National Initiatives for Open Science in Europe

Fair data and principles
Research data management

Milutin Radonjić
University of Montenegro
Faculty of Electrical Engineering
mico@ucg.ac.me

Outline

- ✓ Fair data and principles
- ✓ Initiatives concerning FAIR
- ✓ Research data life cycle
- ✓ Tools for FAIRness
- ✓ Research data management
- ❑ Some slides are taken from:
  - ➢ René van Horik (rene.van.horik@dans.knaw.nl), Cees Hof (cees.hof@dans.knaw.nl)

    Data Archiving and Networked Services (DANS), The Netherland, Both active in RDM Training Work package of EOSC-hub project

    and
  - ➢ Judit Fazekas-Paragh, Edit Görögh, Ádám Száldobágyi, NI4OS-Europe,

    Train the trainers event: 'FAIR data principles", 19-20, February, 2020

FAIR data principles

- ❑ What are FAIR principles?
- ❑ The FAIR data principles are created to make research data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable
- ❑ The FAIR data principles allow us to facilitate navigation in open science, sharing knowledge and make scientific work collaborative using digital technologies

FAIR data principles

- ❑ **Findable** – Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets
- ❑ **Accessible** – Stored for long term such that they can be easily accessed and/or downloaded with well-defined license and access conditions (Open Access when possible), whether at the level of metadata, or at the level of the actual data content
- ❑ **Interoperable** – Ready to be combined with other datasets by humans as well as computer systems
- ❑ **Re-usable** – Ready to be used for future research and to be processed further using computational methods

FAIR data principles

- The principles apply to three types of entities:
  - data (or any digital object),
  - metadata (information about this digital object),
  - infrastructure (search resource)

Findable

- The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.
- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

Source: https://www.go-fair.org/fair-principles/

F1 - globally unique and persistent identifier

- ❑The most important principle
- ❑Globally unique and persistent identifiers remove ambiguity by assigning a unique identifier to every element
- ❑In this context, identifiers consist of an internet link (e.g., a URL that resolves to a web page that defines the concept such as a particular human protein)
- ❑Identifiers can help other people understand exactly what you mean, and they allow computers to interpret your data in a meaningful way
- ❑Identifiers are essential to the human-machine interoperation that is key to the vision of Open Science
- ❑ Identifiers will help others to properly cite your work when reusing your data

Source: https://www.go-fair.org/fair-principles/

F1 - globally unique and persistent identifier

- F1 stipulates two conditions for your identifier:

1. It must be globally unique
2. It must be persistent

- Examples of globally unique and persistent identifiers:
  - One particular person on planet earth has this globally unique and persistent identifier: https://orcid.org/0000-0002-9670-2431
  - Identifier that uniquely links to the results of a study estimating the FAIRness of different data repositories: doi:10.4121/uuid:5146dd06-98e4-426c-9ae5-dc8fa65c549f
  - The human polycystin-1 protein has a globally unique and persistent identifier given by the UniProt database: http://www.uniprot.org/uniprot/P98161
  - Polycystic kidney disease Type 1 has a globally unique and persistent identifier given by the OMIM database: http://omim.org/entry/173900

Source: https://www.go-fair.org/fair-principles/

F1 - globally unique and persistent identifier

- ❑ Example services that supply globally unique and persistent identifiers:
  - ❑ Identifiers.org provides resolvable identifiers in the form of URIs: http://identifiers.org
  - ❑ Persistent URLs: http://www.purlz.org
  - ❑ Digital Object Identifier: http://www.doi.org
  - ❑ Research Resource Identifiers: https://scicrunch.org/resources
  - ❑ Universally unique identifier: https://en.wikipedia.org/wiki/Universally_unique_identifier

Source: https://www.go-fair.org/fair-principles/

F2 - Data are described with rich metadata

- ❑ Metadata can (and should) be generous and extensive, including descriptive information about the context, quality and condition, or characteristics of the data
- ❑ Rich metadata allow a computer to automatically accomplish routine and tedious sorting and prioritising tasks
- ❑ Someone should be able to find data based on the information provided by their metadata, even without the data’s identifier
- ❑ Rich metadata implies that you should not presume that you know who will want to use your data, or for what purpose
- ❑ So, as a rule of thumb, you should never say ‘this metadata isn’t useful’; be generous and provide it anyway!

Source: https://www.go-fair.org/fair-principles/

F3 - Metadata clearly and explicitly include the identifier

- ❑Simple and obvious principle, but of critical importance to FAIR
- ❑The metadata and the dataset they describe are usually separate files
- ❑The association between a metadata file and the dataset should be made explicit by mentioning a dataset’s globally unique and persistent identifier in the metadata
- ❑As stated in F1, many repositories will generate globally unique and persistent identifiers for deposited datasets that can be used for this purpose

Source: https://www.go-fair.org/fair-principles/

F4 - (Meta)data are registered or indexed in a searchable resourc

- ❑Identifiers and rich metadata descriptions alone will not ensure 'findability' on the internet
- ❑If the availability of a digital resource such as a dataset, service or repository is not known, then nobody (and no machine) can discover it
- ❑There are many ways in which digital resources can be made discoverable, including indexing (for example, Google spiders)
- ❑For scholarly research data, we need to be more explicit about indexing
- ❑Principles F1-F3 will provide the core elements for fine-grained indexing by some current repositories and future services

Source: https://www.go-fair.org/fair-principles/

Accessible

- Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation
- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

Source: https://www.go-fair.org/fair-principles/

A1 - (Meta)data are retrievable by their identifier using a standardised communications protocol

- Most users of the internet retrieve data by ‘clicking on a link’
- This is a high-level interface to a low-level protocol called tcp, that the computer executes to load data in the user’s web browser
- Principle A1 states that FAIR data retrieval should be mediated without specialised or proprietary tools or communication methods
- This principle focuses on how data and metadata can be retrieved from their identifiers
- Most data producers will use http(s) or ftp

Source: https://www.go-fair.org/fair-principles/

A1.1 - The protocol is open, free, and universally implementable

- ❑To maximise data reuse, the protocol should be free (no-cost) and open (-sourced) and thus globally implementable to facilitate data retrieval
- ❑Anyone with a computer and an internet connection can access at least the metadata
- ❑This criterion will impact your choice of the repository where you will share your data.
- ❑Examples:
  - ❑ HTTP, FTP, SMTP, …
  - ❑ A counter-example would be Skype, which is not universally-implementable because it is proprietary
  - ❑ Microsoft Exchange Server protocol is also proprietary

Source: https://www.go-fair.org/fair-principles/

A1.2 - The protocol allows authentication and authorisation

- ❑The ‘A’ in FAIR does not necessarily mean ‘open’ or ‘free’
- ❑It implies that one should provide the exact conditions under which the data are accessible
- ❑Even heavily protected and private data can be FAIR
- ❑Machine can automatically understand the requirements, and then either automatically execute the requirements or alert the user to the requirements
- ❑Users to create a user account for a repository to authenticate the owner (or contributor) of each dataset, and to potentially set user-specific rights

Source: https://www.go-fair.org/fair-principles/

A2 - Metadata are accessible, even when the data are no longer available

- ❑Datasets tend to degrade or disappear over time because there is a cost to maintaining an online presence for data resources
- ❑Links become invalid and users waste time hunting for data that might no longer be there
- ❑Storing the metadata generally is much easier and cheaper
- ❑Principle A2 states that metadata should persist even when the data are no longer sustained
- ❑Metadata are valuable when planning research, especially replication studies
- ❑If the original data are missing, tracking down people, institutions or publications associated with the original research can be extremely useful

Source: https://www.go-fair.org/fair-principles/

Interoperable

- The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.
- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

Source: https://www.go-fair.org/fair-principles/

I1 - (Meta)data use a formal, accessible, … language

- ❑Humans should be able to exchange and interpret each other’s data
- ❑Data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings
- ❑Interoperability typically means that each computer system at least has knowledge of the other system’s data exchange formats
- ❑To ensure automatic findability and interoperability of datasets, it is critical to use:
  - ❑ commonly used controlled vocabularies, ontologies, thesauri
  - ❑ a good data model (a well-defined framework to describe and structure (meta)data)

Source: https://www.go-fair.org/fair-principles/

I2 - (Meta)data use vocabularies that follow FAIR principles

- The controlled vocabulary used to describe datasets needs to be documented and resolvable using globally unique and persistent identifiers
- This documentation needs to be easily findable and accessible by anyone who uses the dataset

Source: https://www.go-fair.org/fair-principles/

I3 - (Meta)data include qualified references to other (meta)data

- A qualified reference is a cross-reference that explains its intent
- For example, „*X is regulator of Y*“ is a much more qualified reference than „*X is associated with Y*“, or „*X see also Y*“
- The goal is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data
- You should specify if one dataset builds on another data set, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset
- All datasets need to be properly cited (i.e., including their globally unique and persistent identifiers)

Source: https://www.go-fair.org/fair-principles/

Reusable

- The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.
- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage license
  - R1.2. (Meta)data are associated with detailed provenance
  - R1.3. (Meta)data meet domain-relevant community standards

Source: https://www.go-fair.org/fair-principles/

R1 - (Meta)data are richly described with attributes

- ❑ It will be much easier to find and reuse data if there are many labels are attached to the data
- ❑ Focuses on the ability of a user (machine or human) to decide if the data is actually USEFUL in a particular context
- ❑ The data publisher should provide not just metadata that allows discovery, but also metadata that richly describes the context
- ❑ This may include the experimental protocols, the manufacturer and brand of the machine or sensor that created the data, the species used, the drug regime, etc.
- ❑ The term ‘plurality’ indicates that the metadata author should be as generous as possible in providing metadata, even including information that may seem irrelevant

Source: https://www.go-fair.org/fair-principles/

R1.1 - (Meta)data are released with a clear and accessible data usage license

❑ Under ‘I’, are covered elements of technical interoperability. R1.1 is about legal interoperability.

❑ Ambiguity about usage rights could severely limit the reuse of your data by organisations that struggle to comply with licensing restrictions

❑ Clarity of licensing status will become more important with automated searches involving more licensing considerations

❑ The conditions under which the data can be used should be clear to machines and humans

Source: https://www.go-fair.org/fair-principles/

R1.2 - (Meta)data are associated with detailed provenance

- ❑For others to reuse your data, they should know where the data came from (i.e., clear story of origin/history), who to cite and/or how you wish to be acknowledged
- ❑Include a description of the workflow that led to your data:
  - ❑ Who generated or collected it?
  - ❑ How has it been processed?
  - ❑ Has it been published before?
  - ❑ Does it contain data from someone else that you may have transformed or completed?
- ❑Ideally, this workflow is described in a machine-readable format

Source: https://www.go-fair.org/fair-principles/

R1.3 - (Meta)data meet domain-relevant community standards

- ❑It is easier to reuse data sets if they are similar: same type of data, data organised in a standardised way, well-established and sustainable file formats, documentation (metadata) following a common template and using common vocabulary
- ❑If community standards or best practices for data archiving and sharing exist, they should be followed
- ❑If submitter have valid and specified reasons to divert from the standard good practice for the type of data to be submitted, this should be addressed in the metadata

Source: https://www.go-fair.org/fair-principles/

Initiatives concerning FAIR

❑FAIRsFAIR (Fostering FAIR data practices) - www.fairsfair.eu

(aims to supply practical solutions for the use of the FAIR data principles throughout the research data life cycle)

❑GO FAIR initiative - www.go-fair.org

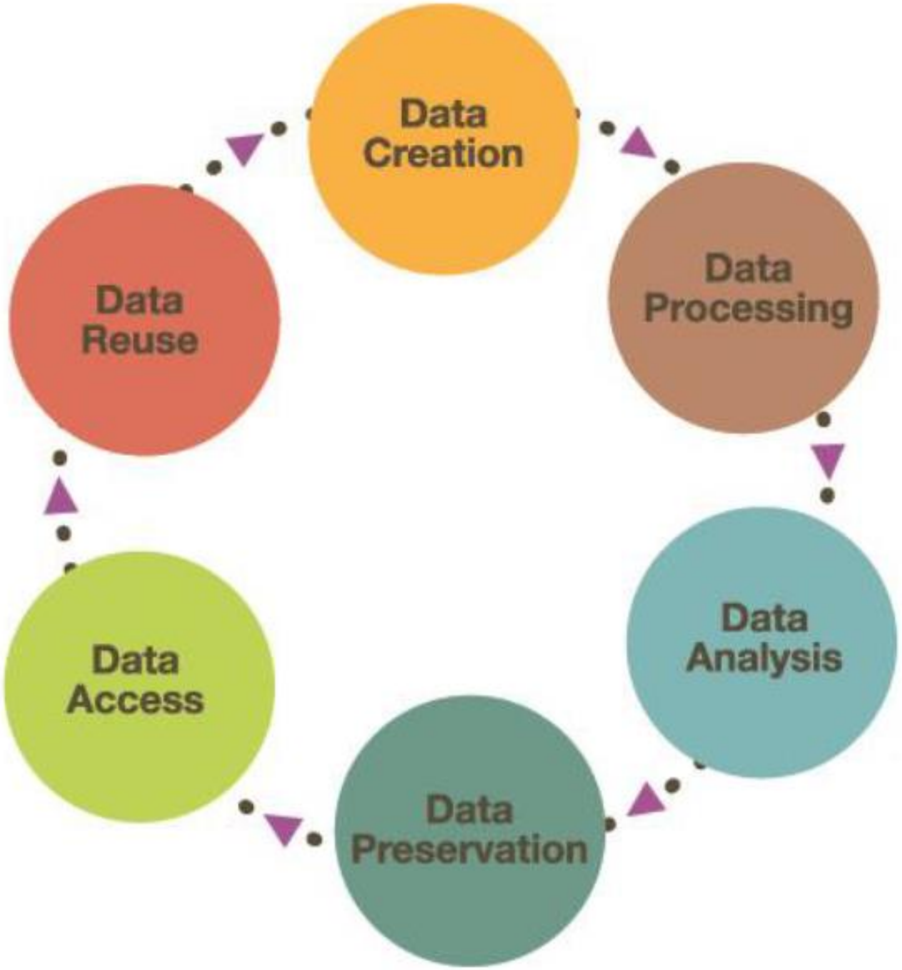
(aims to implement the FAIR data principles)

❑FAIRsharing - fairsharing.org

(A curated , informative and educational resource on data and metadata standards inter-related to databases and data policies)

What aspects to consider when looking for the right tool?

- ❑ Research data life cycle
- ❑ Disciplines
- ❑ The possibilities and expectations of the given institution
- ❑ Funder and journal expectations



Picture: https://blogs.ntu.edu.sg/lib-datamanagement/data-lifecycle/

Tools for all disciplines:

- RDA Metadata Standards Directory - The overriding goal is to develop a collaborative, open directory of metadata standards applicable to scientific data can help address infrastructure challenges

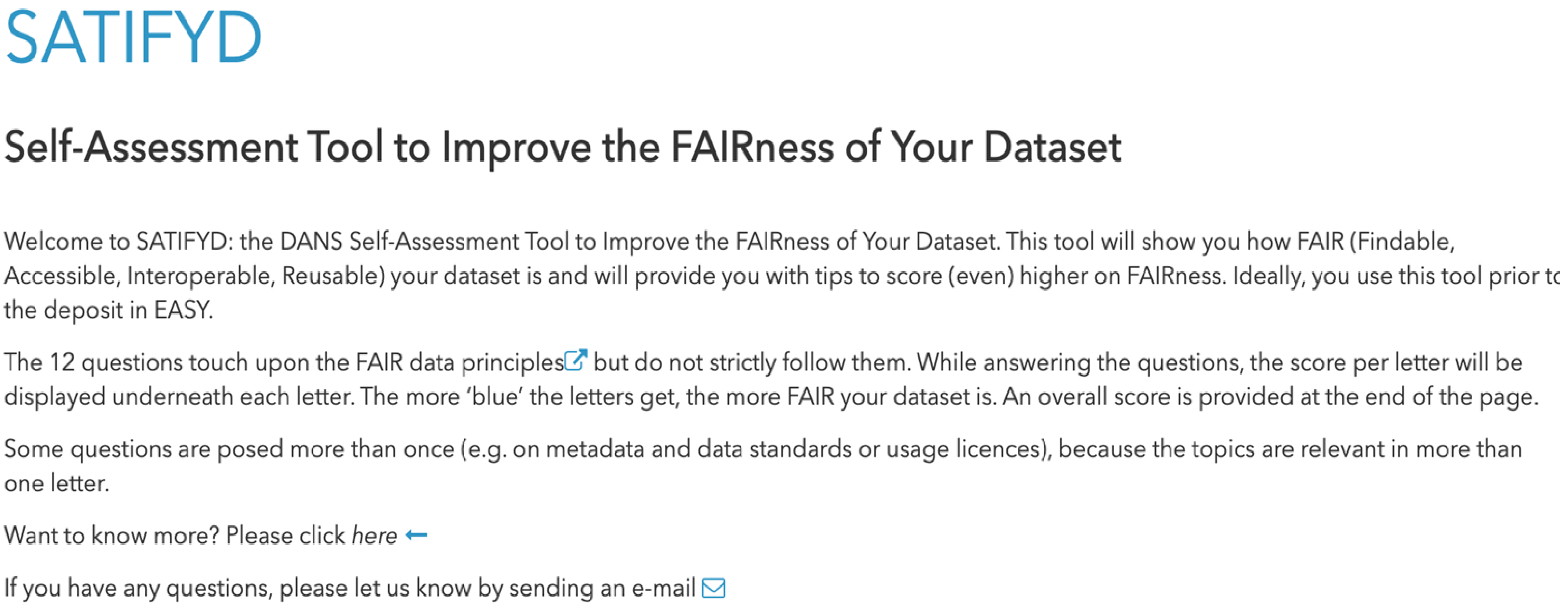
FAIR Evaluation Services

- ARGOS - An online tool in support of automated processes to creating, managing, sharing and linking DMPs with research artifacts they correspond to
- AMNESIA - data anonymization tool, that allows to remove identifying information from data. Amnesia not only removes direct identifiers like names, but also transforms secondary identifiers like birth date and zip code so that individuals cannot be identified in the data.

Tools for all disciplines:

- ❑B2SHARE - user-friendly, reliable and trustworthy way for researchers, scientific communities and citizen scientists to store and publish small-scale research data from diverse contexts
- ❑Zenodo - an open repository for all scholarship, enabling researchers from all disciplines to share and preserve their research outputs
- ❑re3data.org - registry of research data repositories
- ❑Checklist to evaluate FAIRness of data - helps you assess the quality (FAIRness) of your dataset(s) and the trustworthiness of the repository that you have chosen

Example: Tool to assess the FAIRness of a dataset



SATIFYD

Self-Assessment Tool to Improve the FAIRness of Your Dataset

Welcome to SATIFYD: the DANS Self-Assessment Tool to Improve the FAIRness of Your Dataset. This tool will show you how FAIR (Findable, Accessible, Interoperable, Reusable) your dataset is and will provide you with tips to score (even) higher on FAIRness. Ideally, you use this tool prior tc the deposit in EASY.

The 12 questions touch upon the FAIR data principles↗ but do not strictly follow them. While answering the questions, the score per letter will be displayed underneath each letter. The more 'blue' the letters get, the more FAIR your dataset is. An overall score is provided at the end of the page.

Some questions are posed more than once (e.g. on metadata and data standards or usage licences), because the topics are relevant in more than one letter.

Want to know more? Please click *here* ←

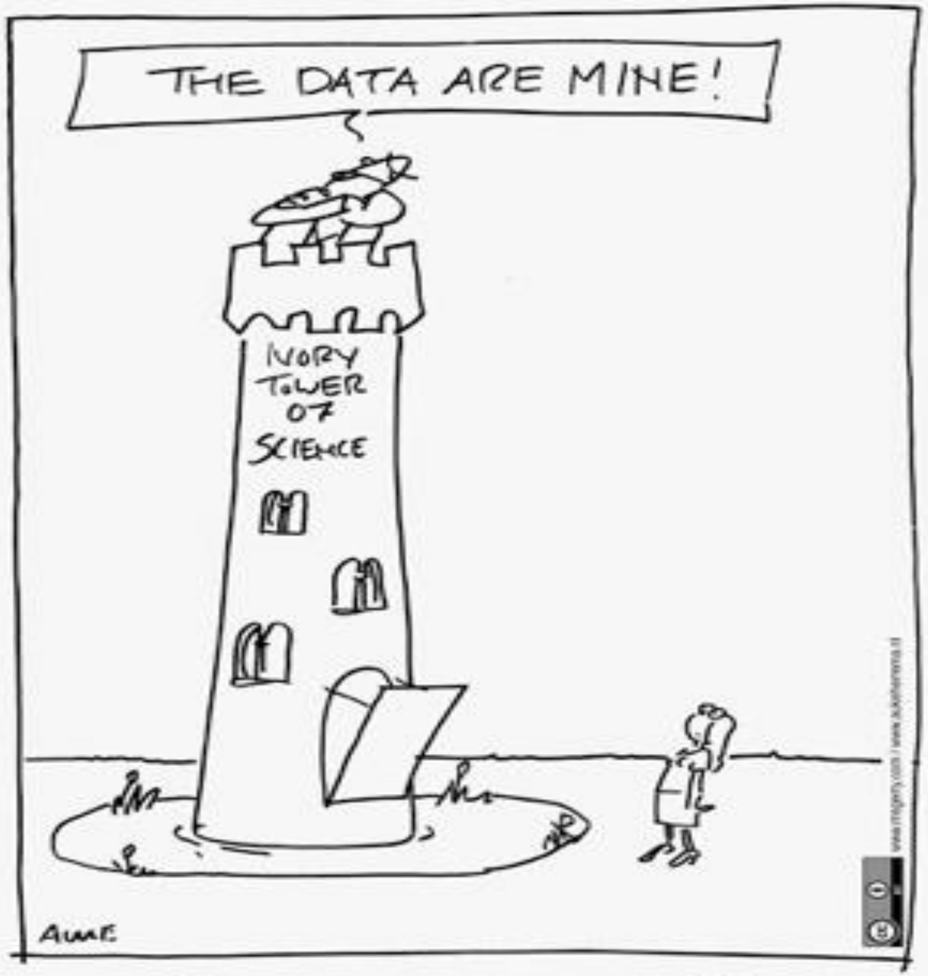
If you have any questions, please let us know by sending an e-mail ✉

Source: https://satifyd.dans.knaw.nl/

Research Data Management

- ❑ Research data management refers to the development, execution and supervision of (research) plans, policies, programs and practices that control, protect, deliver and enhance the value of (research) data and information assets
- ❑ Why is it important:
  - ❑ Saves time and resources
  - ❑ It helps to prevent errors and increases quality of research
  - ❑ Allows to validate and replicate findings
  - ❑ Facilitates sharing of data
  - ❑ ….

Research Data Management



SCENE FROM THE PAST?

MY DATA
MY TAX MONEY
SOCIETY
MONEY FOR RESEARCH
AUKE

DATA SHARING

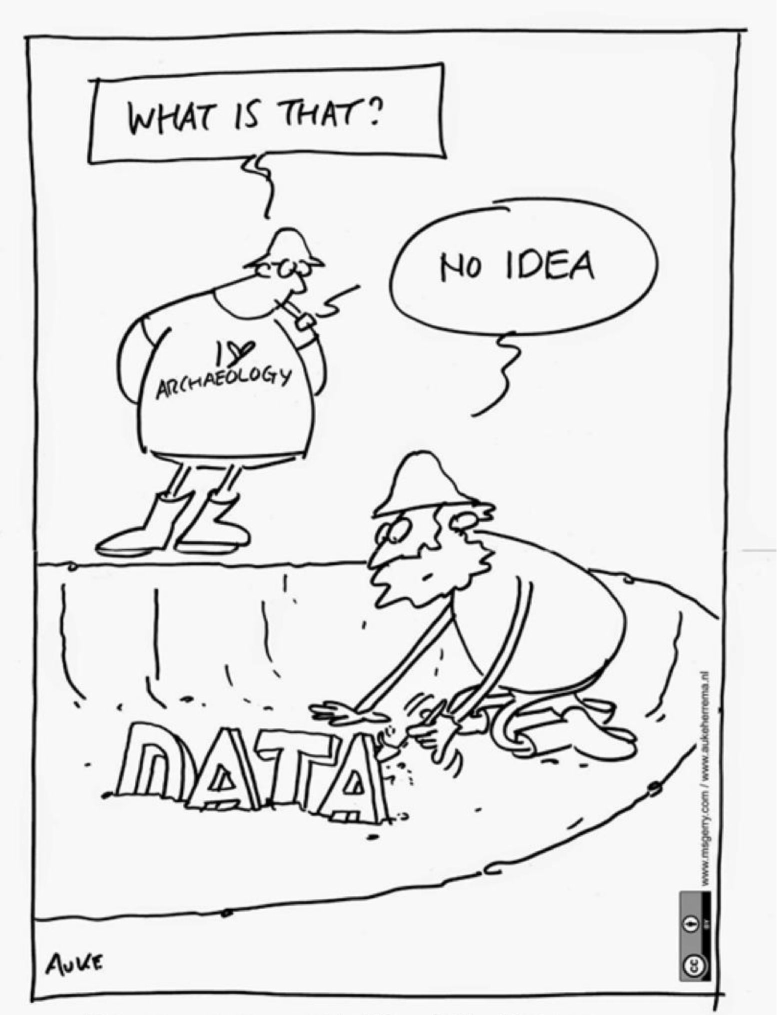
Research Data Management

DATA?
UHH
OVER THERE
PUBLICATION
1
2
3
AUKE
www.magerry.com / www.aukeherrema.nl

PUBLICATIONS AND DATA

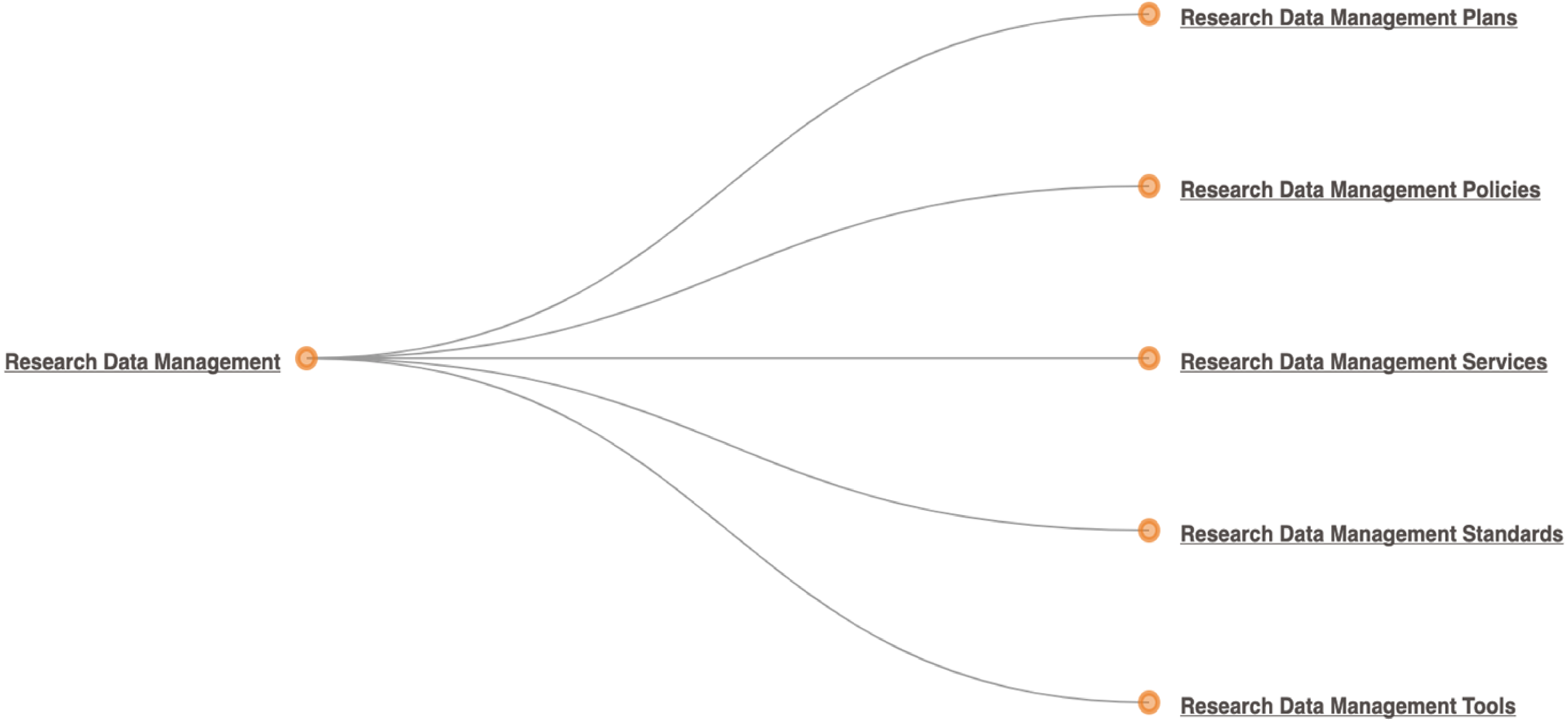
DATA TREE
FREE
PICKING
www.magerry.com / www.aukeherrema.nl

THE FUTURE?



DATA FOR FUTURE GENERATIONS

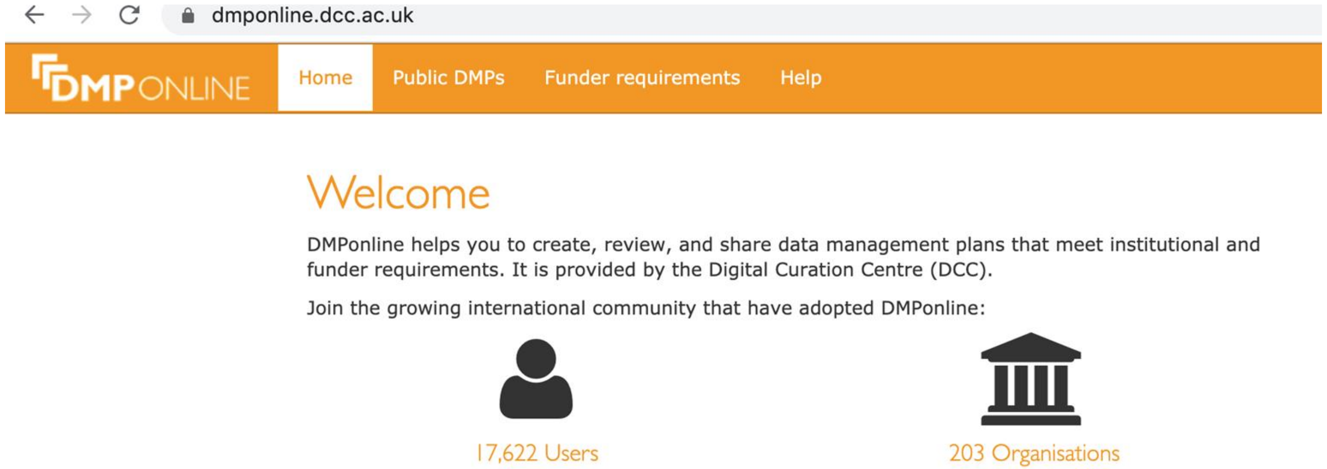
Research Data Management


Research Data Management
Research Data Management Plans
Research Data Management Policies
Research Data Management Services
Research Data Management Standards
Research Data Management Tools

source: https://www.fosteropenscience.eu/resources#tab-9

Research Data Management Plan

- ❑ Document that contains information about handling, organising, documenting and enhancing research data, and enabling their sustainability and sharing for a research project
- ❑ A DMP Describes and analyzes workflows along the Research Data Lifecycle
- ❑ A DMP can be a few paragraphs short up to several pages long
- ❑ Example of online tool to create a DMP: https://dmponline.dcc.ac.uk/



Welcome

DMPonline helps you to create, review, and share data management plans that meet institutional and funder requirements. It is provided by the Digital Curation Centre (DCC).

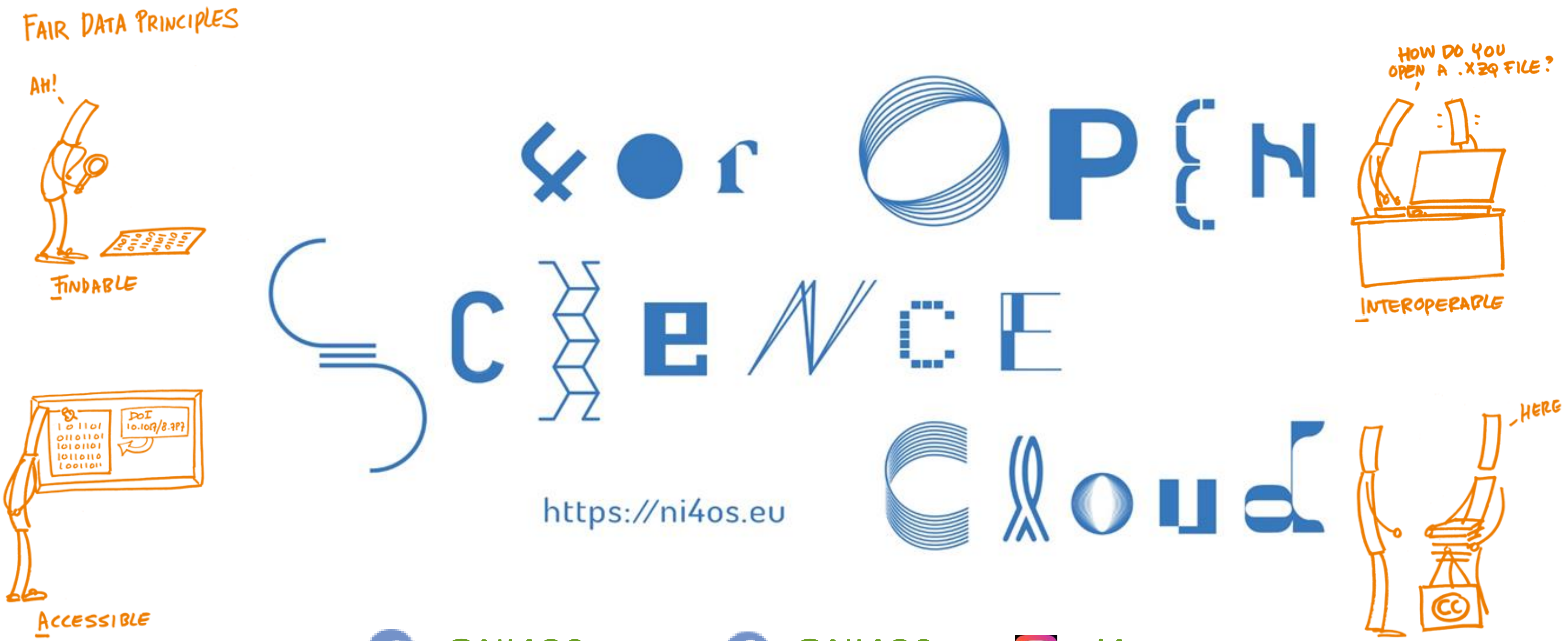
Join the growing international community that have adopted DMPonline:

17,622 Users

203 Organisations

Thanks!

National Initiatives for Open Science in Europe



@NI4OS_eu @NI4OS ni4os.eu