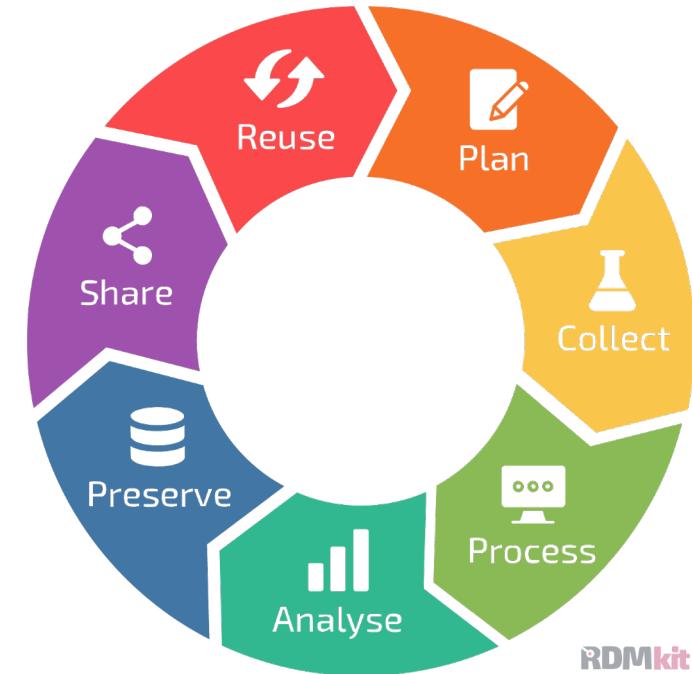


Secure



Reusable

Data life cycle: RDM stages



<https://rdmkit.elixir-europe.org>

Why do we do RDM?  
Let's write possible reasons

*use pens and papers*



# Funders' and journals' data policy: focus on data

Before



RDA Plenary Cartoons by Auke Herrema in CC-BY

After (now)

**nature**

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

nature > editorials > article

EDITORIAL | 13 March 2018

## Everyone needs a data-management plan

They sound dull, but data-management plans are essential, and funders must explain why.

[Twitter icon](#) [Facebook icon](#) [Email icon](#)

Keep your research data organized with a management plan. Credit: Jasper Juinen/Bloomberg/Getty

doi: <https://doi.org/10.1038/d41586-018-03065-z>

## Funders requirements:

- Data management plan (DMP)
- Data availability policy:  
*“as open as possible, as closed as necessary”*



# Data management plan (DMP)

Sessions:

- “Planning for efficiency” session

What's “data” in DMP ? How are you going to manage it during and after the project?



Digital or digitalised information



Physical materials

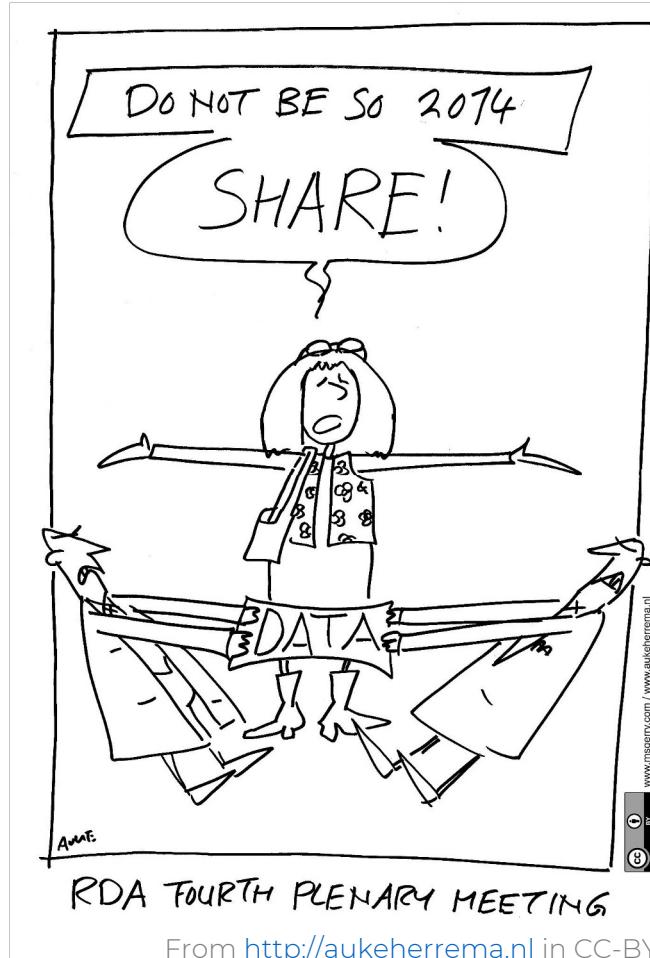


Software, other research output

# Data availability policy: as open as possible, as closed as necessary

Sessions:

- “Ethical and legal constraints on the sharing of personal data”



## *Open Definition*

“Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness).

[From Open Definition](#)



# Journals: no available data, no paper

## Data availability

Please provide a Data Availability statement in the Methods section under "Data Availability"; detailed guidance can be found in our [data availability and data citations policy](#). Certain data types must be deposited in an appropriate public structured data depository (details are available [here](#)), and the accession number(s) provided in the manuscript. Full access is required at publication. Should full access to data be required for peer review, authors must provide it.

We encourage provision of other source data in unstructured public repositories such as [Dryad](#) or [figshare](#), or as supplementary information. To maximize data reuse, we encourage publication of detailed descriptions of datasets in [Scientific Data](#).

## Computer code

Any previously unreported custom computer code used to generate results reported in the manuscript and that are central to the main claims must be made available to editors and referees upon request. Any practical issues preventing code sharing will be evaluated by the editors who reserve the right to decline the manuscript if important code is unavailable. At publication, *Nature* journals consider it best practice to release custom computer code in a way that allows readers to repeat the published results.

For all studies using custom code that is deemed central to the conclusions, a statement must be included in the Methods section, under the heading "Code availability", indicating whether and how the code can be accessed, including any restrictions.

Journals: no available data, no paper



**Cell Press STAR★Methods**

Our STAR Methods format (Structured, Transparent, Accessible Reporting) [viewed](#) and is now used in all Cell Press life science journals, as well as in *iScience*.

### Submit your featured protocols to STAR Protocols

Did you know you could quickly and easily convert the key protocols from your STAR Methods section into a published article?

***STAR Protocols*** is an open access protocol journal from Cell Press. The primary criteria for publication in ***STAR Protocols*** are usability and reproducibility. ***STAR Protocols*** welcomes protocols—from basic to advanced—

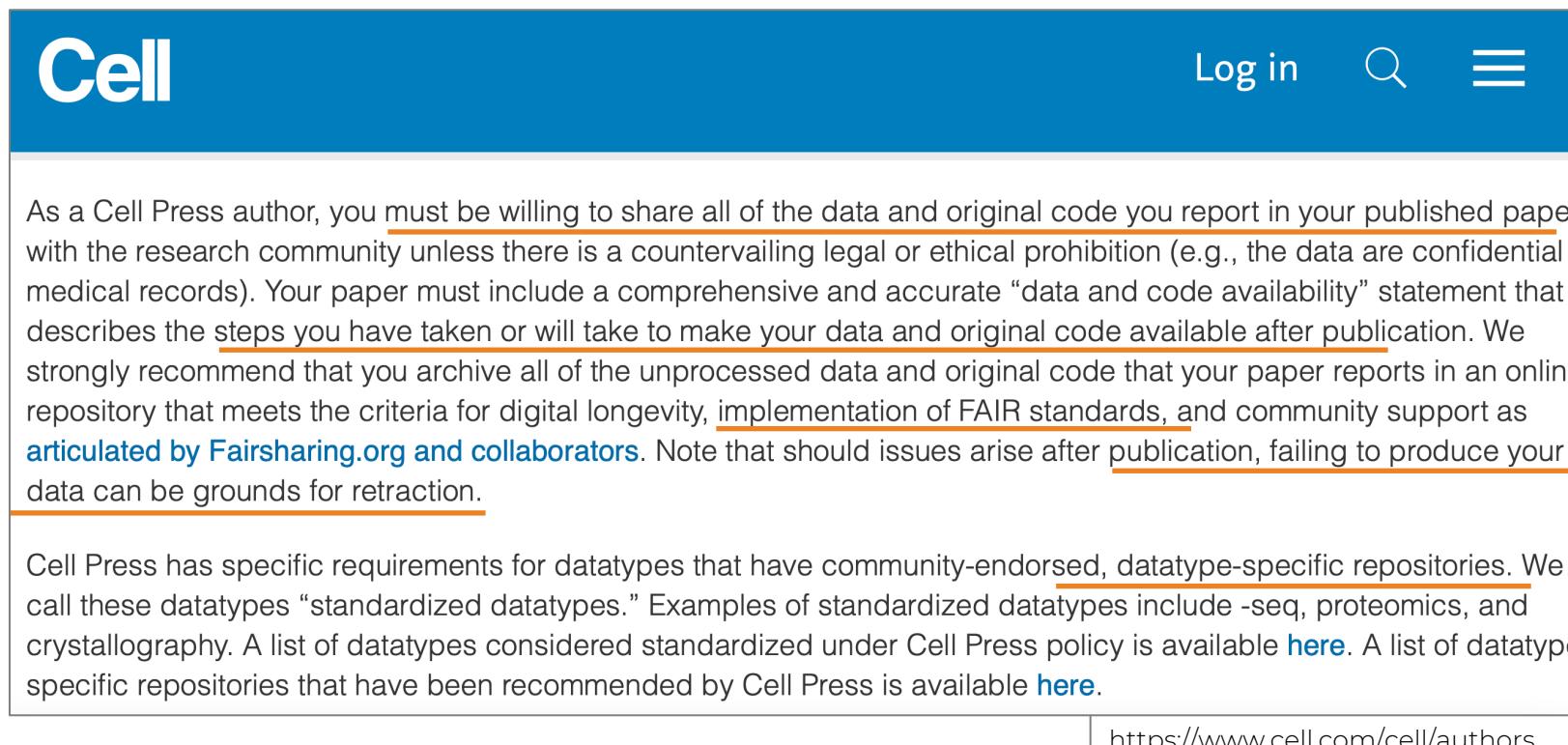
<https://www.cell.com/star-authors-guide>



# Journals request FAIR data



The screenshot shows the Science journal homepage. At the top right, there are navigation links: Current Issue, First release papers, Archive, About (with a dropdown arrow), and a button labeled "Submit manuscript". Below the header, a bullet point in a box states: "• Specification of where all data underlying the study are available, or will be deposited, and whether there are any restrictions on data availability such as an MTA." At the bottom right of the page, a URL is provided: <https://www.science.org/content/page/science-information-authors>.



The screenshot shows the Cell journal homepage. On the left, the word "Cell" is displayed. On the right, there are links for "Log in" and a search icon. Below the header, a large block of text outlines Cell Press's data sharing policy:

As a Cell Press author, you must be willing to share all of the data and original code you report in your published paper with the research community unless there is a countervailing legal or ethical prohibition (e.g., the data are confidential medical records). Your paper must include a comprehensive and accurate “data and code availability” statement that describes the steps you have taken or will take to make your data and original code available after publication. We strongly recommend that you archive all of the unprocessed data and original code that your paper reports in an online repository that meets the criteria for digital longevity, implementation of FAIR standards, and community support as articulated by Fairsharing.org and collaborators. Note that should issues arise after publication, failing to produce your data can be grounds for retraction.

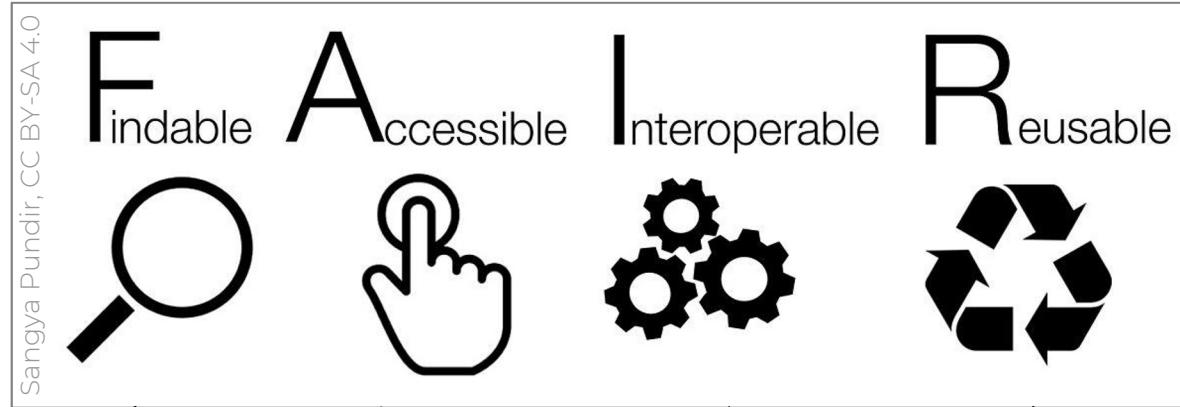
Cell Press has specific requirements for datatypes that have community-endorsed, datatype-specific repositories. We call these datatypes “standardized datatypes.” Examples of standardized datatypes include -seq, proteomics, and crystallography. A list of datatypes considered standardized under Cell Press policy is available [here](#). A list of datatype specific repositories that have been recommended by Cell Press is available [here](#).

At the bottom right of the text block, the word "FAIR?" is written with an arrow pointing towards the "implementation of FAIR standards" text in the original text. At the very bottom right, a URL is provided: <https://www.cell.com/cell/authors>.



# FAIR principles for data and metadata

<https://www.go-fair.org/fair-principles/>



## Metadata

- keywords
- identifiers
- machine-actionable (SEO, etc)

## Access procedure

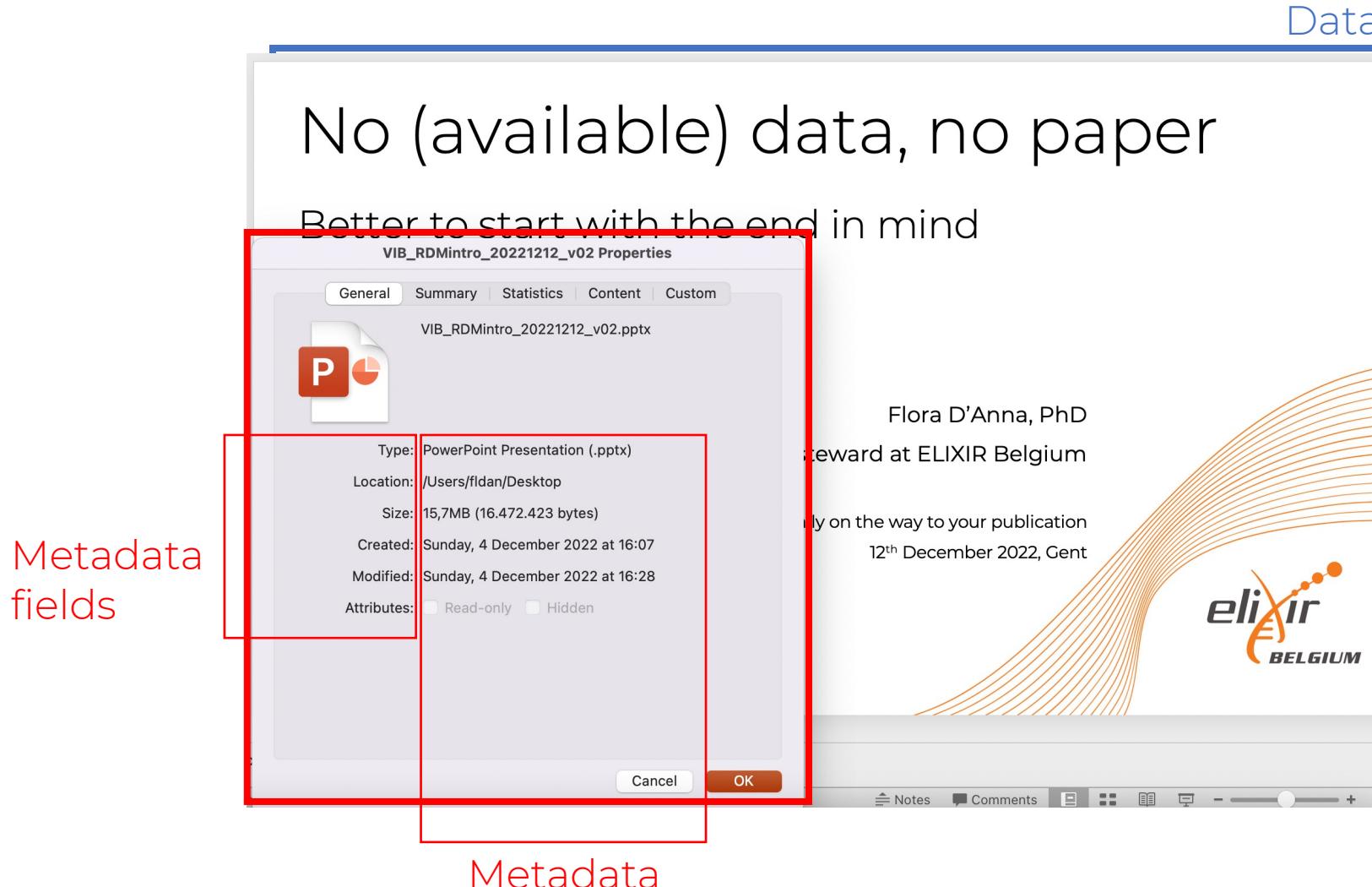
- e.g., DAC
- authentication, identification
- ...

- Standards
- Ontologies
- For humans and machines

- Documentation
- Licence

# Metadata is highly structured documentation

Example 1: machine-actionable metadata of a PowerPoint file



# Example 2: metadata of a DNA sequencing experiment

Metadata fields or attributes					Raw Data File * ↕
	Input (library name) *	nucleic acid sequencing * ↕	sequencing instrument * ↕	file type * ↕	file checksum
	<input type="checkbox"/> library 1 x	sequencing protocol	MinION	bam	9840f585055afc37de353706fd31a377 fake3.bam

Metadata      Data

From <https://datahub.test.elixir-belgium.org>

! WARNING:  
Data VS Metadata

- The difference is often dependent on the research question
- What is considered metadata by someone can be used as data by others in a different context

# Example 3: metadata for a study about the most used file type in sequencing

Metadata fields or attributes					
Metadata			Data		
<input type="checkbox"/> Input (library name) *	nucleic acid sequencing * ↕	sequencing instrument * ↕	<input type="checkbox"/> file type * ↕	file checksum	Raw Data File * ↕
<input type="checkbox"/> library 1 x	sequencing protocol	MinION	bam	9840f585055afc37de353706fd31a377	fake3.bam

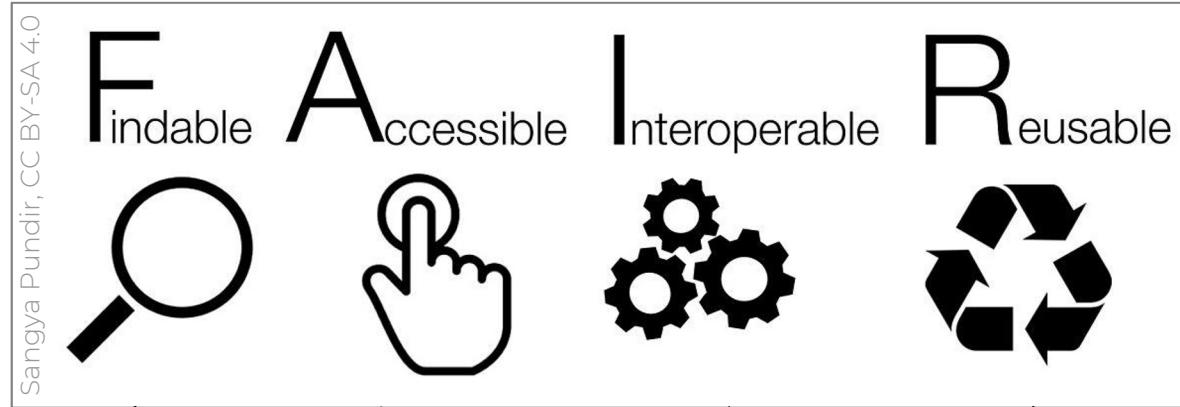
From <https://datahub.test.elixir-belgium.org>

! WARNING:  
Data VS Metadata

- The difference is often dependent on the research question
- What is considered metadata by someone can be used as data by others in a different context

# FAIR principles for data and metadata: explained

<https://www.go-fair.org/fair-principles/>



## Metadata

- keywords
- identifiers
- machine-actionable (SEO, etc)

## Access procedure

- e.g., DAC
- authentication, identification
- ...

- Standards
- Ontologies
- For humans and machines

- Documentation
- Licence

# Findable

Your (meta)data can be discovered by others



## Where?

Searchable resources such as

- search engines
- databases or repositories
- systems for shared storage, etc
- registries of metadata

## When?

- During the project
- After the end of the project

## By whom?

- Collaborators, colleagues
- Reviewers
- Other scientists, society at large
- Registries, web crawler

## How?

- Machine-actionable\* metadata (SEO, etc)
- Machine-actionable keywords
- Unique and persistent identifiers

\*Machine-actionable: following standards for machines

# Sharing data – Repositories

Sessions:

- “Data publication 101” session



Repositories or deposition databases

- Machine-actionable metadata (SEO, etc)
- Machine-actionable keywords
- Unique and persistent identifiers

# Sharing data – Appropriate restrictions

Sessions:

- “Ethical and legal constraints on the sharing of personal data”
- “A closer look at the repositories world”



Ethical and legal constrains



Repositories with restricted access

# Accessible

Your (meta)data can be made available to others

Where?

- Databases or repositories
- Systems for shared storage, etc
- Registries for metadata

When?

- During the project
- After the end of the project

To whom?

People and/or machines with the right access permission or authorization

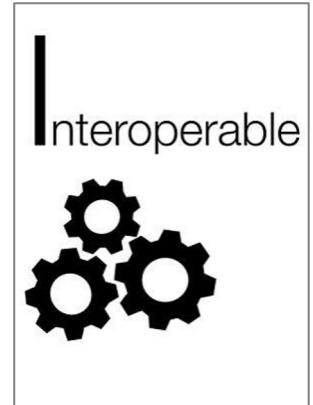
How?

- Open, free, universally implementable standard protocol to retrieve the (meta)data (e.g. via a “link”)
- It could include authentication and/or authorization protocol
- Data Access Committee (DAC), etc

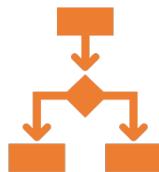


# Interoperable

Humans and machines can exchange, interpret and integrate each other's (meta)data



Non-proprietary file formats that are data exchange formats, for structured data (e.g. JSON, XML, CSV, TSV, FASTQ, FASTA etc.).



(Meta)Data schema defining the relations, such as hierarchy, of the elements that constitute the (meta)data model or structure.



Controlled vocabularies or ontologies to convey unambiguous meaning or semantics (e.g. EFO, OBI, Gene Ontology).

# Interoperable: file formats for exchange of structured data

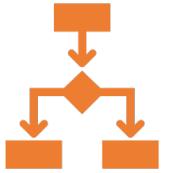


<https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-11728/samples/>

Display summary							Export table in Tab-delimited format
Source Name	Characteristics[organism]	Characteristics[cell line]	Characteristics[sex]	Characteristics[age]	Unit[time unit]	Characteristics[developmental st]	CSV, TSV
24 rows							
CRL-7815_Control_1	Homo sapiens	Hs 483.Sk cell	male	83	year	adult	
CRL-7815_Control_1	Homo sapiens	Hs 483.Sk cell	male	83	year	adult	
CRL-7815_Control_2	Homo sapiens	Hs 483.Sk cell	male	83	year	adult	
CRL-7815_Control_2	Homo sapiens	Hs 483.Sk cell	male	83	year	adult	
CRL-7815_RB_1	Homo sapiens	Hs 483.Sk cell	male	83	year	adult	
CRL-7815_RB_1	Homo sapiens	Hs 483.Sk cell	male	83	year	adult	
CRL-7815_RB_2	Homo sapiens	Hs 483.Sk cell	male	83	year	adult	
CRL-7815_RB_2	Homo sapiens	Hs 483.Sk cell	male	83	year	adult	
HS-27_Control_1	Homo sapiens	Hs27 cell	male	1	month	infant	

- Most machines (computers, software) read and open CSV or TSV by “arranging” the information in columns and rows
- Easy to readable for humans as well

# Interoperable: hierarchically structured metadata model



```
{ isa_seek-investigation_p19.json • 
Users > fldan > Nextcloud > Flora > ELIXIR > BH2022 > {} isa_seek-investigation_p19.json > {} investigation > [ ] studies > {} 0 > {} materials > [ ] samples
410
411
412
413     "samples": [
414         {
415             "@id": "#sample/331",
416             "name": "leaf 1",
417             "derivesFrom": [
418                 {
419                     "@id": "#source/330"
420                 }
421             ],
422             "characteristics": [],
423             "factorValues": [
424             ]
425         }
426     },
427     "protocols": [
428         {
429             "@id": "#protocol/18_10",
430             "name": "sample collection",
431             "protocolType": {
432                 "annotationValue": "sample collection",
433                 "termAccession": "",
434                 "termSource": ""
435             },
436             "description": "",
437             "uri": "",
438             "version": ""
439         }
440     ]
441 }
```

## ISA-JSON

- Most machines (computers, software) read and open JSON
  - key : value
- Not easy to read for humans





# Interoperability: controlled vocabularies or ontologies to convey unambiguous meaning or semantics

The screenshot shows the OLS (Ontology Search) homepage. The logo 'OLS' is at the top left, followed by 'ONTOLOGY SEARCH'. Below the logo are four navigation links: 'Home', 'Ontologies', 'Documentation', and 'About'. The main content area features a teal background with a network graph of nodes and connections. The title 'Gene Ontology' is displayed prominently. A sub-section below it states: 'The Gene Ontology (GO) provides a framework and set of concepts for describing the functions of gene products from all organisms.' There is a search bar labeled 'Search GO' with a 'Search' button, and three buttons below it: 'Terms', 'Download', and 'Request a Term'.

This screenshot shows the search results for 'mouse' in the Gene Ontology. On the left, there's a sidebar with 'Browse Terms' and 'Browse Properties' buttons. The main content area has a box titled 'Ontology term:' containing '- definition' and '- identifier'. Below this box, the text 'NCBI Taxonomy for mouse:' and 'Taxonomy ID: 10090' is displayed next to icons of a mouse with a green checkmark and a red X.

This screenshot shows the detailed ontology information for the Gene Ontology. It includes the 'Ontology IRI' (http://purl.obolibrary.org/obo/go.owl), 'Version IRI' (http://purl.obolibrary.org/obo/go/releases/2022-07-01/go.owl), 'Ontology ID' (go), and 'Version' (2022-07-01).

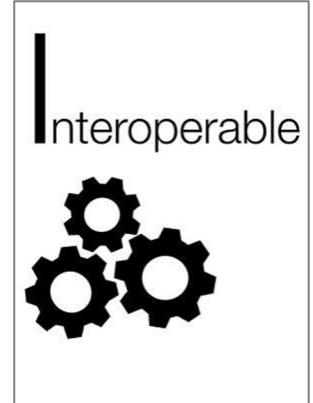
More info at:

- [https://rdmkit.elixir-europe.org/metadata\\_management#how-do-you-find-appropriate-vocabularies-or-ontologies](https://rdmkit.elixir-europe.org/metadata_management#how-do-you-find-appropriate-vocabularies-or-ontologies)
- <https://faircookbook.elixir-europe.org/content/recipes/interoperability/introduction-terminologies-ontologies.html>

# Interoperability: standards and standardization for humans and machines

Sessions:

- “Organising and standardising research data that underpin your publication”
- “Make writing easier: Document & describe your data”



Use of existing international or domain-specific standards



ISO STANDARDS for dates:  
20220715 or 2022-07-15

**DCMI Schemas**

The Dublin Core™ Metadata Initiative provides access to schemas defining DCMI term declarations represented in various schema languages. Schemas are machine-processable specifications which define the structure and syntax of metadata specifications in a formal schema language.

**DublinCore**



- [XMLS Schemas](#)
- [RDFS Schemas](#)



# Reusable

Your (meta)data can be reused by others



Data usage licence



Documentation and metadata (e.g., data provenance, etc)

README.txt

# Documentation and metadata

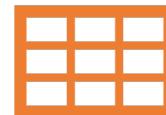
Sessions:

- “Organising and standardising research data that underpin your publication”
- “Make writing easier: Document & describe your data”



README.txt

Documentation is (unstructured) information about the data.



Metadata.json

Metadata is highly structured documentation.

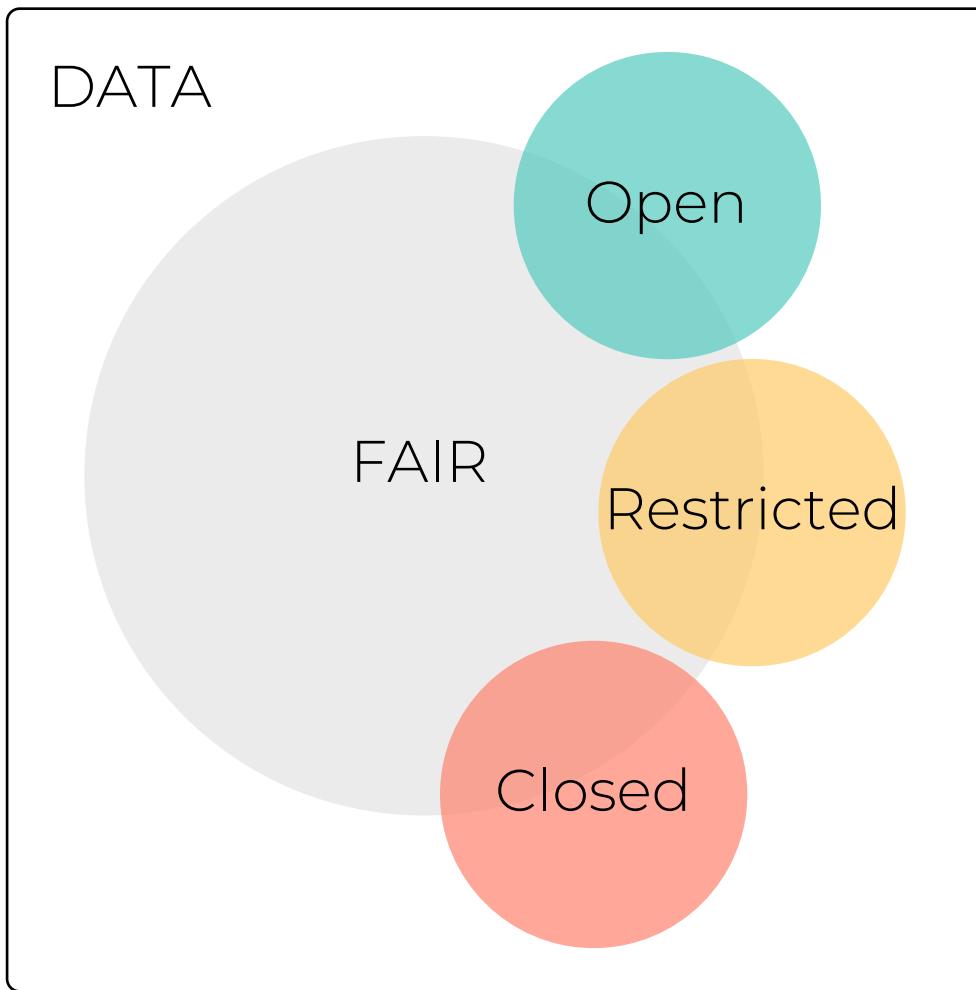
# Data reuse

Sessions:

- “Reusing data”
  - Where to find data to reuse
  - Permission to reuse for specific purpose
  - Data quality
  - How to handle versioning
  - How to cite
  - ...



# Open vs. FAIR data: Open data is different from FAIR data



What are the benefits of good RDM?

Let's write possible reasons

*use pens and papers*



# Benefits of good RDM (or FAIR data)

## During the project

- Reduce the risk of data loss
- Less time (and money) spent
  - looking for unfindable data
  - redoing not documented experiments/analysis
- Increase efficiency and smooth collaboration in a team (easier and faster way for sharing information)
- More control over data access (who, when, how, etc.)
- More clear documentation for people onboarding the team and by people leave the team
- Easier data reuse (samples, raw data, scripts, analysis etc.)



# Benefits of good RDM (or FAIR data)

After the project

- Results are reproducible by others
- Transparent and trustworthy data and science (and paper) among your peers
- Higher impact in scientific community and possibly more citations
- Integration of several data sources from different disciplines could lead to new discoveries
- New discoveries at a lower cost by reusing high quality existing data
- Less duplication of work



# How to approach RDM starting with “the end” (FAIR metadata) in mind

1. Choose the right repositories for your data
2. Learn the requirements of the selected repositories
3. Start writing and applying your DMP document



# How to choose the right repositories for your data

Sessions:

- “Ethical and legal constraints on the sharing of personal data”
- “A closer look at the repositories world”

1. Ethical and  
legal issues



2. Data availability  
requirements



3. Consultation  
with experts



4. Access restrictions  
and/or reuse limitations



5. Appropriate  
repositories



# Learn the requirements of the selected repositories

Sessions:

- “Data publication 101” session
- “Organising and standardising research data that underpin your publication”
- “Make writing easier: Document & describe your data”
- Reusing data

1. Data policies



2. File formats



3. Costs



4. Documentation & metadata requirements



# Write and implement your DMP

Sessions:

- “Planning for efficiency” session

What's “data” in DMP ? How are you going to manage it during and after the project?



Digital or digitalised information



Physical materials



Software, other research output

# Useful resources for RDM best practices and guidelines

- [RDMkit](#): the research data management toolkit for life sciences by ELIXIR (<https://elixir-europe.org>)



- ELIXIR research data management guidelines and best practices (<https://elixir-europe.org/what-we-offer/guidelines> )

A screenshot of the RDMkit interface. At the top, there's a navigation bar with the text "Research data management". Below it are four main sections: "Overview of good data management practices" (with a "RDMkit" icon), "Step-by-step instructions" (with a "FAIR cookbook" icon), "FAIR cookbook" (with a "DSW" icon), and "Data management plan wizard" (with a "DSW" icon). Each section contains a brief description of its purpose.

- ELIXIR Belgium services for RDM ([https://www.elixir-belgium.org/research\\_data\\_management](https://www.elixir-belgium.org/research_data_management))



- [RDM Guide](#) by ELIXIR Belgium



# RDMkit: research data management best practices and guidelines for life sciences



The Research Data Management toolkit for Life Sciences  
Best practices and guidelines to help you make your data FAIR (Findable, Accessible, Interoperable and Reusable)

**What can we help you find?**

Search RDMkit

**Browse all topics by**

- Data life cycle**  
Start here to get an overview of research data management based on stages in the data life cycle.
- Your role**  
Identify your role in research data management, find data management resources relevant for you, and information to help you progress in your career path.
- Your domain**  
Learn about data management tasks that affect your domain or research community, and the solutions adopted to address them.
- Your tasks**  
Find guidelines and solutions for tackling common data management tasks.
- Tool assembly**  
Find concrete combinations of tools and resources assembled into an ecosystem for research data management.
- National resources**  
Find pointers to country specific information resources and national research data management practices.

<https://rdmkit.elixir-europe.org>



# ELIXIR research data management guidelines and best practices

The screenshot shows the ELIXIR website's 'WHAT WE OFFER' page. The main navigation bar includes links for ABOUT US, WHAT WE OFFER, HOW WE WORK, EVENTS, NEWS, and INTRANET. Below the navigation is a breadcrumb trail: Home > What we offer >. The main heading is 'Guidelines and best practices'. A sub-section title 'Research data management' is shown with an orange icon. Three cards provide details on RDMkit, FAIR cookbook, and DSW.

**WHAT WE OFFER**

Guidelines

Web portals

Services

Partnerships with Industry and SMEs

Opportunities to work together

For ELIXIR members

## Guidelines and best practices

### Research data management

#### Overview of good data management practices

**RDMkit**

The Research Data Management Kit (RDMkit) guides you through the whole data management life cycle and includes advice specific to your domain, your role and your country.

#### Step-by-step instructions

**FAIR cookbook**

The FAIR Cookbook contains step-by-step recipes to accomplish specific data management tasks and to make your data FAIR (Findable, Accessible, Interoperable, Reusable).

#### Data management plan wizard

**DSW**

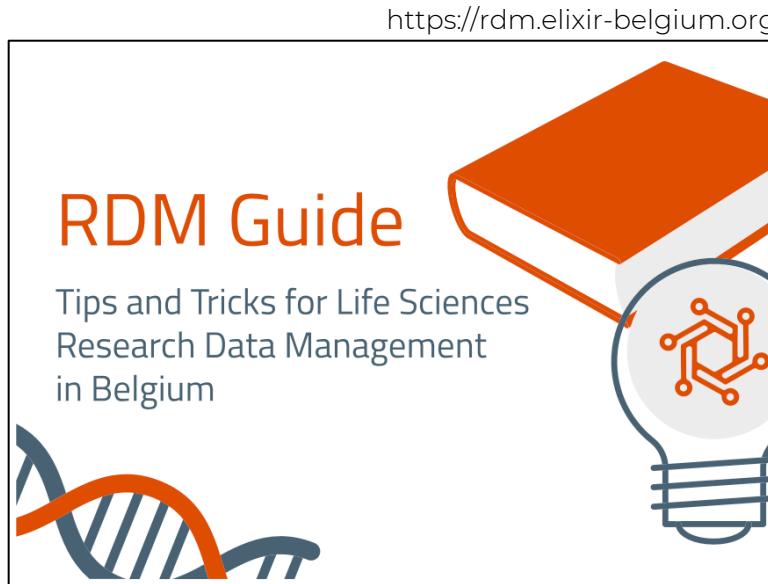
The Data Stewardship Wizard (DSW) is an online tool that guides researchers and data stewards through their data management planning.

<https://elixir-europe.org/what-we-offer/guidelines>



# ELIXIR Belgium services for RDM – RDM Guide

<https://www.elixir-belgium.org/services#term-25>



The screenshot shows the "Research Data Management" section of the ELIXIR Belgium website. At the top, there is a navigation bar with links for HOME, ABOUT, SERVICES (highlighted in orange), PROJECTS, TRAINING, EVENTS, and NEWS, along with a search bar. Below the navigation, there are several tabs: Data Analysis Platforms, Human Data, Plant & Biodiversity, Functional Genomics, and Research Data Management (which is the active tab). A main text block explains the importance of FAIR data management for innovation and societal challenges, mentioning guidelines, tools, and resources for effective data management at all phases of research projects. It also notes the collaboration with data management experts from partners in Belgium and a broad network from ELIXIR Europe. Below the text, there are eight service cards, each with a circular background graphic and a "MOOS SERVICE" badge. The services listed are RO-Crate, WorkflowHub, RDMkit, RDM Guide, ENA Data Submission Toolbox, and DATAHUB.

https://rdm.elixir-belgium.org

elixer  
BELGIUM

HOME ABOUT SERVICES PROJECTS TRAINING EVENTS NEWS Search

Data Analysis Platforms Human Data Plant & Biodiversity Functional Genomics **Research Data Management**

Connecting life sciences data sets is key to foster innovation and address major societal challenges. For this to succeed, research data needs to be Findable, Accessible, Interoperable, and Reusable (FAIR), which starts with good data management. Here you can find guidelines, tools and resources for effective data management at all phases of your research project. This material has been created and compiled in collaboration with data management experts from our partners in Belgium and a broad network from ELIXIR Europe.

RO-Crate

WorkflowHub

RDMkit

RDM Guide

ENA European Nucleotide Archive Data Submission Toolbox

DATAHUB

# Conclusions

1. RDM includes all tasks needed to make (meta)data well described, easy to find, secure and reusable
2. Funders and scientific journals require DMP, open and FAIR data
3. Open data is different from FAIR data
4. Benefits of good RDM for the research and the researchers
  - During the project
  - After the project
5. Starting your DMP and your research project with FAIR (meta)data in mind eases data publication
6. Use available resources for RDM in life sciences provided by ELIXIR (Belgium)





# Get in touch with us.



[info@elixir-belgium.org](mailto:info@elixir-belgium.org)



<https://www.elixir-belgium.org>



This work is licensed under the Creative Commons Attribution 4.0 International License.  
To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

