Research Data Management - Introduction and basic concepts

Ana Rodrigues (aprodrigues@icnas.uc.pt)
Bruno Direito (bruno.direito@uc.pt)

March 15th 2022

Formações iii UC
ERRAMENTAS DE APOIO À INVESTIGAÇÃO

CIBIT
Coimbra Institute for Biomedical Imaging and Translational Research

UNIVERSIDADE D COIMBRA
INSTITUTO DE CIÊNCIAS NUCLEARES APLICADAS À SAÚDE

Overview

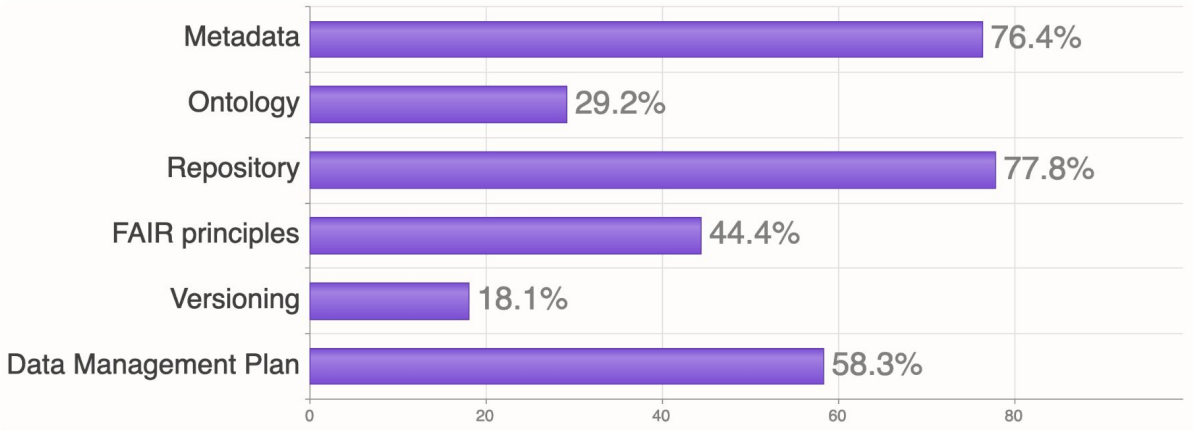
- Data, Information & Knowledge
- Data reuse
- Open Science & FAIR Principles
- Research Data Management
- What should be in a DMP?
- The Science Europe DMP Template

Interact with us!

live.voxvote.com

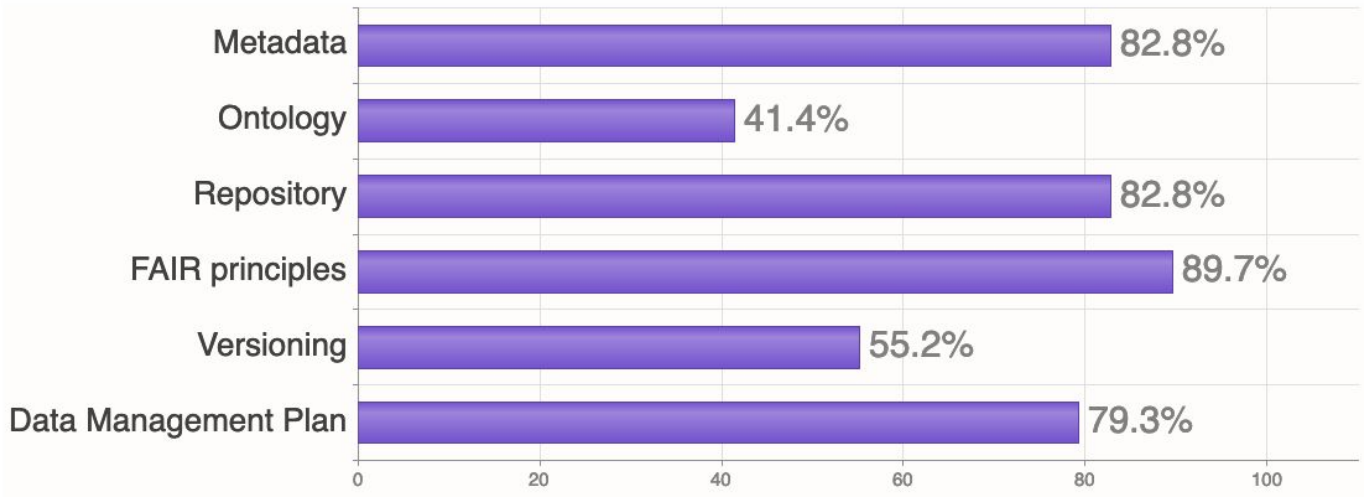
PIN: 189128

1. Have you ever heard about:



72 users voted

7. Do you now feel confortable with these concepts?



29 users voted

Estes slides são adaptados da formação “Ready for Biodata Management” BioData.pt|ELIXIR PT.

Podem ser consultados na sua versão original em *https://github.com/BioData-PT/Ready4BioDataManagement/*

BioData.pt

Data, Information & Knowledge

Learning Outcome 1:

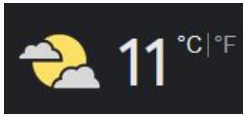
Distinguish between **Data**, **Information** and **Knowledge**

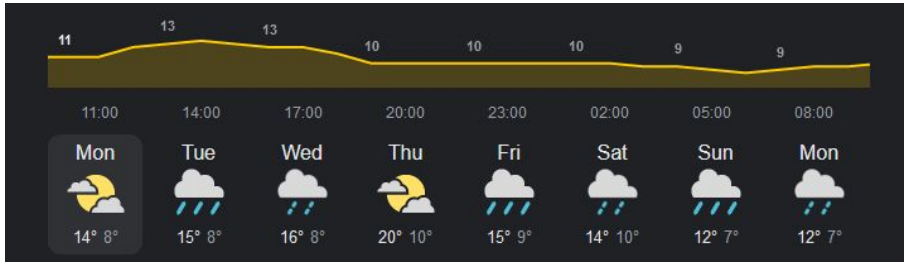
Introduction

- Science is a knowledge discovery paradigm predicated on data acquisition and analysis
- Distinguishing between **data**, **information** and **knowledge** is critical for understanding the need for **research data management**!

Data (plural form)

- Datum (singular): an atomic fact or piece of “information”
  - Temperature
- Dataset: a collection of data that share an object or scope





Information

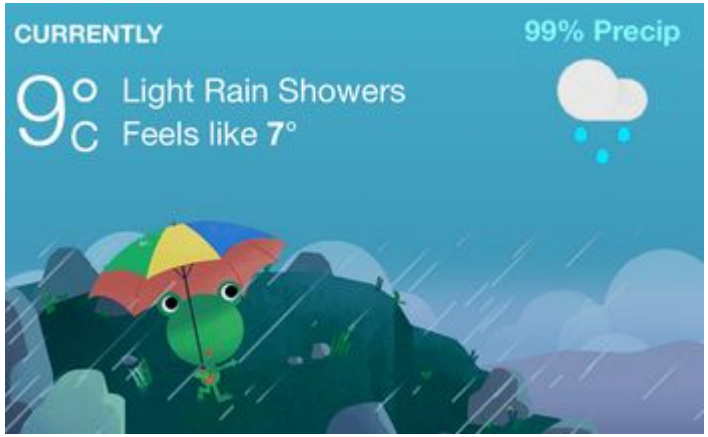
- Information: data + context (metadata)

- Data

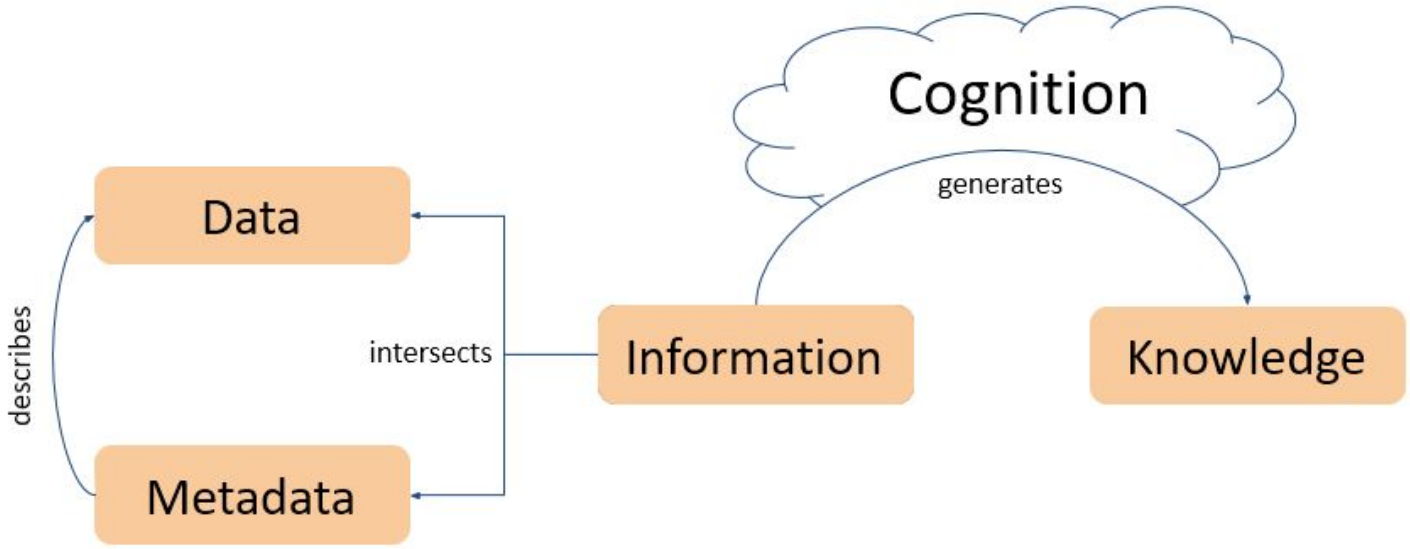
- Metadata (is data about data, providing context):
  - Who produced the data?
  - when?
  - how?
  - why?
  - How can the data be used? (license)
- ! https://www.epa.gov/ceam/metadata-weather-data-management

Knowledge

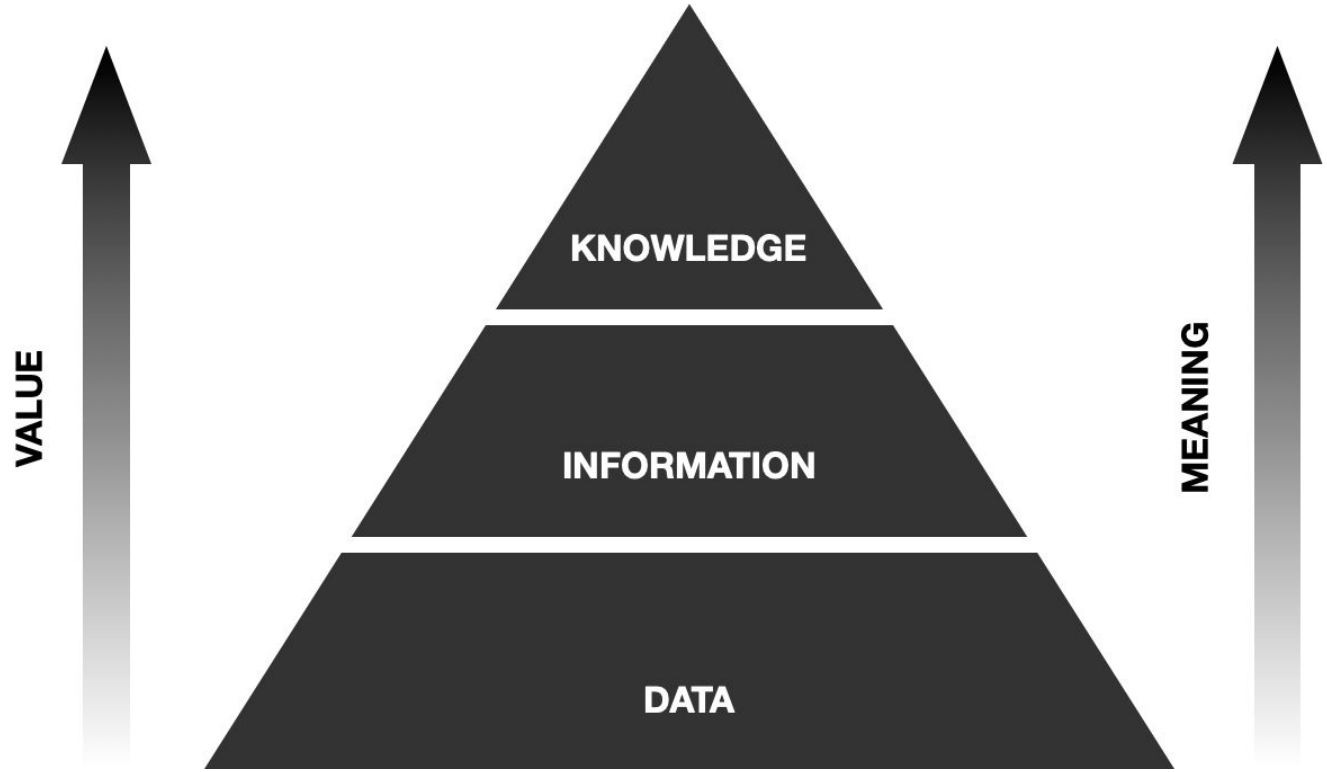
- Knowledge: *information* + (actionable) *understanding*
  - Take action based on the weather data, e.g. umbrella, coat, etc.



Data, Information & Knowledge



Data, Information & Knowledge



Data Reuse

Learning Outcome 2:

Identify the **challenges** and **solutions** to the **data reuse problem**

Introduction

- Scientists acquire data to discover knowledge, and are assessed for sharing knowledge in the form of scientific publications
- But the data itself has value for science:
  - It can be reused to discover further knowledge
  - New techniques or theories can require it to be reexamined

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

Introduction

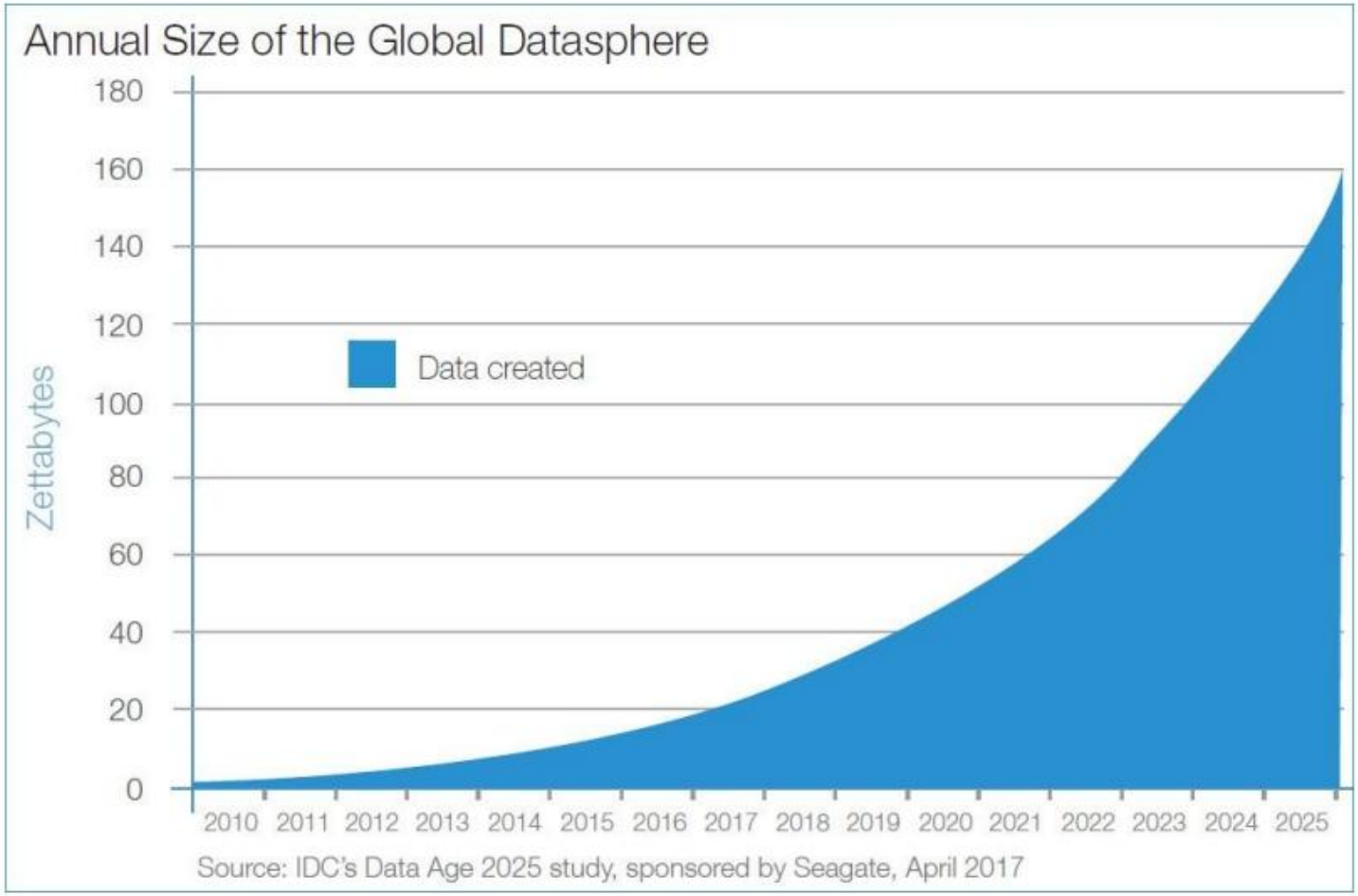
- Data sharing has been an afterthought for most scientists
- For a few types of data, the norm is deposition in public databases, while in some cases the data is included as an appendix to the (digital) publication itself
- However, it is still not uncommon that you need to contact the author of a scientific publication to request the data

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

Introduction

- To reuse research data, we typically must:
  - Read the publication wherein it was described
  - Figure out if the data is relevant
  - And then extract the metadata needed for interpreting it
- This is not a scalable approach!

Problem: Exponential Data Production



Problem: Exponential Data Production

Findability:

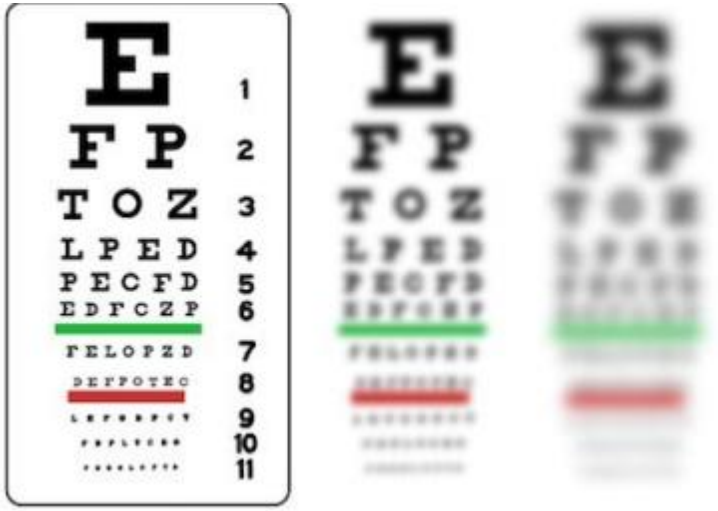
- More data ⇒ harder search
- Things can get lost amid a sea of things
- If it is not findable, it might as well not exist.

By Martin Handford, retrieved from: https://exploringyourmind.com/how-does-our-brain-find-waldo/

Problem: Exponential Data Production

Interpretability:

- More data ⇒ more costly to interpret
- We become myopic by necessity
  - can´t afford the time to read the fine-print ("to *the best of our knowledge*")
- If we cannot interpret it readily, then it is nearly useless.



Problem: Exponential Data Production

Interoperability:

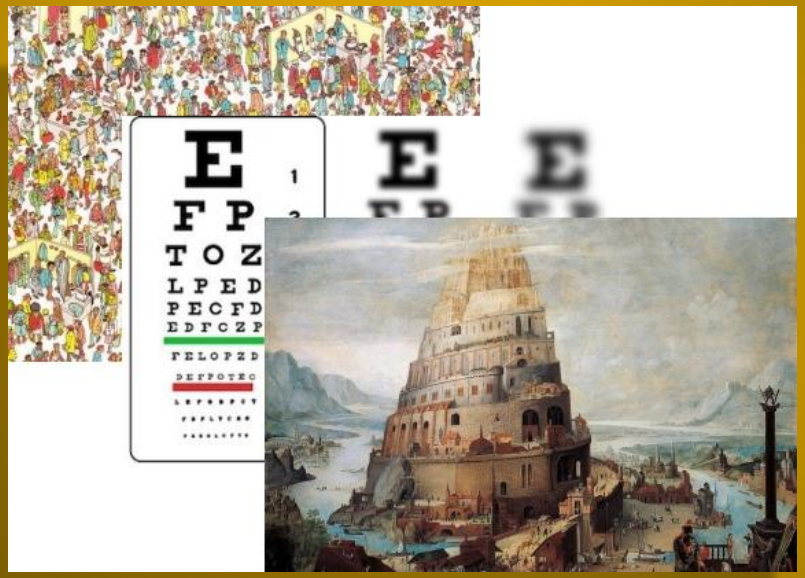
- More data & specialization ⇒ vocabulary and viewpoint divergence
- Use of local dialects leads to sundered data and knowledge
- If we don't find common ground, we cannot integrate data from related domains

By Abel Grimmer, retrieved from: http://cbcpnews.net/cbcpnews/the-tower-of-babel/

The Data Reuse Problem

Wrap-Up:

- Publishing data only in scientific papers is not enough
  - Papers are not efficient vehicles for knowledge transfer
- If we want our data to be reusable, we must publish it in a form that is:
  - Findable
  - Interpretable
  - Interoperable



Group Discussion Q2

How to make data:
Findable?
Interpretable?
Interoperable?

live.voxvote.com

PIN: 189128

2. How to make data Findable, Interpretable, Interoperable?

WordCloud | Frequency

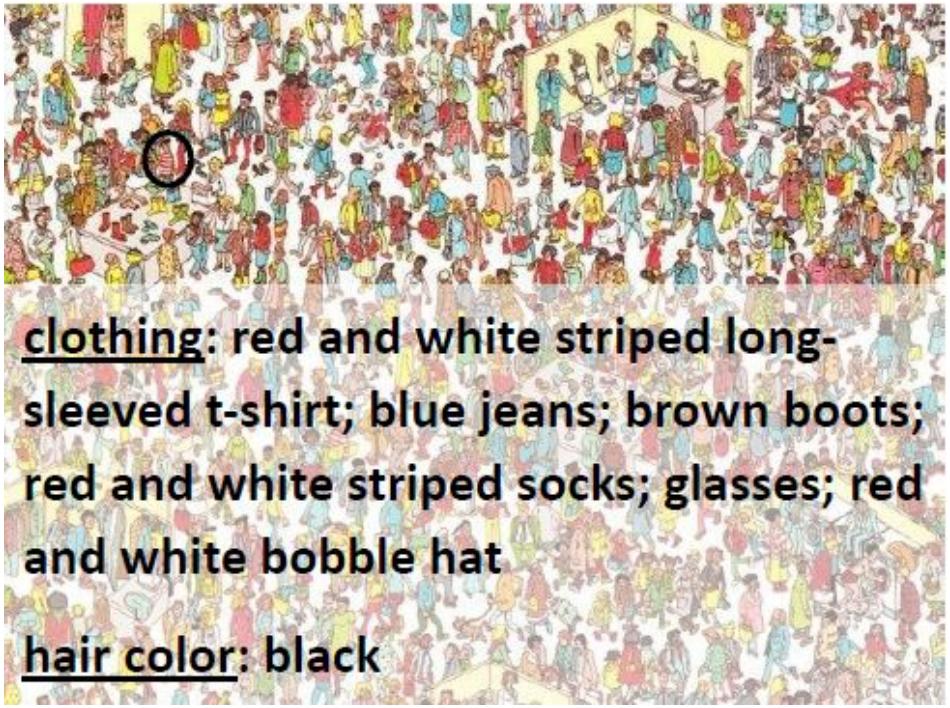
60 users voted

(2) Repositórios online
188128
aplicando princípios fair
Armazenamento em bases de dados abertas; Metadados suficientes para compreensão do contexto de recolha
Através de banco de dados armazenados na nuvem
Base de dados internacional
Base de dados
Base de dados, interpretação, e consideração de elementos comuns para interoperabilidade
Base de dados, uso de sistemas de banco de dados
bases de dados comuns; "formatos universais"
colocá-los em bases de dados e adicionar metadados (info de apoio que ajude a caracteriza-los)
Com uma boa organização
Common database
common databases
database
Deixar os dados claros no texto
Denilza
Depósito em repositórios de dados
Disponibilizar em plataformas de acesso aberto, simplificar a linguagem, adicionar documentos descritivos da dataset
Disponibilizar os dados de forma mais clara no corpo do texto do artigo e/ou disponibilizá-los em formato de anexo
Either attach the data as supplementary material or upload it to an open access repository
em bases, com recursos a conceitos universais e transversais a todas as áreas
Eventos científicos facilitam a divulgação dos dados
faço a menor ideia
Fiabilidade = ref fonte / Interpretabilidade = base / interoperabilidade = web
Findable - public access Interpretable - metadata Interoperable - good metadata!
Haver bases de dados assim como existem plataformas bibliográficas.

Solutions

Findability:

- Describe data with precise metadata useful for searching
- Use a common structured controlled vocabulary for metadata fields and values
- Put data in a repository that
  - Uses persistent unique identifiers
  - Indexes metadata and allows searches

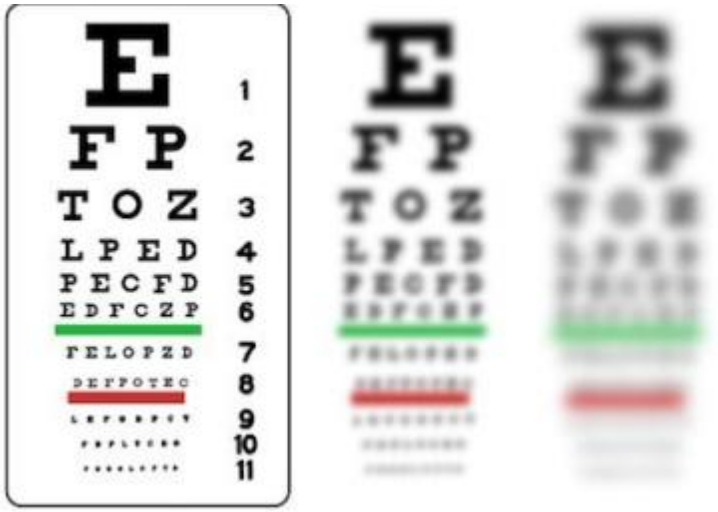


By Martin Handford, retrieved from: https://exploringyourmind.com/how-does-our-brain-find-waldo/

Solutions

Interpretability:

- Describe data with **sufficient metadata** for interpreting it and understanding the experimental context
- Use a **common (structured) controlled vocabulary** for metadata fields and values
    - e.g. medical terms and units vary
    - SNOMEDCT, ICD10

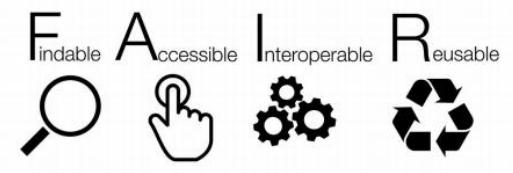


Solutions

Interoperability:

- Use a common (structured) controlled vocabulary for metadata fields and values
- Include cross-references to external data objects whenever suitable (e.g. NCBI taxon ID)

By Abel Grimmer, retrieved from: http://cbcpnews.net/cbcpnews/the-tower-of-babel/



Open Science & FAIR Principles

Learning Outcome 3:

Recognize the **demands of science funders** and debate their **pros and cons**

Introduction

- The need to improve scientific dissemination has been recognized by research communities and publishers
  - Leading to initiatives such as Open Science and the FAIR principles
- Funders recognized and are endorsing these initiatives
  - H2020 projects now require FAIR compliance
  - FCT will have a position on these issues soon
    - DMP mandatory

TO DO OPEN SCIENCE AND PUBLISH FAIR DATA!

What is Open Science?

Goals:

- Scientific research and its dissemination accessible to all levels of society
  - publications
  - data
  - physical samples
  - software
  - etc.
- Transparent and accessible knowledge shared and developed through collaborative networks

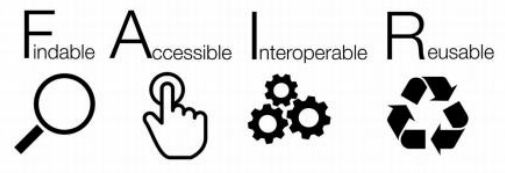
What is Open Science?

Layers:

- Open Access: research outputs distributed online, free of cost or access barriers
- Open Research: data, result and methodology clearly documented and freely available online
- Open-Notebook Science: primary record of a research project publicly available online as it is recorded—no insider information

What are the FAIR Data Principles?

- A set of four principles detailed in fifteen guidelines, that establish what Research output should aim for
  - Findability – (Meta)data should be easy to find for both humans and computers
  - Accessibility – (Meta)data should have a defined access protocol with authentication and authorization rules
  - Interoperability – (Meta)data should be integratable with other similar datasets and interpretable by applications or workflows for analysis, storage, and processing
  - Reusability – (Meta)data should be well described so that it can be interpreted and reused

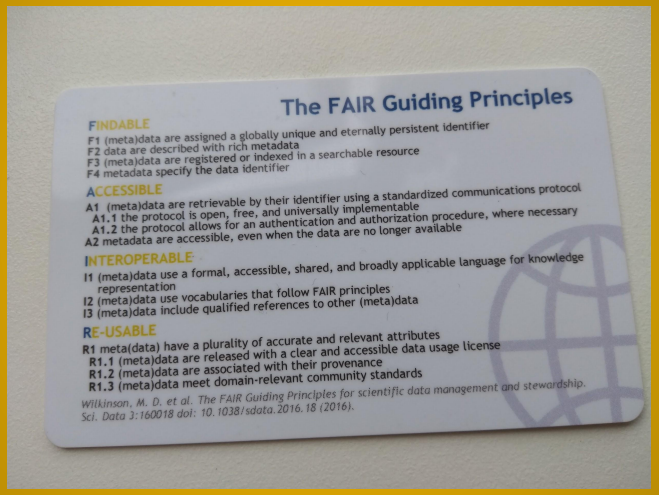


The FAIR Data Principles

=

The Solution for the Data Reuse Problem

- Wait, we talked about **Interpretability** but not **Accessibility** or **Reusability**…
  - **Reusability** is the end-goal, not the problem—it is contingent on *Interpretability* and *Interoperability*.
  - **Accessibility** concerns data repositories, not really researchers, and it is already well addressed. As long as you publish your data in a well-established repository and define an authorization policy (when applicable, such as for sensitive data) you are well off.



Group Discussion Q3

To be or not to be Open & FAIR? Pros and Cons!

live.voxvote.com

PIN: 189128

3. To be or not to be Open & FAIR? Pros and Cons.

WordCloud | Frequency

25 users voted

cons-Não receber o devido crédito pela sua contribuição à ciência (com datasets)
é uma forma de retribuição `sociedade que nos finacia - Pró Implica custos - Contra Visibilidade - Pró
balancing both variables
Be Open an Fair. Prós: visibilidade e usabilidade dos dados. Cons: Direitos dos dados; custos.
Con: Alguém utilizar os nossos dados e desenvolver uma ideia que já tinhamos de antemão Pro: Impacto pessoal em citações etc
Con: Encargos maiores para o investigação. Pro:Liberalização da ciência e aproximação à comunidade
É uma forma de retribuir à sociedade, que nos financia. Fazer o conhecimento chegar às pessoas.
FAIR principles are Ok, but Editorial policies relative to open access are unfair for researchers
Garantir retorno dessa prática para o investigador/grupo de investigação
implica custos
Incremental research, as a benefit
Industrial confidentiality (new patents and already existing products and processes)
O envolvimento das empresas é mais complicado com os princípios OPEN, por causa das questões proprietárias.
Para estudos de estudantes, muitas vezes, não há recursos financiados para publicações em revistas open acess - que apresentam altos custos. Ficando então mais fácil para investigações financiadas.
Penso que são mais Prós do que contras o uso Open & FAIR.
Possibilidade de divulgação massiva de dados sensíveis
pro : reproducibilidade, con : + fácil violar questões ética, mesmo que inadvertidamente
pros: contribuição para a sociedade e fomento de políticas públicas. Contra: vulneração da privacidade das pessoas
Pros: maior acesso à investigação Cons: custos de publicação
publicar os dados antes de publicar o artigo pode levar a que outros consigam evoluir as ideias e publicar primeiro, afetando o impacto da nossa possível publicação
questões éticas / RGPD
To be... Creio que é um retorno do investimento feito pela sociedade na ciência. No entanto, os dados pessoais devem ser protegidos por uma questão de ética e cumprimento do RGPD.

FAIR & Open Science—Pros & Cons

- Pros
  - Facilitates knowledge discovery
  - Promotes reproducibility / impedes fake science
  - Enables networking
  - Helps demystify science for the general public
- Cons
  - Care with sensitive data and with knowledge that has dangerous misuse potential
  - Harder to make money off of your research
  - Harder to stay ahead of your competitors

FAQ

- Can I receive credit for publishing data?
  - This is not yet well established, but we are amidst a shift towards crediting data publishers as much as paper publishers.
- Can’t someone publish a paper ahead of me if I release my data?
  - If someone can write a paper using your data ahead of you that supersedes yours, shame on you. If it does happen, you at least get credit for the use of your data, and will likely still be allowed to publish your paper as the original author of the data.
- What if someone uses my data without giving me credit?
  - The same can happen with paper publication. Reviewers and editors are expected to police this. Authors that do so can be red flagged.

To Be or Not to Be Open & FAIR

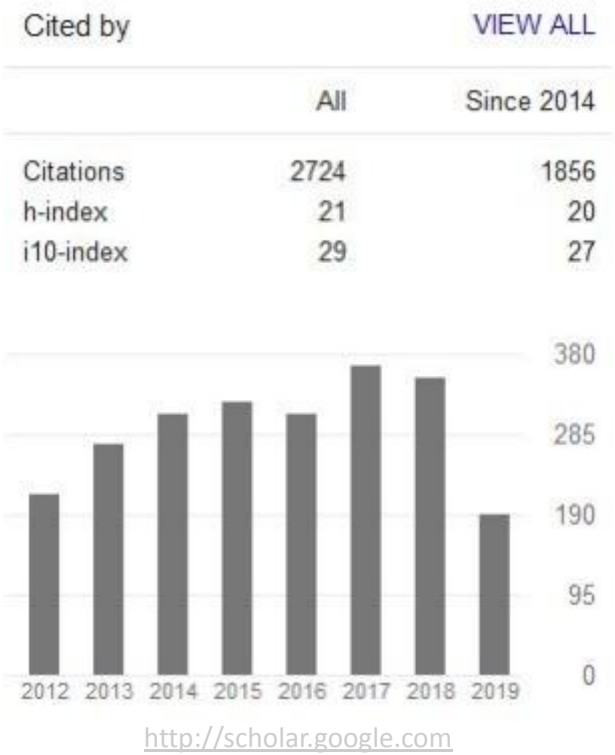
It Helps Science!

- Enables others to apply your knowledge in contexts beyond your foresight
- Enables others to reuse your data to make new research

To Be or Not to Be Open & FAIR

It Helps You!

- It is easier to find and reuse your own data
- It is easier to write and submit a research paper
- If others apply or reuse your research, you get more citations (citing aor crediting datasets is becoming common practice)



http://scholar.google.com

To Be or Not to Be Open & FAIR

You'll Need It To Get Funded!

- Soon it will be impossible to get public funding in Europe without adherence to Open Science and FAIR
- FAIR compliance is starting to be verified
- A good track record will contribute to project approval
    - also, it will help you getting noticed outside academia

To Be or Not to Be Open & FAIR

You'll Need It To Get Funded!

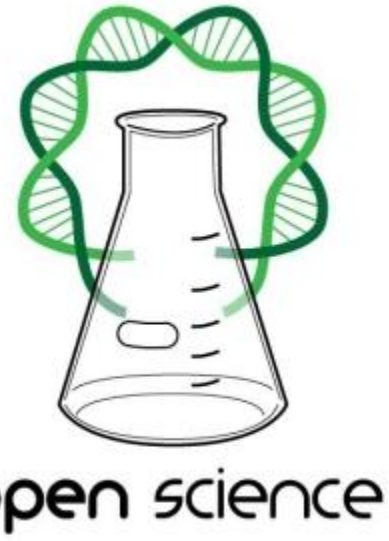
Recomendações para entidades produtoras, avaliadoras e financiadoras de Ciência

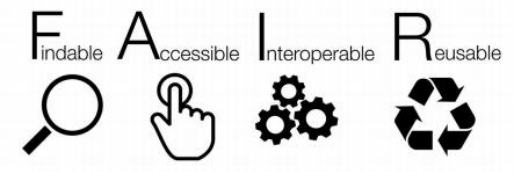
EVA-5 | Em qualquer processo de avaliação científica os indicadores quantitativos, quando utilizados, devem ser sempre entendidos como complementares do processo de avaliação qualitativa realizada por especialistas nas áreas disciplinares e o uso de métricas como o *Journal Impact Factor* (JIF) não deve ser considerado.

EVA-6 | A qualidade e do impacto da investigação, as práticas da área disciplinar e o contexto em que se realiza a avaliação científica devem orientar a escolha do conjunto de métricas a utilizar, se apropriado, que deverá ser abrangente e significante, no sentido de ser multifacetado e ser claramente apreendido o significado estatístico dos dados utilizados.

EVA-7 | Os procedimentos adotados num processo de avaliação científica devem ser claros e transparentes para os avaliados, que devem ter acesso aos dados, à semântica subjacente e às fórmulas de cálculo que tenham sido utilizados.

EVA-8 | Os diferentes intervenientes nos processos de avaliação científica devem ser envolvidos no desenho, monitorização e revisão dos indicadores quantitativos que





Open Science & FAIR Principles

Learning Outcome 4:

Comply with the **demands** of science **funders**

Introduction

- We've seen that adherence to Open Science and compliance with the FAIR principles are being **increasingly demanded by funding agencies**
  - Debated the **merits** and **demerits** of compliance
- **What must be done in practice to comply with these demands?**

TO DO OPEN SCIENCE AND PUBLISH FAIR DATA!

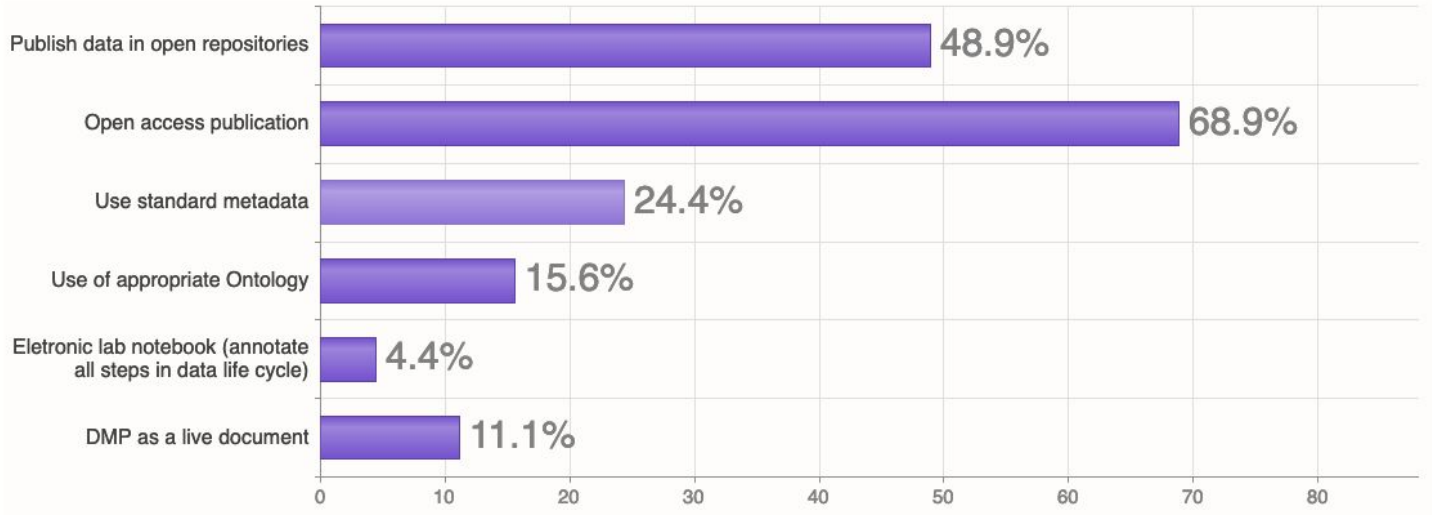
Group Discussion Q4

What are you doing in your daily activities towards Open Science and FAIR principles?

live.voxvote.com

PIN: 189128

4. What are you doing in your daily activities towards Open Science and FAIR principles?



45 users voted

How to be Open & FAIR?

Step 1 – Do Your Homework

- Consult the data steward of your institution
  - UC is preparing its policy framework and appointing data stewardship team
- Learn the basics in RDMkit
- Learn specific recipes in the FAIR Cookbook
- Lookup the best examples of FAIR data publication in your domain
- Consult information hubs about existing standards, such as FAIRsharing.org
- Search for key concepts through ontology lookup services, such as BioPortal

How to be Open & FAIR?

Step 1 – Do Your Homework

- Is there a default public database or repository for your research domain?
  - Does it have a metadata schema?
- Are there community metadata standards?
  - Do they cover your use case?
- Are there adequate ontologies?
  - If more than one, which is best?
- Are there default data (open) file formats (e.g. avoid .xls formats and use .csv instead)?

How to be Open & FAIR?

Step 2 – Do Your Work-Work

- Organize, Document & Annotate:
  - Your code / scripts / workflows,
  - Your protocols
  - Your data & metadata
- According to the applicable guidelines / standards or the repository where you're depositing your data / materials
- Using domain ontologies, recommended file formats
- Cross-referencing all relevant information objects

Photo by cottonbro Photo by cottonbro from Pexels

How to be Open & FAIR?

Step 3 – Deposit

- Deposit your data and materials in an appropriate public repository:
  - Code / scripts / workflows: GitHub, BitBucket
  - Protocols: Zenodo, FAIRDOMHub, Dataverse
  - Data: Domain database, one of the above
  - Metadata: Together with the data (as an accessory file, in the form of the repository)
- Under a declared usage license
- With a clear versioning policy

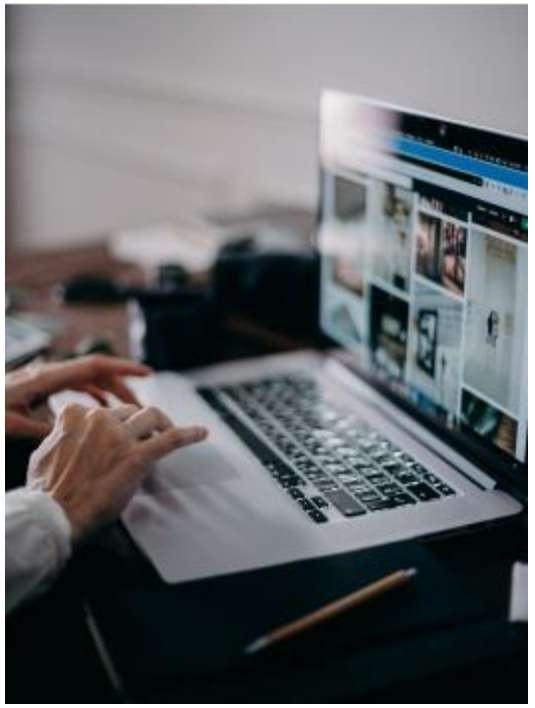


Photo by cottonbro Photo by cottonbro from Pexels

How to be Open & FAIR?

Step 3 – Repositories, how to select?

- Deposit your data and materials in an appropriate public repository:
  - Criteria for the Selection of Trustworthy Repositories (minimum criteria)
    - Provision of Persistent and Unique Identifiers (PIDs)
    - Metadata
    - Data access and usage license
    - Preservation

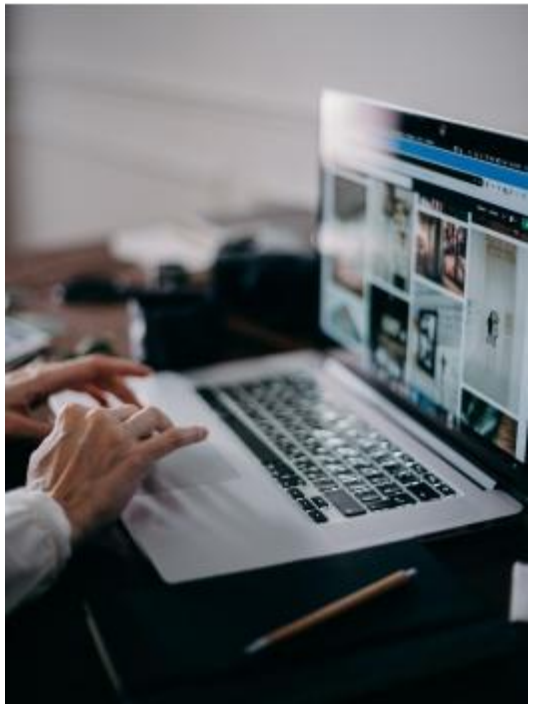


Photo by cottonbro Photo by cottonbro from Pexels

How to be Open & FAIR?

The Main Hurdles

- The Ontology landscape is complex and hard to navigate:
  - There are often overlapping ontologies for a given domain
  - And worse, the same concepts appear in several ontologies, sometimes with the same URI!!!
  - But there are also domains with no (suitable) ontology
- Metadata standards exist only for a few domains, and not all specify a data format for publication
- Generic data repositories (e.g. FAIRDOMHub, Zenodo, Dataverse) have rigid data models that are not compatible with all domains / standards

How to be Open & FAIR?

That sounds like a lot of work!

- It is, especially if you only do it at the time of publication:
  - Have to trace all the data—risk of data loss
  - Have to recall all the details about the experiment—risk of metadata loss, compromises reproducibility
- It is a lot of boring work to do at once—inertia and rush lead to poor job

Photo by Oladimeji Ajegbile Photo by Oladimeji Ajegbile from Pexels

How to be Open & FAIR?

The Data Lifecycle

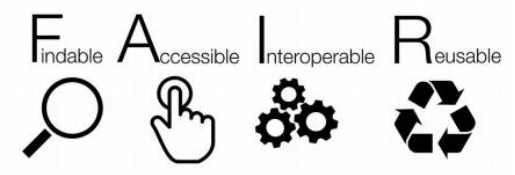
https://rdmkit.elixir-europe.org/

Manage research data across its whole lifecycle!

Research Data Management

Learning Outcome 5:

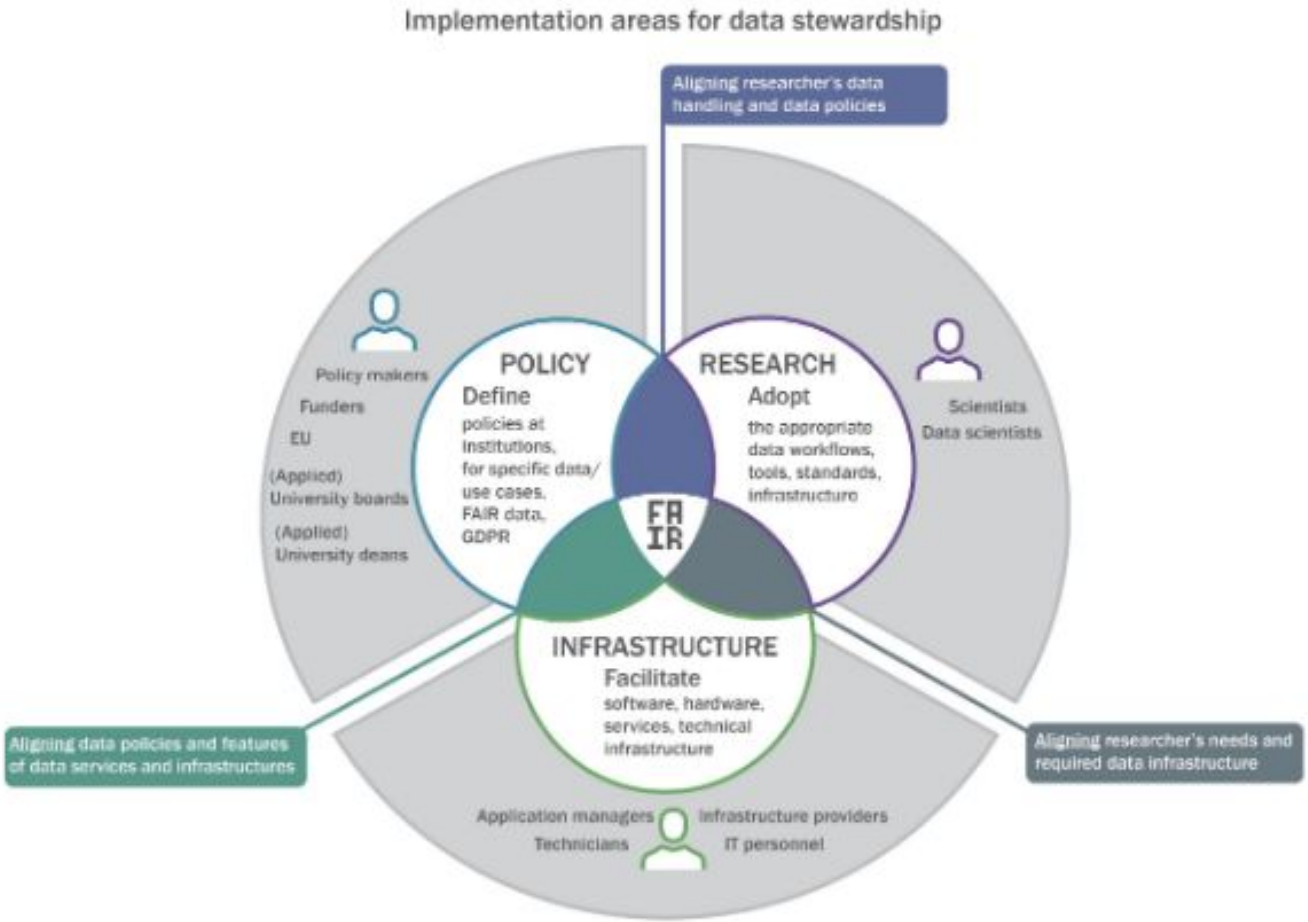
Recognize the **supportive role of data management in science**



Introduction

- Data management is a research domain in each own right
- Devoted to topics such as: Data Architecture, Data Modeling, Data Storage & Maintenance, Data Security, Data Integration, Metadata, Data Quality
- Researchers needn’t be data management experts
- But just like driving or using a computer, basic knowledge of data management is invaluable for a life in research

Data management profiles



Why Should I Care About Data Management?

Improve research:

By LAFS,
CC BY 2.0

Effectiveness – obtain more/better results

By Youmena,
CC BY 2.0

Efficiency – improve productivity and cost-efficiency

By ROZMOWA,
CC BY 2.0

Security – reduce data loss / control access to data

By Nithinan Tatah,
CC BY 2.0

Impact – facilitate dissemination and knowledge discovery

Data Management Commandments

- Thou shalt make a Data Management Plan for thy research project, even if it isn't funded by a grant
- Thou shalt allocate some time after each day experimenting to document everything, preferably in a digital platform (e.g. electronic lab notebook, local shared repository)
    - Thou shalt document the documentation process (e.g. Notion)
    - Thou shalt use version control (e.g. git)
    - Thou shalt use controlled vocabularies (public or your own, documented)
- For every data file (or collection thereof) thou shalt create a metadata file

Data Management Resources

- DMP platforms
  - Data Stewardship Wizard
  - Argos
- Electronic lab notebooks
  - Notion, UC?
- Data management platforms
  - Dataverse
  - Zenodo
- Data analysis platforms
  - Jupyter Notebook
  - Google Colabs
- Information hubs
  - RDMKit
  - FAIRCookbook
  - FAIRsharing.org

Take Home Messages

- Do the Best You Can
  - FAIRness is a spectrum, and FAIRer is a step forward
- Reach Out For Help!
  - Data Stewards and Data Managers can provide guidance
    - UC is working on these issues defining a policy and structures that will be able to help you.
- Things Will Get Easier!
  - There are people working towards more user-friendly data management solutions—they need feedback on what can be improved

Demystifying DMPs

CIBIT
Coimbra Institute for Biomedical Imaging and Translational Research

UNIVERSIDADE D COIMBRA
INSTITUTO DE CIÊNCIAS NUCLEARES APLICADAS À SAÚDE

What is a Data Management Plan?

Learning Outcome 1:

Recognize the **purpose** of **Data Management Plans**

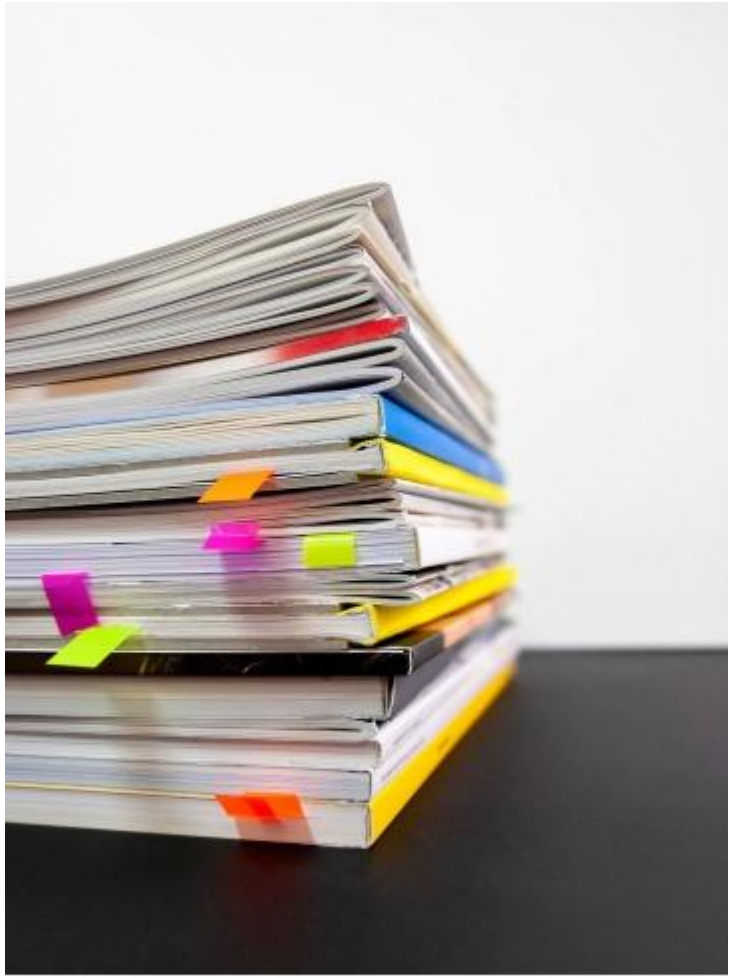
What is a DMP?

- A DMP is a formal document used to **plan and support data management activities** by anticipating **needs and requirements** in a (research) project, facility or institution
- It is the to data management what a blueprint is to construction

What is a DMP?

- A DMP should detail policies and methods pertaining to data:
  - Creation / collection
  - Documentation
  - Access
  - Preservation
  - Dissemination

- And ensure an adequate allocation of resources:
  - Human
  - Computational
  - Financial



Why Do We Need DMPs?

- The stick:
  - Many funding agencies now require that grant proposals be accompanied by a DMP
  - In particular, they require DMPs that demonstrate intent to comply with the FAIR data principles
  - Monitoring of the quality and execution of these DMPs is still light, but expected to tighten

Why Do We Need DMPs?

- The carrot:
  - DMPs are valuable tools in the planning of research activities to ensure the necessary resources are devoted to data management
  - Adequate planning can facilitate the task of ensuring compliance of research outputs with the FAIR principles

What should be in a DMP?

Learning Outcome 2:

List the **main topics** that should be covered by a DMP

What should be in a DMP?

Project Description	Existing Data	Created Data	Data Organisation
Preservation and Access Policies	Licences and Ethics	Resources and Responsibilities	Data Management Costs

What should be in a DMP?

The DMP should include (if not part of a project proposal), a **summary** of the research project or research unit to which it pertains.

This implies describing its **goals**, **specific methodologies, context**, etc.

One of the key aspects is to clearly describe the **sources of funding** for the data management activities described in the DMP.

What should be in a DMP?

Project Description	Existing Data	Created Data	Data Organisation
Preservation and Access Policies	Licences and Ethics	Resources and Responsibilities	Data Management Costs

What should be in a DMP?

If applicable, a description of any **existing data** should be provided.

This implies describing **sources of data**, and its **volume**, any **licenses** that apply or any **costs** associated with its usage.

What should be in a DMP?

Project Description	Existing Data	Created Data	Data Organisation
Preservation and Access Policies	Licences and Ethics	Resources and Responsibilities	Data Management Costs

What should be in a DMP?

All **data created** in the context of the research project or unit, **should be described in the DMP**.

This implies describing the **methodology of how data is created**, what **type of data** is to be created, and in what **volume**.

What should be in a DMP?

Project Description	Existing Data	Created Data	Data Organisation
Preservation and Access Policies	Licences and Ethics	Resources and Responsibilities	Data Management Costs

What should be in a DMP?

Metadata schemas that are applied must be **identified**, and how these metadata schemas are applied should also be characterized within the context of the research.

The **representation** of the data must also be addressed, this implies describing the **data format**.

Any **data structures** that apply should also be defined.

What should be in a DMP?

Project Description	Existing Data	Created Data	Data Organisation
Preservation and Access Policies	Licences and Ethics	Resources and Responsibilities	Data Management Costs

What should be in a DMP?

The DMP should **detail preservation and access policies**. These should be applied to one or multiple of the previously described datasets.

A **preservation and access policy** should define **where** the data will be hosted, **who** can access it, and **how** that access is to be performed.

What should be in a DMP?

Project Description	Existing Data	Created Data	Data Organisation
Preservation and Access Policies	Licences and Ethics	Resources and Responsibilities	Data Management Costs

What should be in a DMP?

In sequence with the definition of preservation and access policies for the datasets, it is essential to consider any existing **ethics issues**.

Selecting the **right licence** for the desired policy, is also fundamental.

- https://chooser-beta.creativecommons.org/

What should be in a DMP?

Project Description	Existing Data	Created Data	Data Organisation
Preservation and Access Policies	Licences and Ethics	Resources and Responsibilities	Data Management Costs

What should be in a DMP?

Having a clear **identification** of the **existing resources**, and how they will be allocated, is key to a good DMP.

Assets responsible for **data management activities** should also be identified.

What should be in a DMP?

Project Description	Existing Data	Created Data	Data Organisation
Preservation and Access Policies	Licences and Ethics	Resources and Responsibilities	Data Management Costs

What should be in a DMP?

The DMP should have a **detailed description of all costs** that are related with **data management activities**.

DMPs: Present and Future

Learning Outcome 3:

Describe the current state and future directions of DMPs

DMPs: Present

- In current practice, DMPs are mainly seen as a bureaucratic hassle
- They are static documents, prepared for grant applications because they are mandatory, but never or rarely updated
- They are generally not validated during the research project, and are never published, which prevents external validation

DMPs: Present

- The fact that **different funding bodies** use **different DMP templates** makes it difficult for researchers to get familiar with them and to **recognize the value**
- Moreover, most templates are **free text questionnaires** that look more like **surveys** than planning documents, and are only **human readable**
- All this results in poor quality DMPs, of **low practical value**

DMPs: Going into the future

The H2020 paradigm

3.1.3 List of Deliverables

Table 10. List of deliverables (Table 3.1c)

No.	Deliverable name	WP #	Delivery month	Type	Dissemination level	Responsible Partner
D1.1	Governance report	1	4	R	CO	UC
D1.2	Project handbook	1	6	R	CO	ACCEL
D1.3	Consortium meeting minutes	1	1, 12, 24, 36, 48, 60	R	CO	ACCEL
D1.4	Periodic reports and final report	1	12, 30, 48, 60 Final M60	R	CO PU	UC
D1.5	Risk assessment reports	1	6, 12, 24, 36, 42, 54	R	CO	CUC
D1.6	Consortium Agreement, Non-Disclosure	1	2	R	CO	UC
D1.7	Project Data Management Plan	1	6, 24	R	PU	ACCEL
	ticipating countries	2	12	R	PU	UHULL
D2.2	Report on women's requirement of PMH services	2	21	R	PU	UHULL
D2.3	Report on women's experiences of receiving					

2.2.1.3 Research data management

Table 5. Data management for [redacted]

What standards will be used?	
What types of data will the project generate/collect?	Clinical reports, psychological (self-reports; observational [quantitative and qualitative]), developmental, neurocognitive, neuroelectrophysiological measures, genetic, epigenetic, stress and inflammatory molecules
What standards will be used?	Guidelines for Data Management in Horizon 2020 and ECRIN data centers standards, compliant with ICH/GCP, E6 (R2), will be applied. All data will be collected and stored in OpenClinica. OpenClinica implements CDISC ODM XML representations in its Extract Data and Import Data modules as well as in other parts of the software. https://docs.openclinica.com/3.1/technical-documents/openclinica-and-cdisc-odm-specifications.
How will this data be exploited and/or shared/made accessible for verification and re-use?	Data generated during the project will be collected, stored and shared and archived according to a Data Management Plan (DMP) to be agreed within the Consortium in the first 6 months of the project. The provisions of the DMP can be summarized as follows: For data management and storage partner UC (coordinator) has a Data Center (CRU2C-UC) which will be responsible for data management and storage. The CRU2C-UC Data Center provides a software (OpenClinica) to support recruitment and secure data collection and data entry (eCRF). All services provided by the Data Center are compliant with Data Protection legislation and all processes are guided by approved Standard Operating Procedures (SOPs). A specific nominated member of the team will hold management responsibility for these activities. Datasets generated by [redacted] to the data center in anonymized format and large datasets will be transferred through secure FTP servers. Data will be stored in safe servers of CRU2C-UC Data Center and all data control procedures will be done according to standardized operating procedures. All data presented or published will be anonymized. After publication, the original anonymized data sets will be available on appropriate request, adhering to relevant regulatory requirements and ethical use of data approval.
How will this data be curated and preserved?	[redacted] e back up and safely stored for long term preservation and curation onto secure areas on a CRU2C-UC Data Center central server according to standard policies and procedures of the Data Center.

DMPs: Going into the future

The H2020 DMP

- A questionnaire covering the following topics:
  - **Data Summary**
    - Describe the data to be acquired/produced
  - **FAIR data**
    - Detail how you'll comply with the FAIR principles
  - **Allocation of resources**
    - Who does what and what it costs
  - **Data security**
  - **Ethical aspects**
  - Other issues

DMPs: Future

- To be of practical use, a DMP should be:
  - A living document that is updated as needed
  - Both human and machine-readable
  - Comply with a common standard
  - Be shared

DMPs: Going into the future

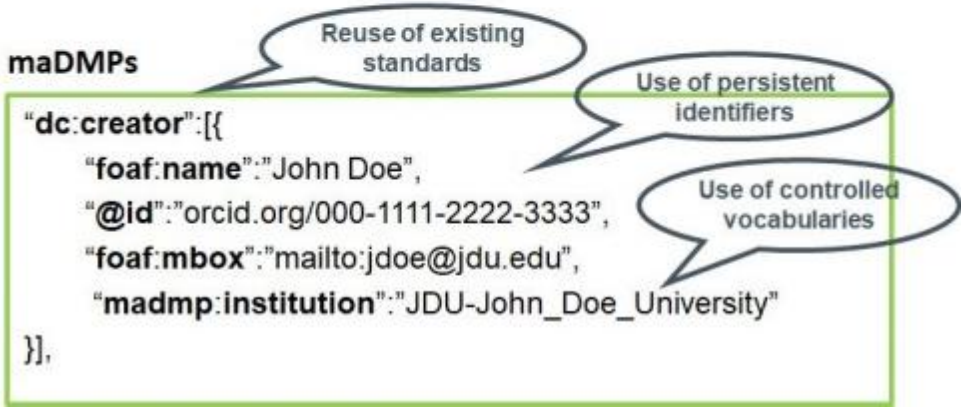
The Machine-Actionable DMP (maDMP):

- **Machine and human readable** descriptions
- **Automated** policy enforcement
- **Interoperable** DMP version
- **Extensible**

Current DMPs

```
<admindata>
    <question>Who is the Principle Investigator?</question>
    <answer>The PI is John Doe from the JDU</answer>
</admindata>
```

maDMPs



DMPs: Going into the future

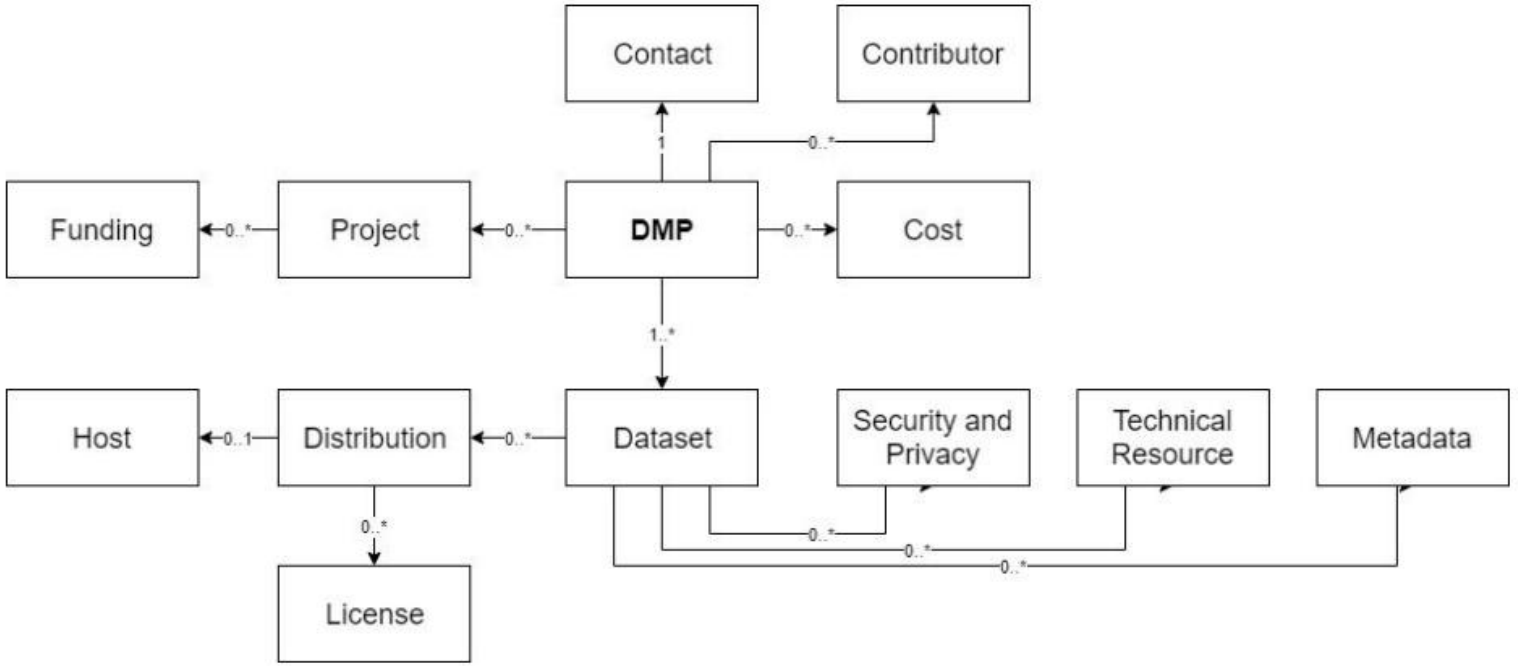
- The **RDA DMP Common Standards Working Group** was created to focus on the **standardization of the knowledge contained in a DMP**.
- Its objective was to **establish a metadata application standard** that defines a core set of elements for a DMP.
- The metadata application standard is modular in design, and allows for extensions.

RDA
RESEARCH DATA ALLIANCE

Scan for more!

DMPs: Going into the future

- A **minimum set of universal terms** to ensure **basic interoperability** of systems using DMPs.



DMPs: Going into the future

Applications of a maDMP:

- One DMP for all templates
- DMP maturity model
- Automation in both creation and monitoring during the project's life-cycle

The Take Home Message

The benefits of **DMPs**:

- Promote **good data management practices**
- Assist in **compliance with FAIR data principles**
- Ensure **adequate allocation or resources** in data management activities.

The benefits of **maDMPs**:

- **Automation** (creation, validation, policy enactment)
- **Increase usefulness**

A good place to start.

Practical Guide to the International Alignment of Research Data Management - Extended Edition

https://doi.org/10.5281/zenodo.4915862

Data Management Plans

Learning Outcome 4:

Understand the options available to fill in the various sections of a DMP and/or where to look for them.

FAIR Data Documentation

- **Metadata Standards**
  - Specify the metadata fields that must be filled in to enable data interpretation
  - Can be looked up in FAIRsharing
  - Not all domains have metadata standards
    - Adapt a similar standard: core metadata fields are essentially common to all domains
    - Adopt a generic standard: probably not rich enough for FAIR data

FAIR Data Documentation

- **Controlled Vocabularies & Ontologies**
  - Used to fill in the metadata fields and/or to annotate the data itself
  - Can be looked up in BioPortal
  - Check metadata standard for ontology recommendations
  - Choose domain-specific ontologies when able
- **Metadata Capturing**
  - Automatically from experimental equipment or software
  - Manually, in electronic lab notebook
  - Manually, on paper
  - **Daily**, throughout the project

Data Quality

- Equipment **calibration** and **verification** practices
- Equipment-provided **quality assessment**
- Service provider's **quality assurance** (e.g. ISO certification)
- Use of **controlled vocabularies**
- **Data validation**
  - Upon entry
  - *A posteriori*
- **Data cleaning**
  - Remove outliers
  - Handle missing values

Storage

- Personal (group) storage
  - Acquired through the project (in budget)
  - Pre-existing (maintenance costs in budget)
- Institutional storage
  - Acquired through the project (in budget)
  - Pre-existing (usage costs covered by overheads)
- Cloud storage
  - National: FCCN, BioData.pt, …
  - International: Google, Amazon, ...

Backups

- Do-It-Yourself
  - Redundancy:
    - Physical: redundant data server, hard-drive, tape
    - Virtual: virtual machine
  - Periodicity:
    - Triggered: periodic/automatic check of changes
    - Manual: upon changes
    - Periodic: e.g. hourly, daily
- Other
  - Check institutional or service provider ’s backup practices and assurances

Security & Protection

- Malicious attacks and accesses are essentially impossible to prevent
- Accidents happen: fire in a data center, early hardware failure
- Sensitive data should **always** be encrypted, so that when access happens, it is not compromised
- All data should be backed-up in a separate physical location (or the cloud) so that when accidents or malicious attacks happen, nothing substantial is lost
- **Access protocols:** who will have access and how access is controlled
- Consult IT experts in your institution or elsewhere

Legal & Ethical Requirements

- **Personal Data**
  - Carefully review GDPR checklist
  - Data anonymization policy for sharing amongst project partners and/or publication
  - Personally identifying information is sensitive and should **always** be encrypted
  - Research subjects always need to sign consent forms
  - Research subjects have the right to request their data and ask you to remove it any point
  - Assign a person responsible for overseeing personal data

Legal & Ethical Requirements

- **Intellectual Property**
  - Typically owned by the host institution
  - Make sure to check your institutional policies and your contract with them

- **Code of Conduct**
  - Avoid gender bias and discrimination
  - Avoid bias and discrimination towards minorities
  - Handle occurrences of inappropriate behaviour
  - Typically deferred to the host institution
    - There must be agreement between projects spanning multiple institutions and countries

Data Sharing & Preservation

- **Data sharing**
  - All non-sensitive data should be made public to comply with FAIR principles
  - You can have an embargo if needed to secure publication of research articles
    - This should not exceed 2 years after the end of the project

- **Data preservation**
  - You are legally bound to preserve research data for a number of years (check institutional, national and funders' policies)
  - Data that is shared in a public repository is essentially preserved (but you may still be legally required to host it as well)
  - Data from failed experiments due to faulty materials, protocols, etc, can be erased

Final Remarks

- Consult institution experts (IT, policy, ethics, etc) and ask for their contribution on the DMP
- Consult national experts if your institution doesn't have in-house expertise
- Consult data management portals
    - e.g. ELIXIR RDMkit

Data Management Plans

-

The **Science Europe DMP** Template

CIBIT
Coimbra Institute for Biomedical Imaging and Translational Research

UNIVERSIDADE D COIMBRA
INSTITUTO DE CIÊNCIAS NUCLEARES APLICADAS À SAÚDE

SCIENCE EUROPE
Shaping the future of research

The Science Europe DMP Template

- Science Europe is the European association for the representation of both public research institutions and fundings bodies
- It comprises 37 member organisations from 27 European countries
- Open Access and Research Data are two of the main priorities of Science Europe.
- Science Europe is also involved in topics that have an impact in research
  - Copyright
  - Data-related legislation

The Organisation of the Science Europe DMP template

Learning Outcome 1:

Understand how the Science Europe DMP template is organised, and how it interlinks with the RDM data lifecycle.

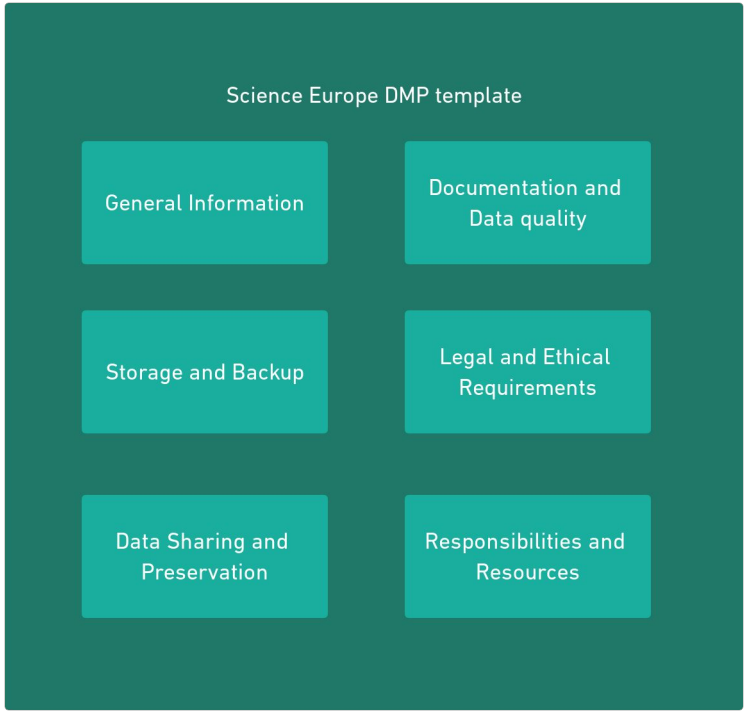
The Science Europe DMP Template

- Science Europe has established a DMP template in an attempt to establish the core requirements for DMPs
- It is organized in two parts

The Science Europe DMP Template

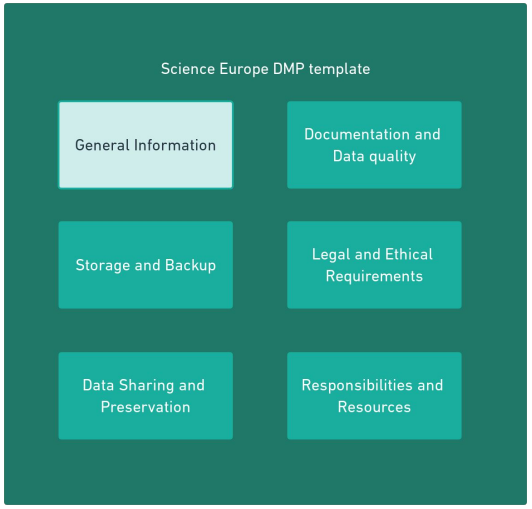
The core requirements for a DMP according to Science Europe template:

- Organized in 6 categories (RDM data lifecycle)
- Comprises 15 questions + administrative info.



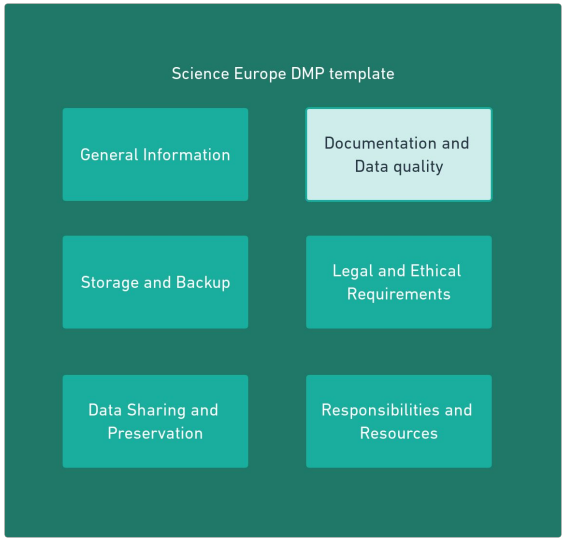
The Science Europe DMP Template

- General Information
  - Information on the DMP document
    - Title , Project, Version, funding agency
- Data description
  - How will new data be collected or produced and/or how will existing data be re-used?
    - Methodologies, constraints on reuse, ...
  - What data (for example the kinds, formats, and volumes) will be collected or produced?
    - Type of data, format, open or proprietary, etc.



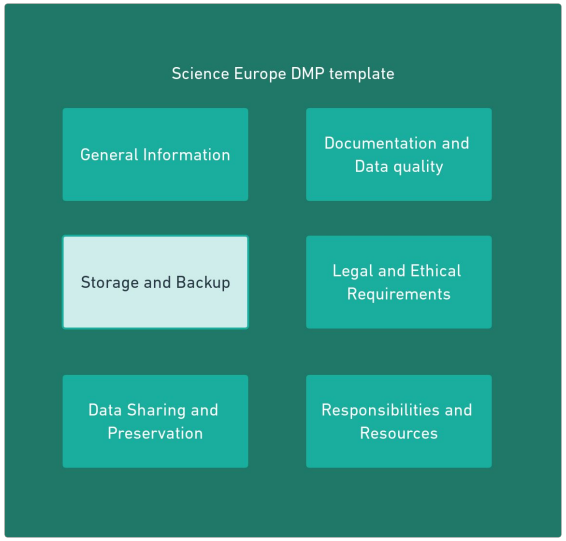
The Science Europe DMP Template

- Documentation and data quality
    - What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?
        - Standards, description of practices, data organisation, how will metadata be captured and recorded
    - What data quality control measures will be used?
        - Calibration, standardised data capture, data entry validation, peer review, controlled vocabularies



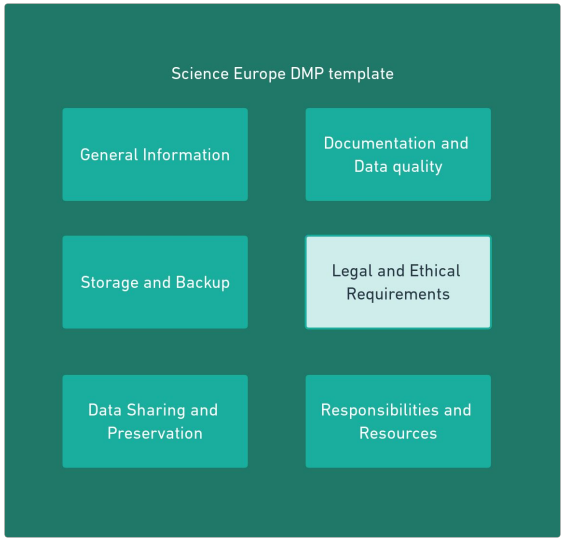
The Science Europe DMP Template

- Storage and backup during the research process
  - How will data and metadata be stored and backed up during the research process?
    - Backup, hosts, ...
  - How will data security and protection of sensitive data be taken care of during the research?
    - Recovery of data, data access, data sensitivity, etc.



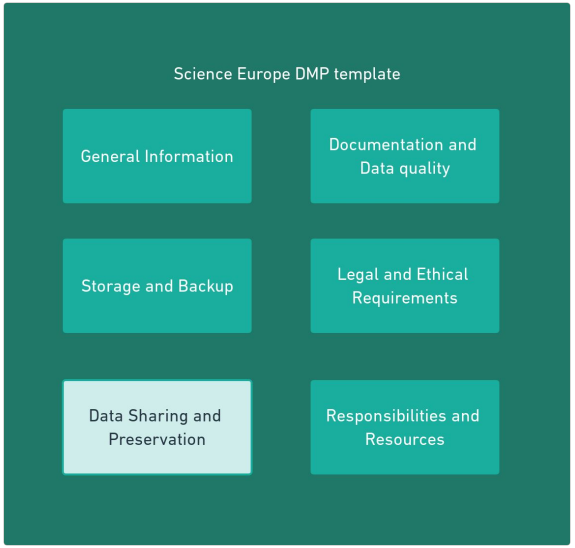
The Science Europe DMP Template

- Legal and ethical requirements, codes of conduct
  - If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?
    - GDPR, anonymisation of personal data, encryption, ...
  - How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?
    - who is the owner of the data, intellectual property rights, Licenses that apply
  - How will possible ethical issues be taken into account, and codes of conduct followed?
    - Institutional, national and international ethical guidelines



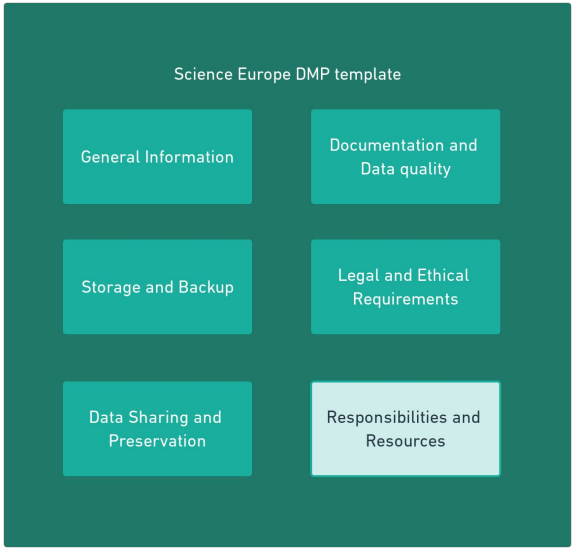
The Science Europe DMP Template

- Data sharing and long-term preservation
  - How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?
    - Data preservation policy, availability of the data,...
  - How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)?
    - What to preserve, what to destroy, rules/guidelines
  - What methods or software tools will be needed to access and use the data?
    - Specific tool, access mechanisms, etc.
  - How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?
    - What type of PUI, and how...



The Science Europe DMP Template

- Data management responsibilities and resources
  - Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?
    - Roles and responsabilities, hierarchy, etc.
  - What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?
    - Costs associated, resources (storage, etc.)



ARGOS and the use of templates

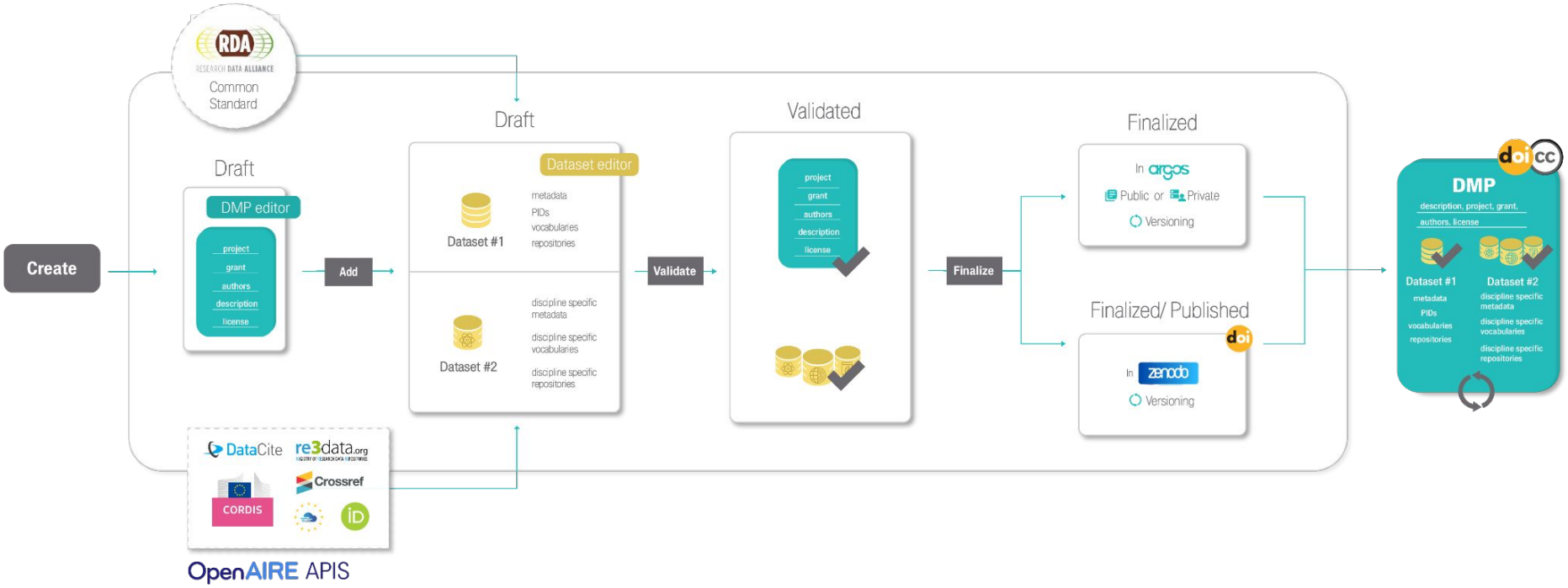
Learning Outcome 3:

Pipeline

The Argos platform

- ARGOS is the joint effort of OpenAIRE and EUDAT to deliver an open platform for Data Management Planning that addresses FAIR and Open best practices and assumes no barriers for its use and adoption.
- Argos is a service for creating and publishing plans that describe data management activities, commonly known as Data Management Plans (DMPs). The plans are produced as machine-actionable outputs (ma-DMPs), in the form of rich text documents, following Open and FAIR practices and are published in Zenodo.

The Argos platform



The Argos platform

- Argos consists of two main editors:
  - DMP editor
  - Dataset editor
- Also, the editors incorporate a mechanism to enable compliance with the RDA DMP Common Standard.