# 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

12th 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

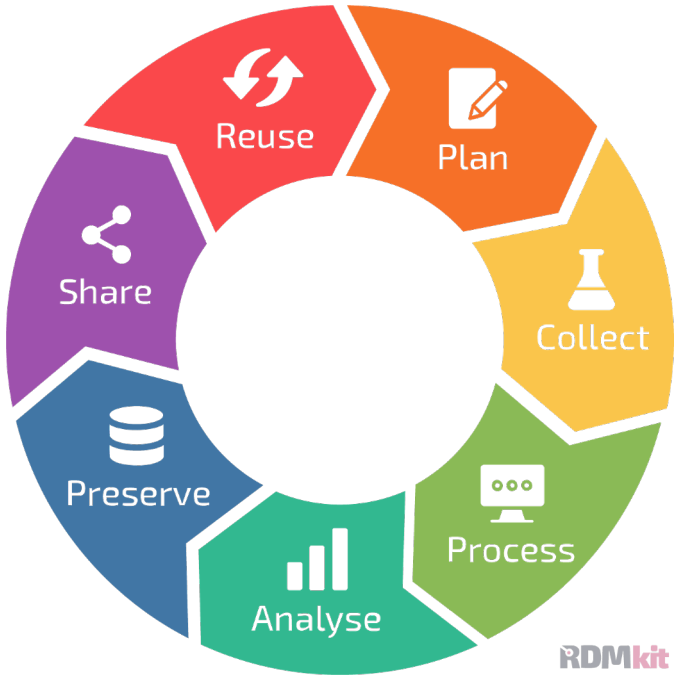
Easy to find

(for humans and machines)

Secure

Reusable

Data life cycle: RDM stages




https://rdmkit.elixir-europe.org

elixir BELGIUM

# Why do we do RDM?

## Let’s write possible reasons

*use pens and papers*

elixir BELGIUM

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

## Before


RDA Plenary Cartoons by Auke Herrema in CC-BY

## After (now)


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

# Funders requiremets:

- 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

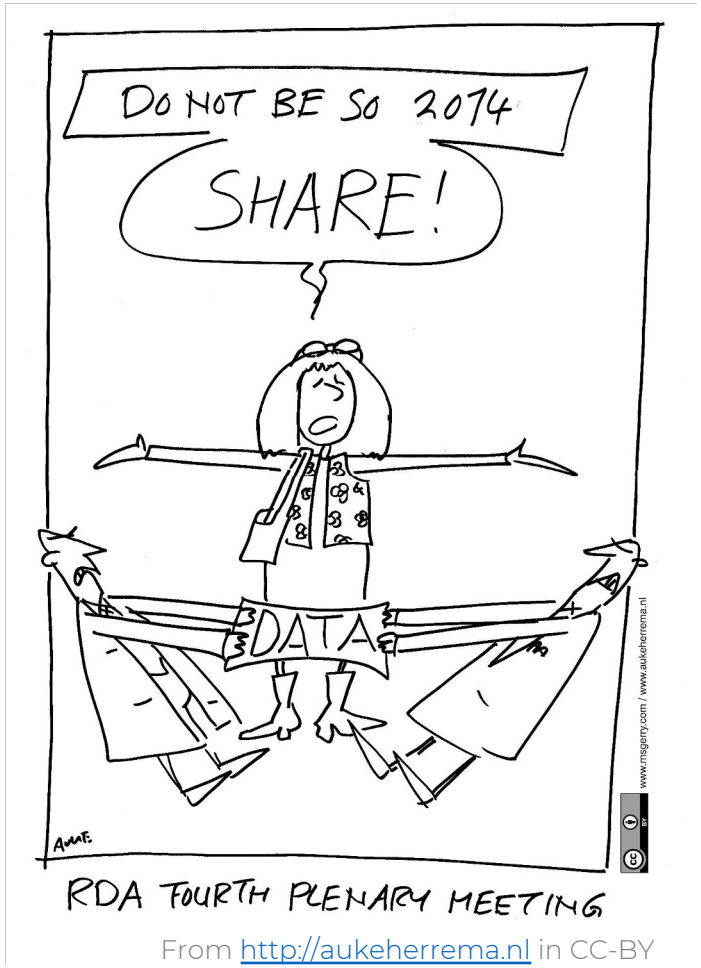
Physical materials

Software, other reseach 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 depositories 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.

https://www.nature.com/nature/for-authors/initial-submission

# Journals: no available data, no paper

CellPress Submit

## Cell Press STAR★Methods

Our STAR Methods format (Structured, Transparent, Accessible Reporting) v
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?

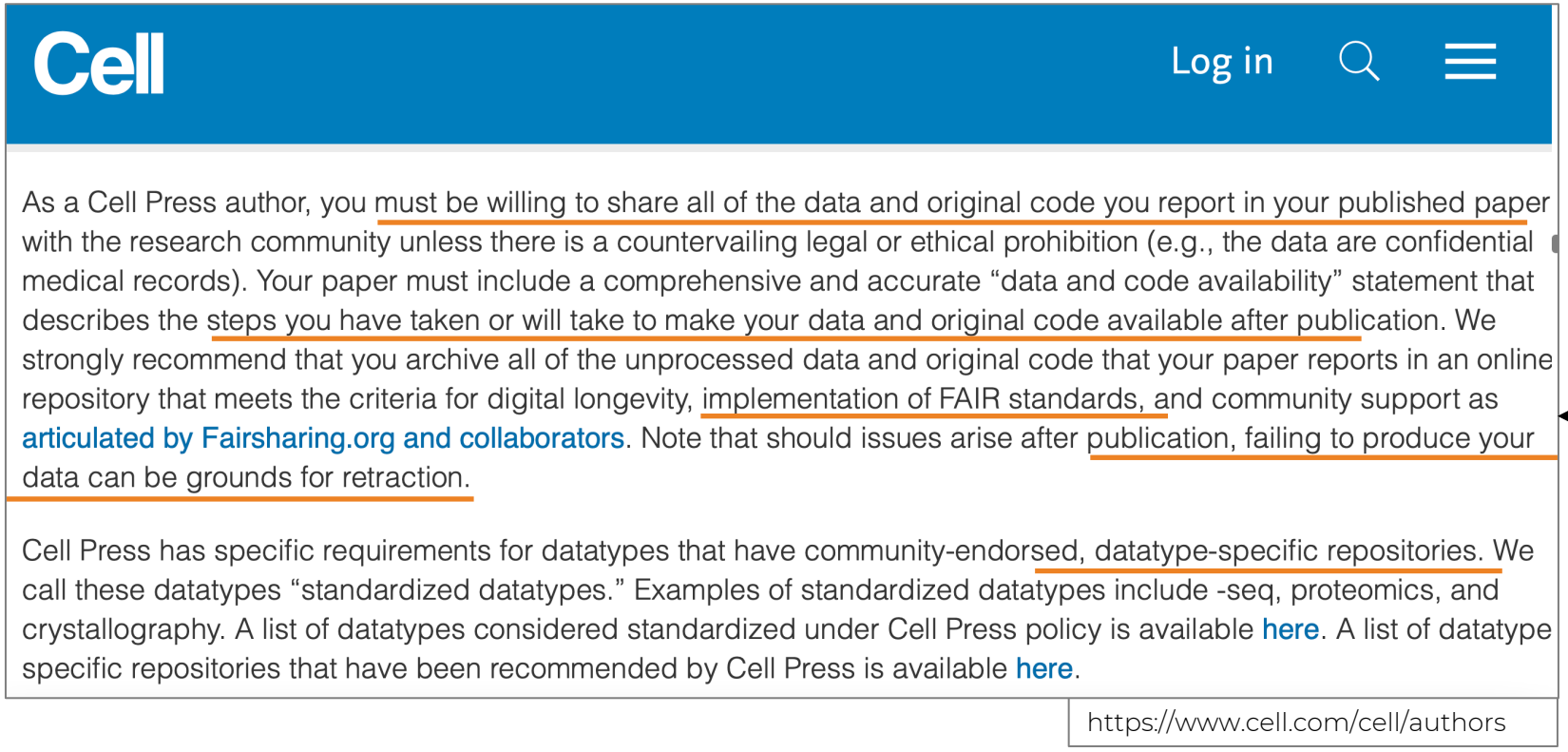
*STAR Protocols* is an open access protocol journal from Cell Press. The primary criteria for publication in *STAR Protocols* are usability and reproducibility. *STAR Protocols* welcomes protocols—from basic to advanced—

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

elixir BELGIUM

# Journals request FAIR data


https://www.science.org/content/page/science-information-authors

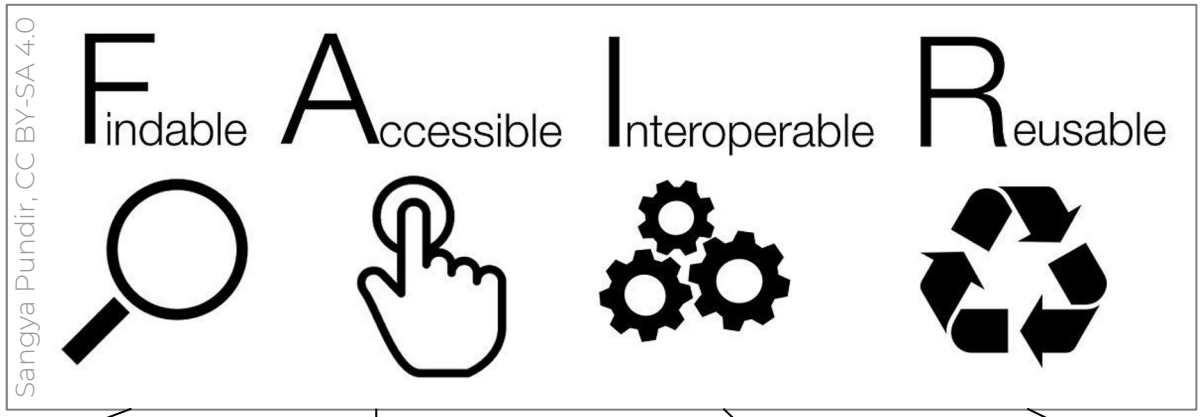



https://www.cell.com/cell/authors

FAIR?

# 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
- …

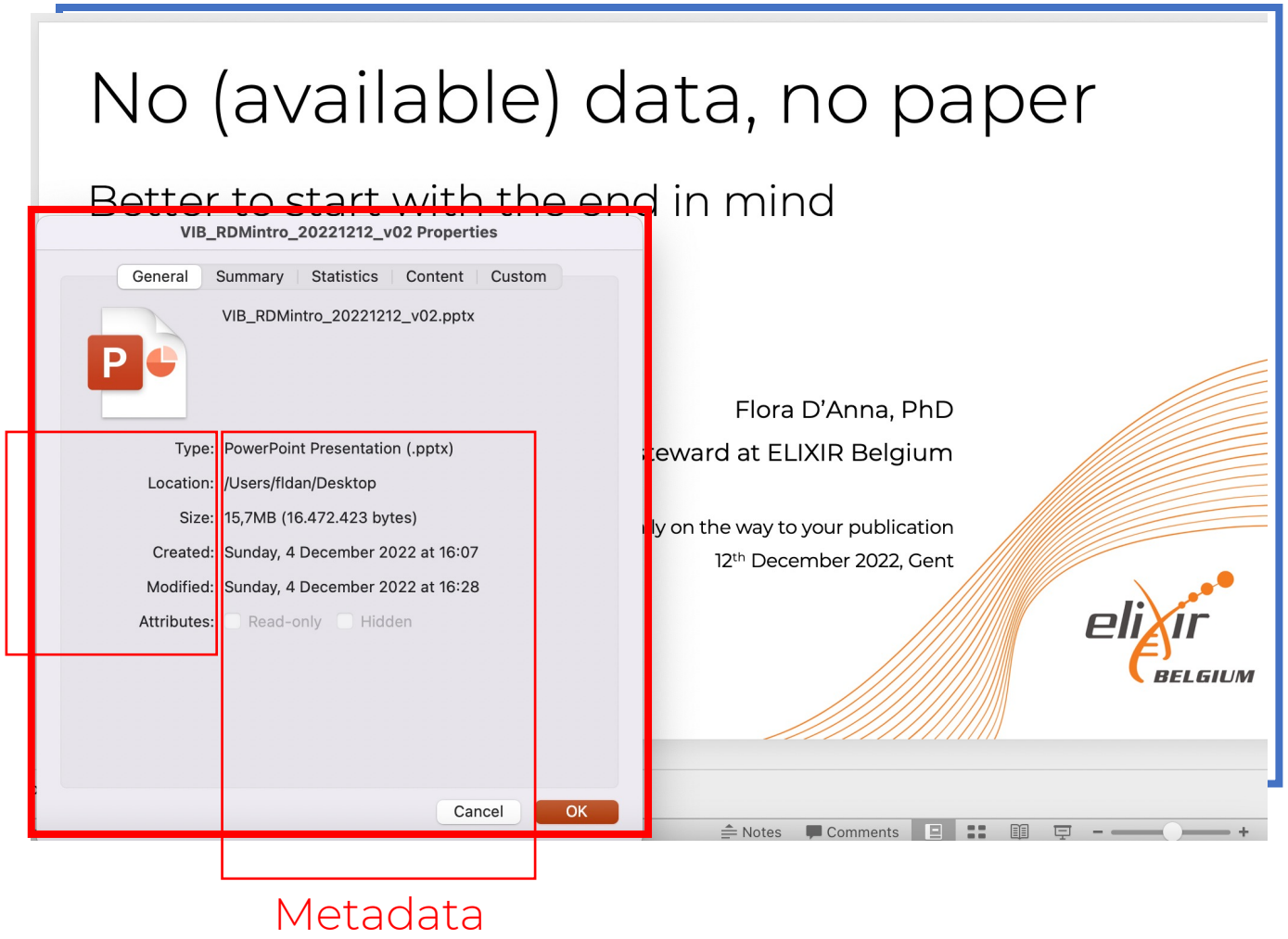
- Standards
- Ontologies
- For humans and machines

- Documentation
- Licence

# Metadata is highly structured documentation

## Example 1: machine-actionable metadata of a PowerPoint file

Data

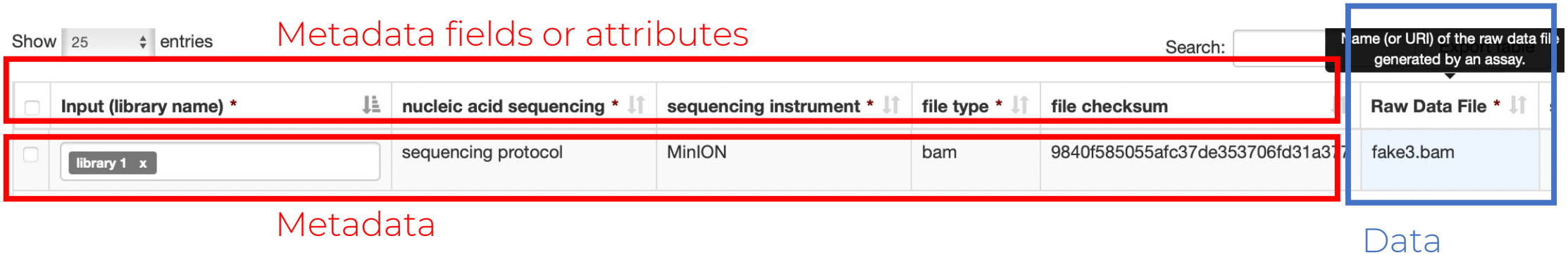



Metadata fields

Metadata

elixir BELGIUM

# Example 2: metadata of a DNA sequencing experiment

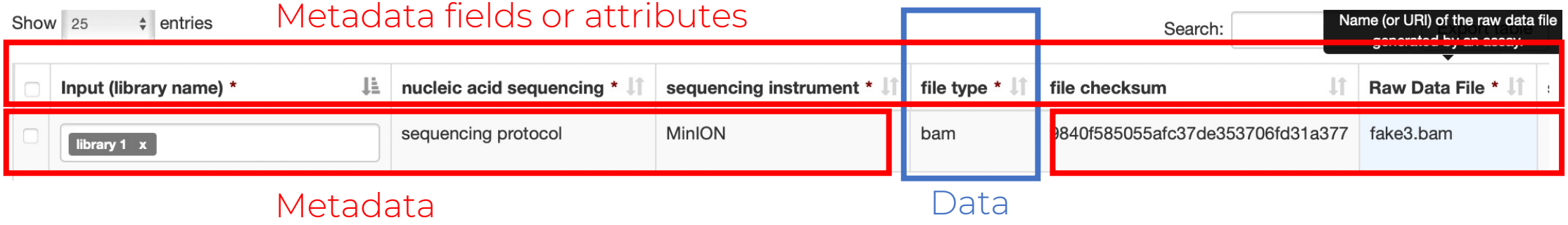



From https://datahub.test.elixir-belgium.org

## ! 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



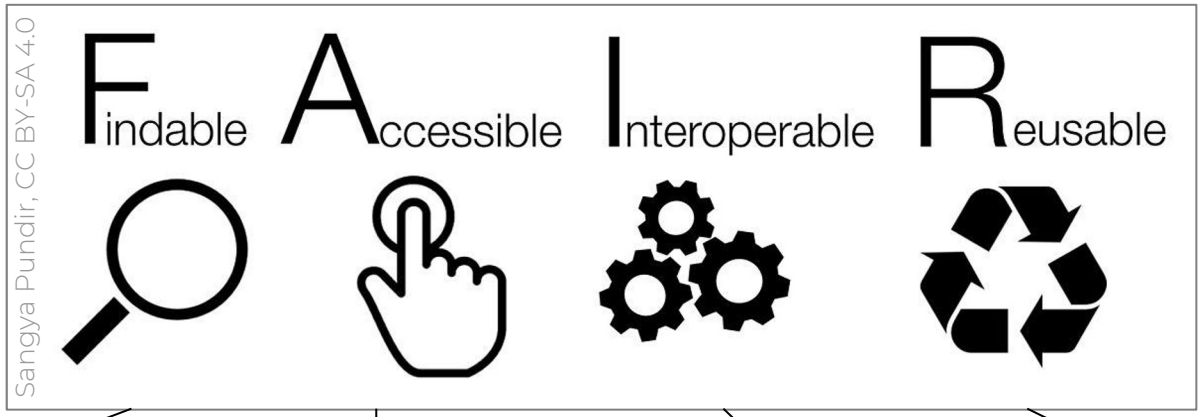

From https://datahub.test.elixir-belgium.org

## ! 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-actionbable* metadata (SEO, etc)
- Machine-actionbable keywords
- Unique and persistent identifiers

*Machine-actionbable: 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?

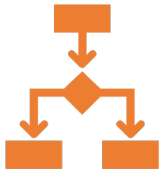
- Open, free, universally implementable standard protocol to retrive the (meta)data (e.g. via a “link”)
- It could include authentication and/or authorization protocol
- Data Access Committee (DAC), etc

elixir BELGIUM

# 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

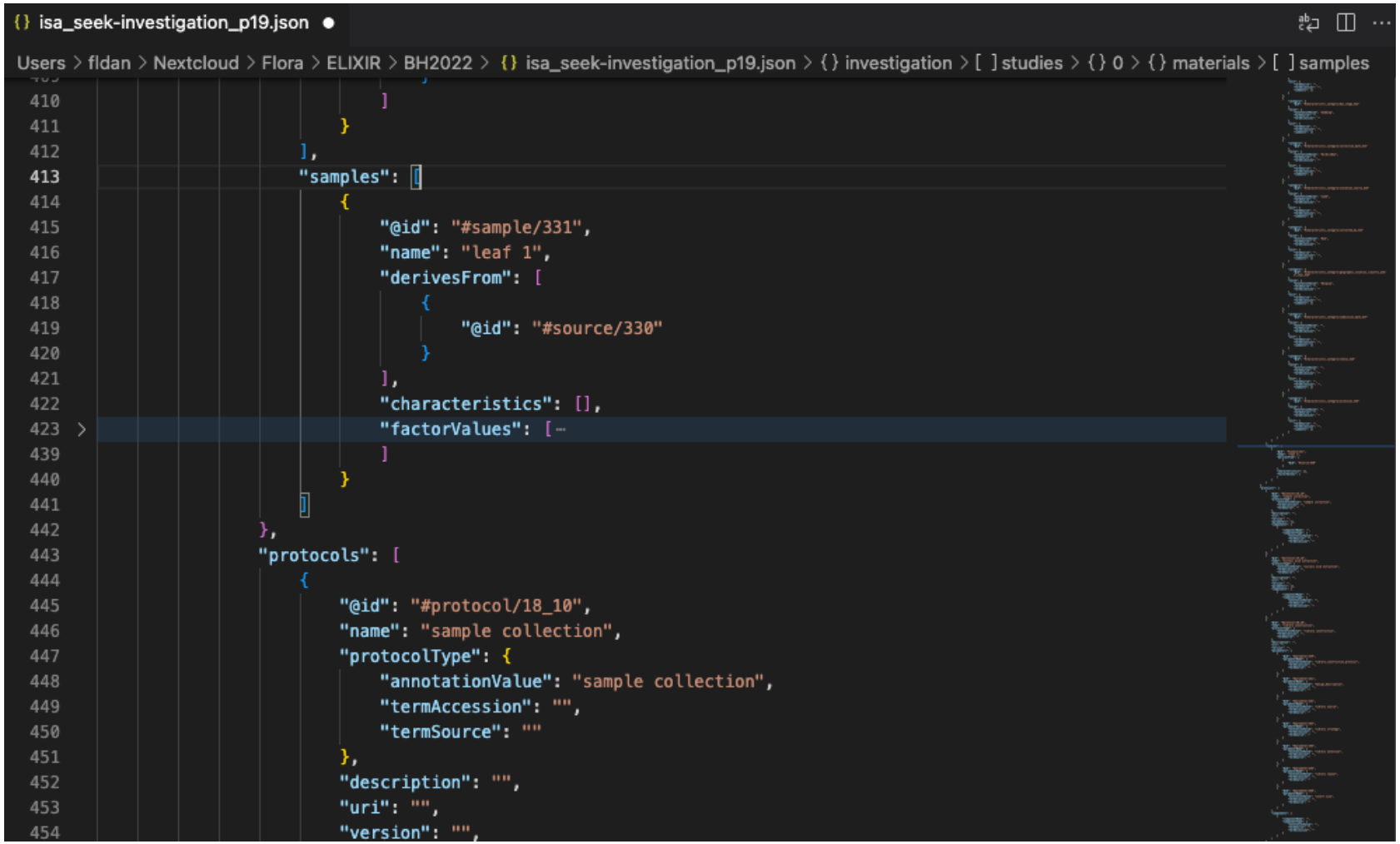
24 rows

| Source Name | Characteristics[organism] | Characteristics[cell line] | Characteristics[sex] | Characteristics[age] | Unit[time unit] | Characteristics[developmental st |
|---|---|---|---|---|---|---|
| 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 |

CSV, TSV

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

# Interoperable: hierarchically structured metadata model

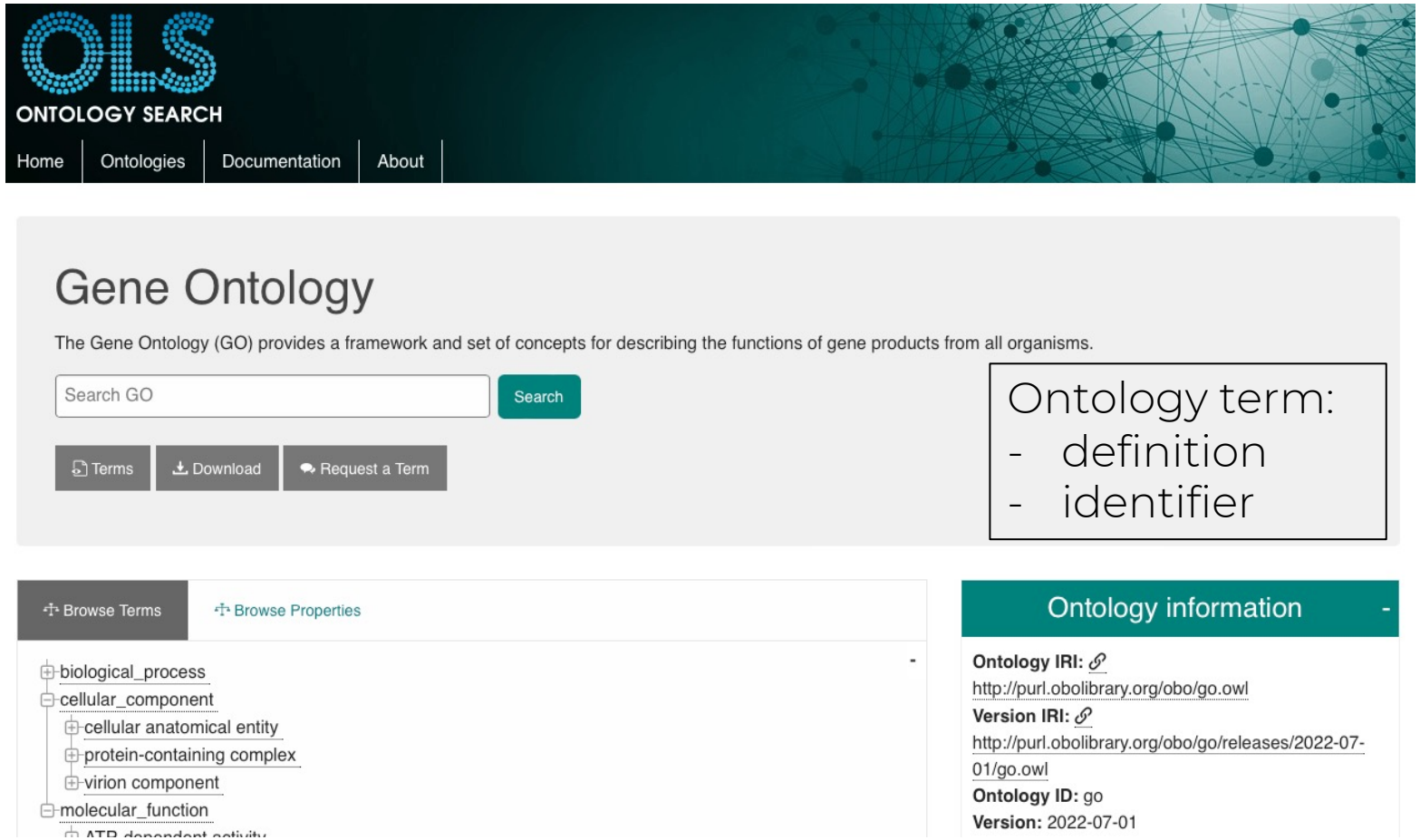



## ISA-JSON

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

Johnson D, et al. ISA API: An open platform for interoperable life science experimental metadata. Gigascience. 2021 Sep 16;10(9):giab060. doi: 10.1093/gigascience/giab060.

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




NCBI Taxonomy for mouse:

Taxonomy ID: 10090

More info at:
- 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.

- 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

elixir BELGIUM

# Documentation and metadata

Sessions:

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

Documentation is (unstructured) information about the data.

README.txt

Metadata is highly structured documentation.

Metadata.json

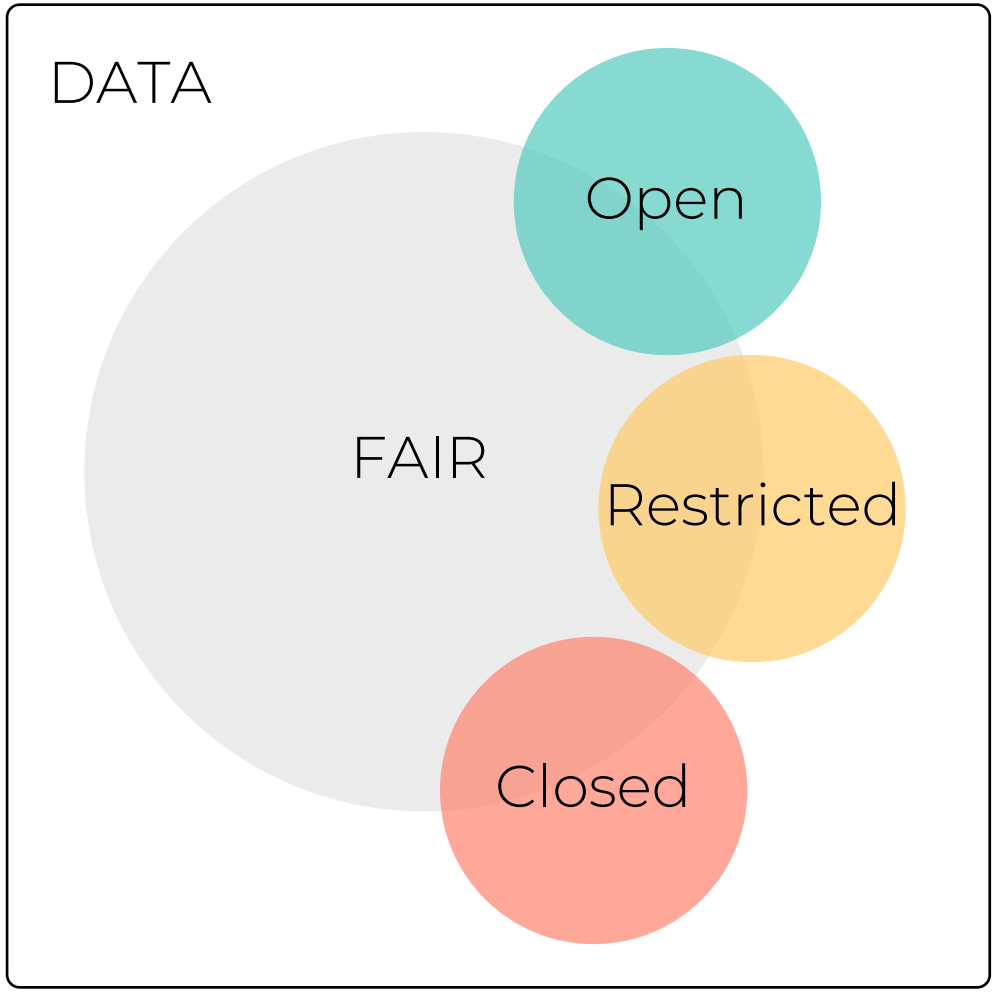
# Data reuse

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

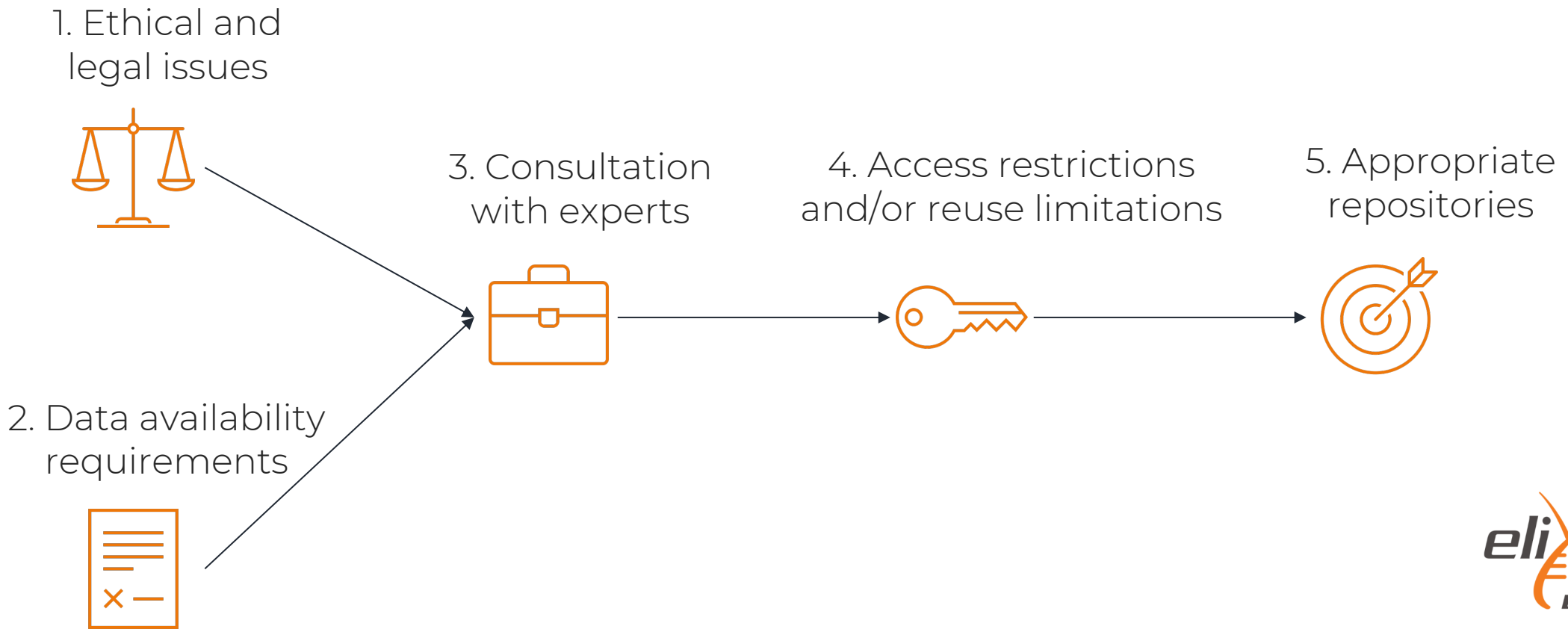
2. Learn the requirements of the selected repositories

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”



# 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

elixir BELGIUM

# 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

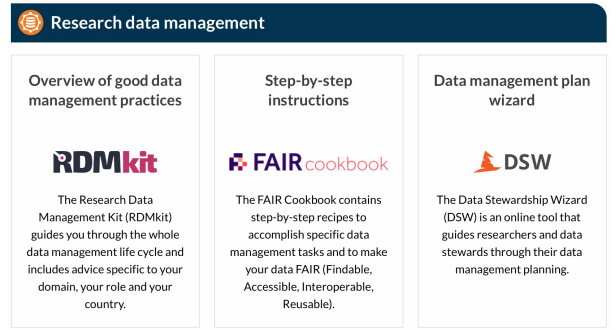
Software, other reseach output

# Useful resources for RDM best practices and guidelines

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

https://rdmkit.elixir-europe.org

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




- ELIXIR Belgium services for RDM (https://www.elixir-belgium.org/research_data_management)

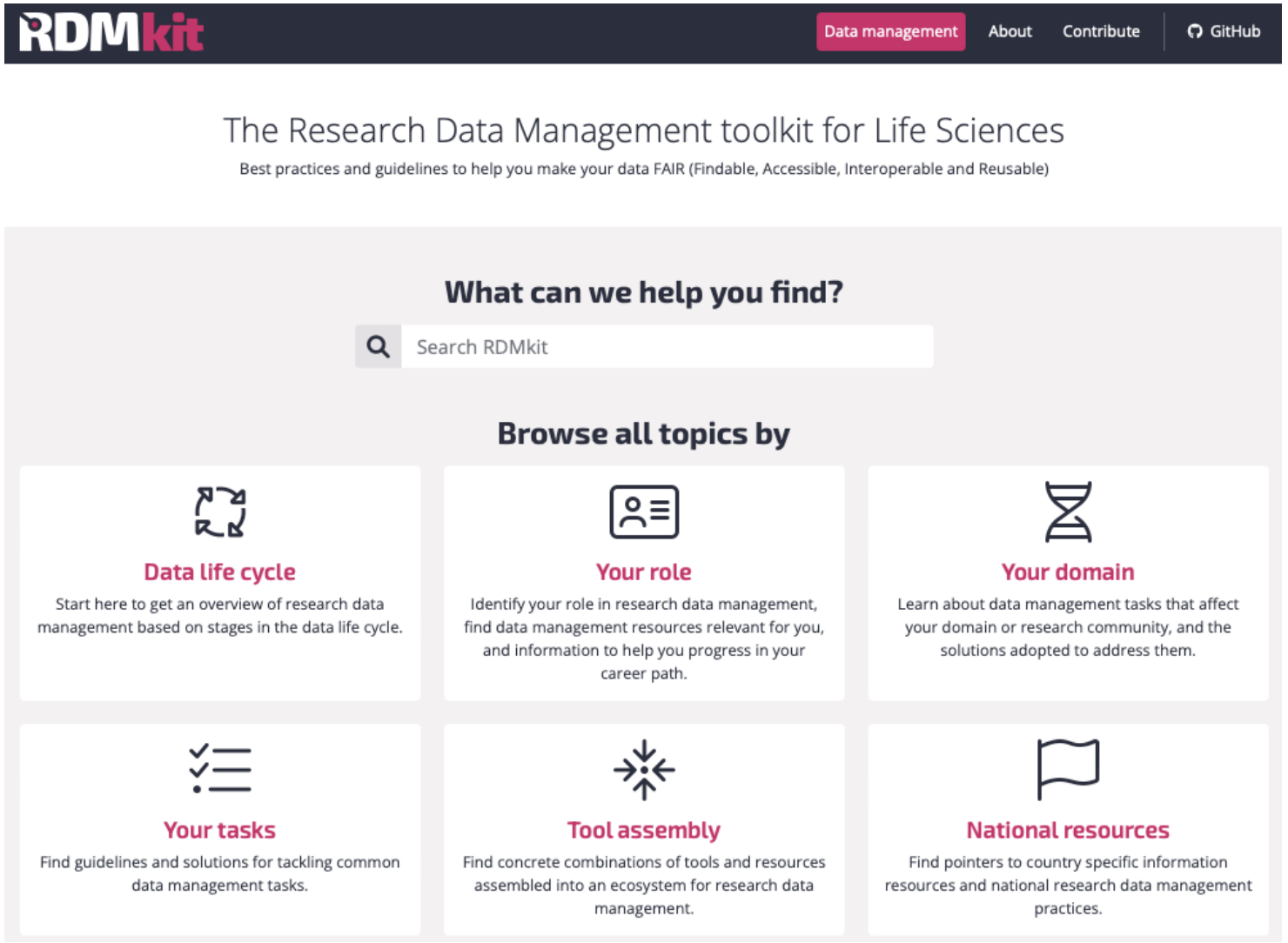
- RDM Guide by ELIXIR Belgium

RDM Guide

https://rdm.elixir-belgium.org

elixir BELGIUM

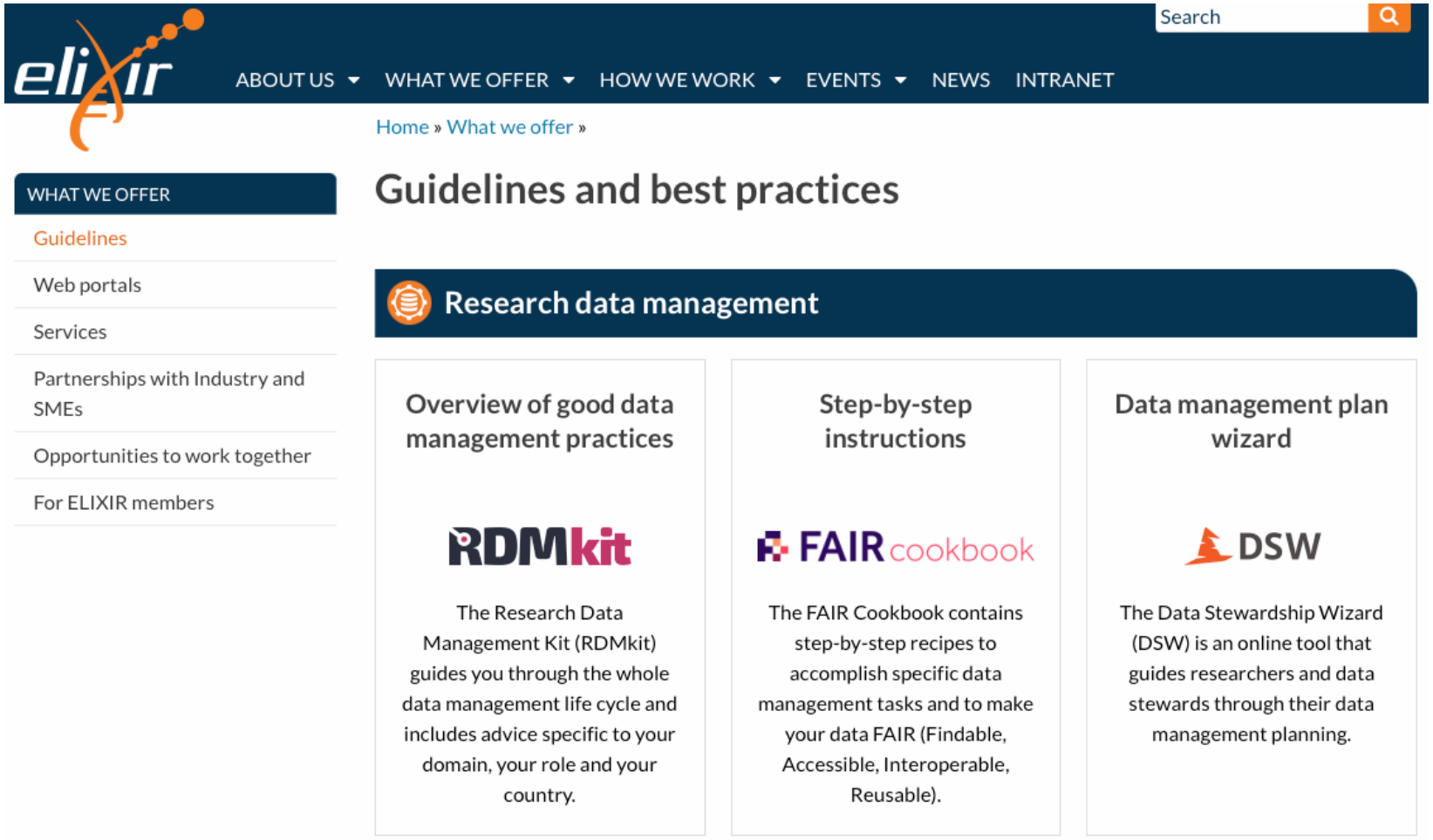
# RDMkit: research data management best practices and guidelines for life sciences




https://rdmkit.elixir-europe.org

elixir BELGIUM

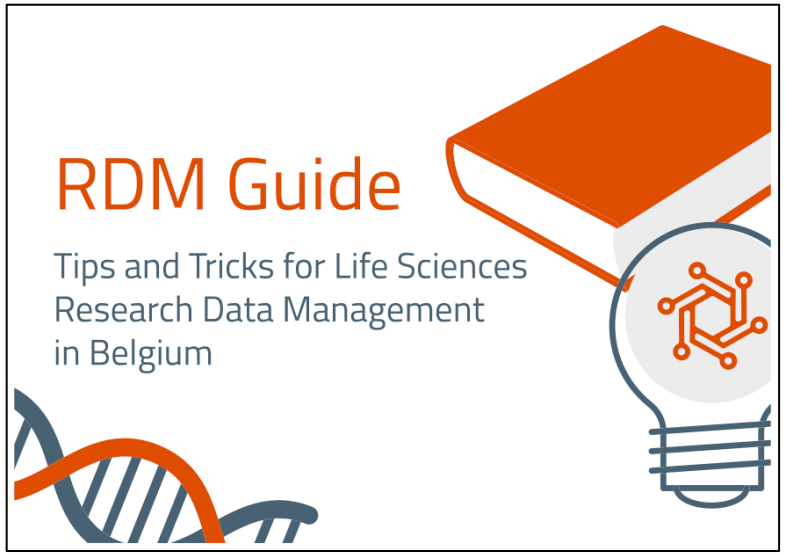
# ELIXIR research data management guidelines and best practices



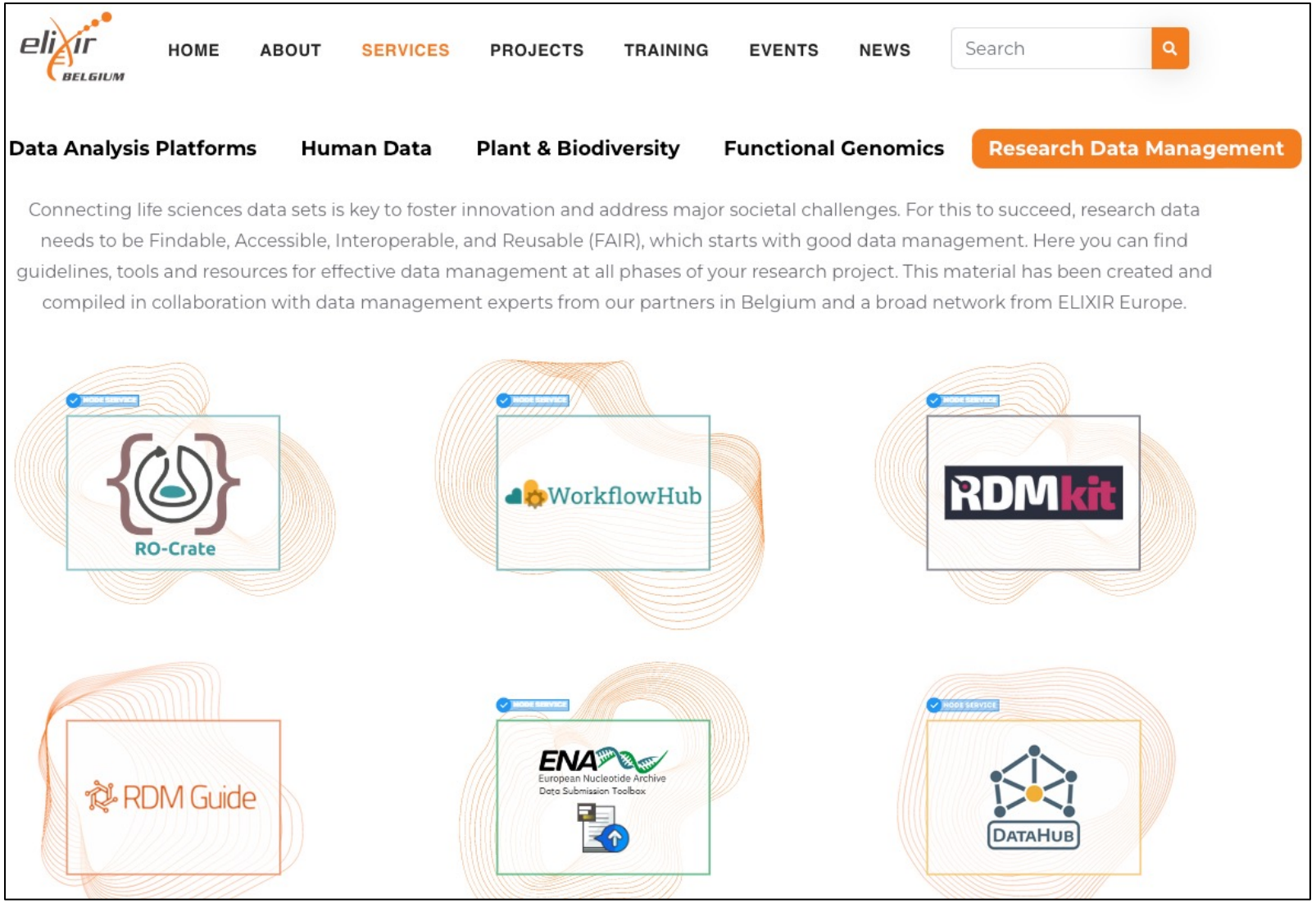

https://elixir-europe.org/what-we-offer/guidelines

# ELIXIR Belgium services for RDM – RDM Guide

https://rdm.elixir-belgium.org




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



# 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

https://www.elixir-belgium.org