An introduction to Open Science and Research Data Management

Marco J Morelli
Center for Omics Sciences
IRCCS Ospedale San Raffaele

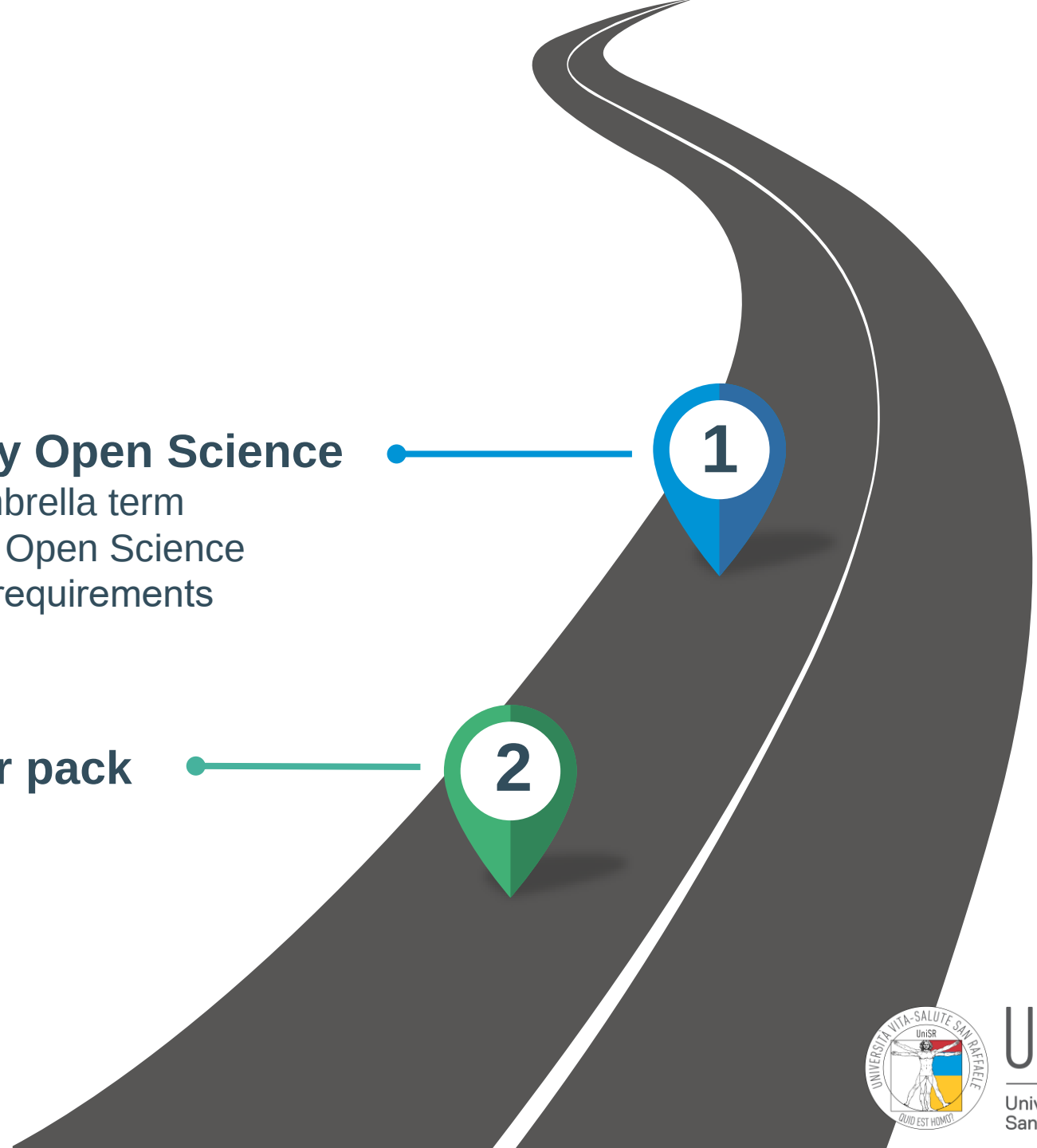
Research Policy Office
Università Vita-Salute San Raffaele

Marco Soriano
Research Policy Office
Università Vita-Salute San Raffaele

Open Science Team
email: open.science@unisr.it

UniSR
Università Vita-Salute San Raffaele

Road map



Setting the context: why Open Science

1

Theory

Open Science: an Umbrella term
Solutions & benefits from Open Science
Funders' & Publishers' requirements

Research Data Management: starter pack

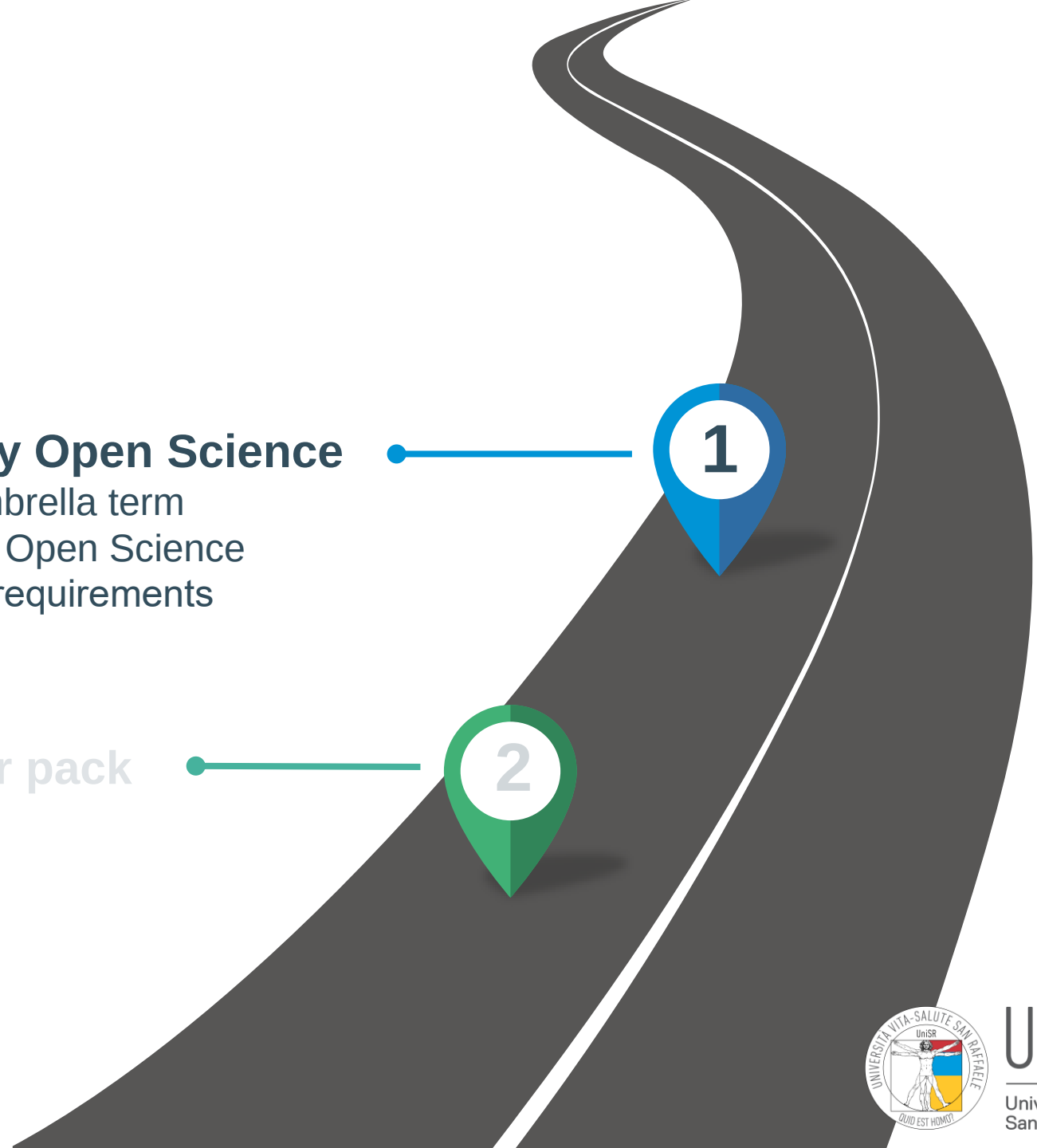
2

Practice

Data Management Plan (DMP)
Data preservation and sharing
FAIR Principles

UniSR
Università Vita-Salute San Raffaele

Road map



Setting the context: why Open Science

1

Theory

Open Science: an Umbrella term
Solutions & benefits from Open Science
Funders' & Publishers' requirements

Research Data Management: starter pack

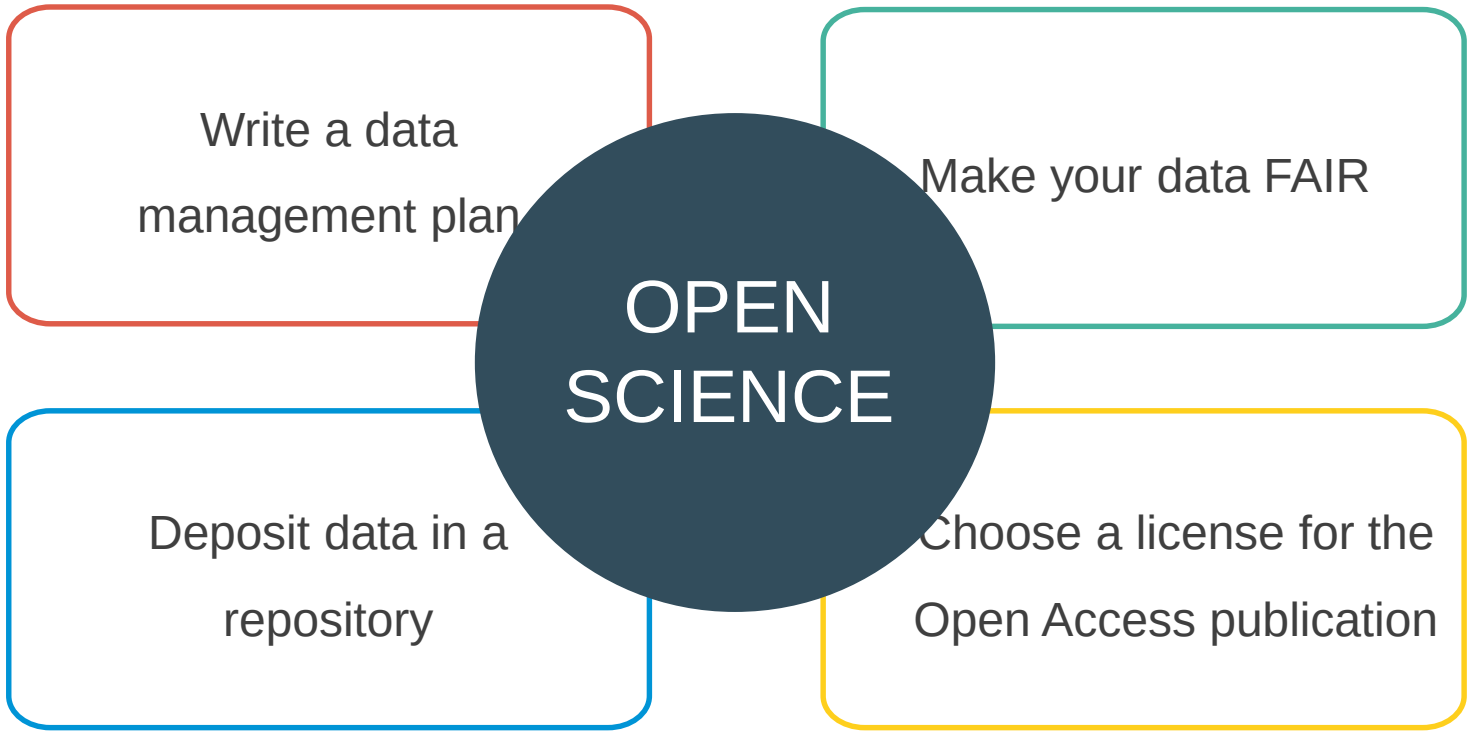
2

Practice

Data Management Plan (DMP)
Data preservation and sharing
FAIR Principles

UniSR
Università Vita-Salute San Raffaele

Have you ever been asked to...



UniSR
Università Vita-Salute San Raffaele

Open Science: a definition

- Open Science describes an on-going change in the way research is performed, researchers collaborate, knowledge is shared, and science is organized.

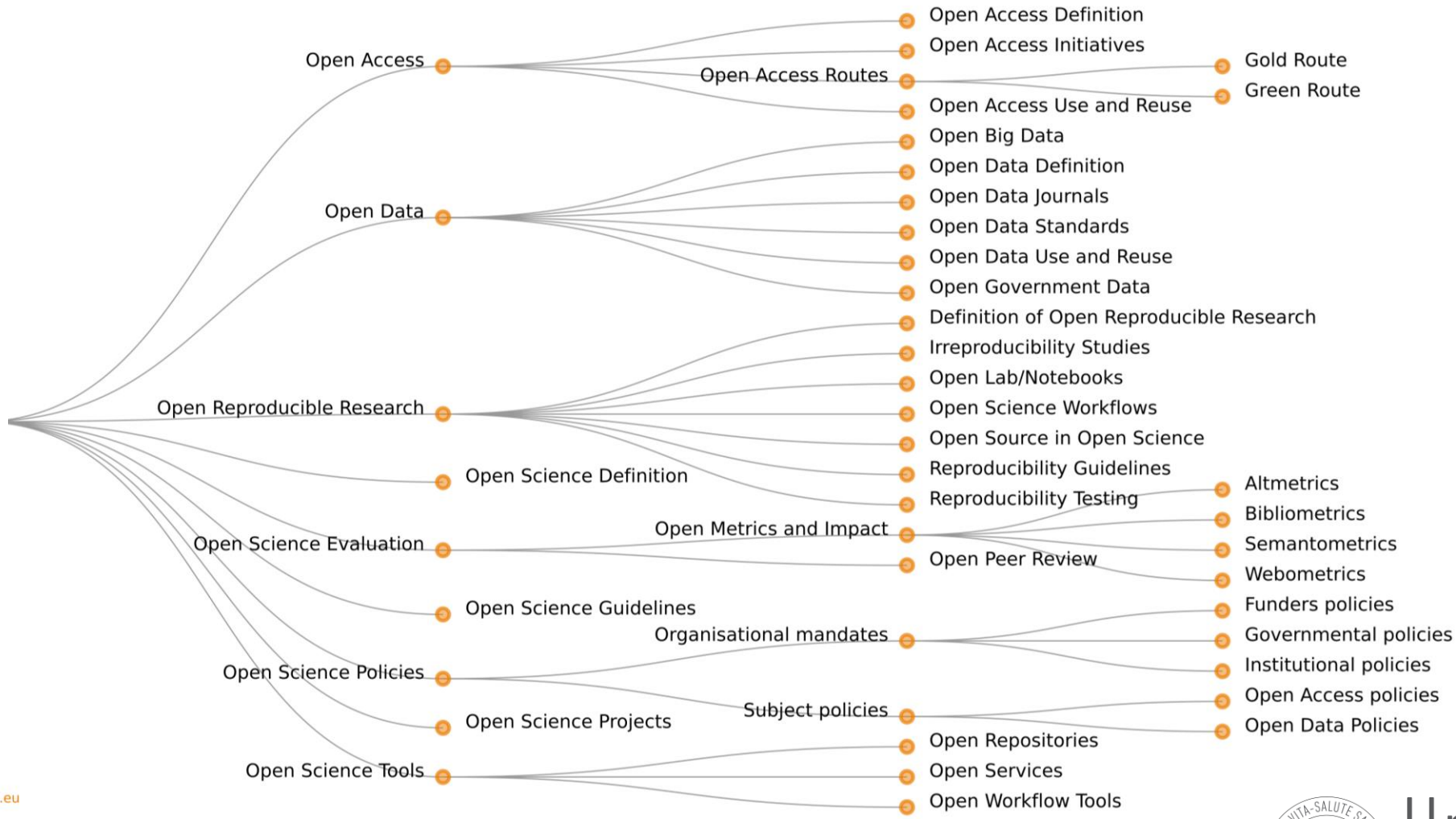
- Open Science opens up scientific processes and products from all levels to everyone.

Source: OpenAire website

Università Vita-Salute San Raffaele

Open Science: an **umbrella** term

OPEN SCIENCE



FOSTER
www.fosteropenscience.eu

Università Vita-Salute San Raffaele

Source: [Foster open science taxonomy]

Open Science: why we **need** it?

Why Open Science?

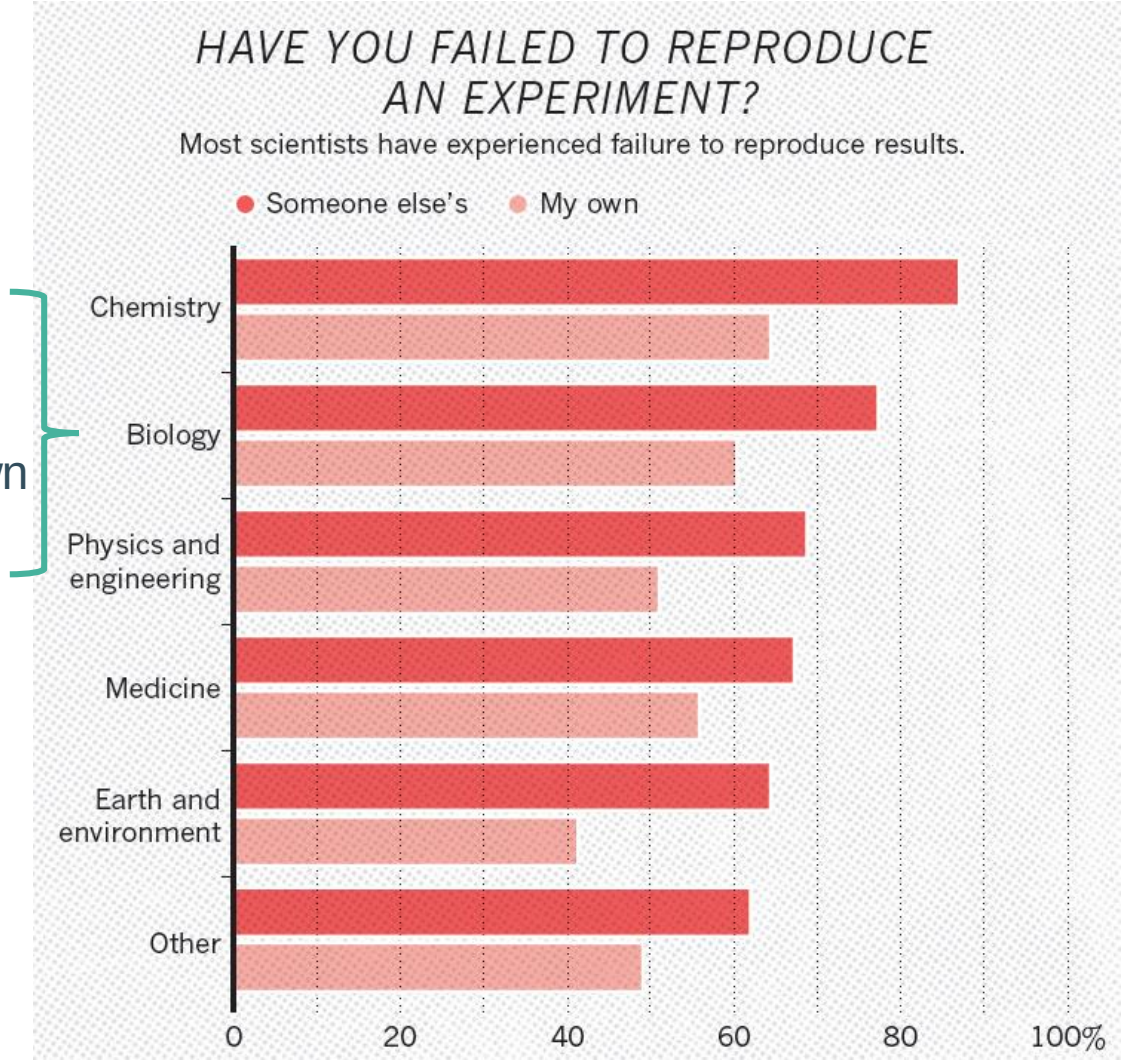
- **Reproducibility Crisis**
- **Fragility of research data**
- **…**

WE HAVE A PROBLEM!

Why Open Science: a **Reproducibility** crisis

Almost **80%** fail to reproduce others' results

60% cannot reproduce their own experiments!

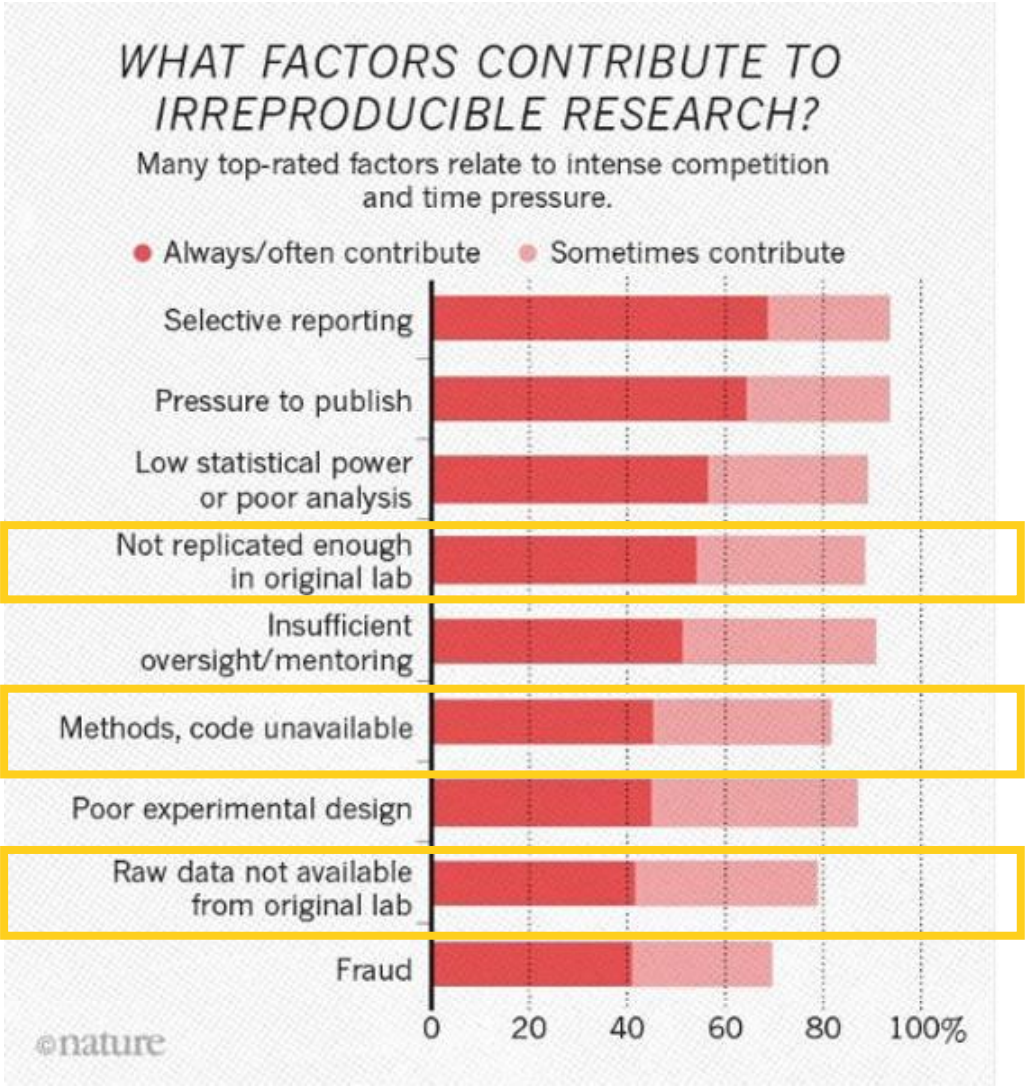


Nature's survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research...

Baker, M., (2016); https://doi.org/10.1038/533452a

UniSR
Università Vita-Salute San Raffaele

Why Open Science: a **Reproducibility** crisis



Modified from: Baker, M., (2016); https://doi.org/10.1038/533452a

"Lack of responsiveness from original authors"

Rodgers, P., Collings, A. (2021). DOI: 10.7554/eLife.75830

UniSR
Università Vita-Salute San Raffaele

Why Open Science: data are **fragile**

Current Biology 24, 94–97, January 6, 2014 ©2014 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2013.11.014

Report

The Availability of Research Data Declines Rapidly with Article Age

Timothy H. Vines,[1,2,*] Arianne Y.K. Albert,[3] Rose L. Andrew,[1] Florence Débarre,[1,4] Dan G. Bock,[1] Michelle T. Franklin,[1,5] Kimberly J. Gilbert,[1] Jean-Sébastien Moore,[1,6] Sébastien Renaut,[1] and Diana J. Rennison[1]

sets (23%) were confirmed as extant. Table 1 provides a breakdown of the data by year.

We used logistic regression to formally investigate the relationships between the age of the paper and (1) the probability

- We examined the availability of data from 516 studies between 2 and 22 years old
- The odds of a data set being reported as extant fell by 17% per year
- Broken e-mails and obsolete storage devices were the main obstacles to data sharing
- Policies mandating data archiving at publication are clearly needed

Datasets '*available upon request*' are often **NOT** available.

https://doi.org/10.1016/j.cub.2013.11.014

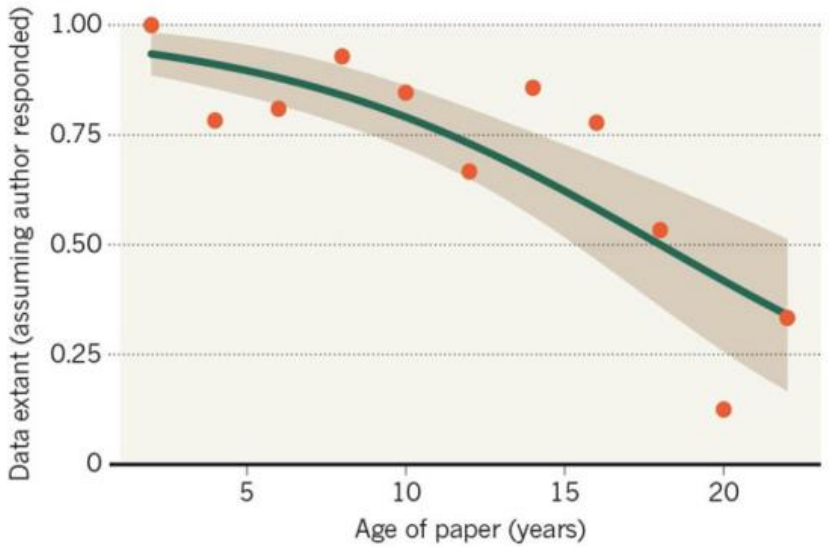
Scientists losing data at a rapid rate

Decline can mean 80% of data are unavailable after 20 years.

Elizabeth Gibney & Richard Van Noorden

MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.



"Decline can mean **80%** of data are **unavailable** after **20 years**"

https://doi.org/10.1038/nature.2013.14416

UniSR
Università Vita-Salute San Raffaele

Why Open Science: solutions and benefits

The problem	Solutions from Open Science
Reproducibility Crisis	**Detailed documentation on data analysis procedures** **Data Management Plan**
Data Fragility	**Best practices on data preservation and sharing** **FAIR principles**

NOT ONLY SOLUTIONS… BUT ALSO BENEFITS

Open Science: **benefits** for all stakeholders

Researchers

- greater visibility & reach
- increased efficiency
- funding
- collaboration/networking

Funders

- increased visibility & reuse of funded research
- greater funding impact
- greater ROI

General Public

- faster knowledge transfer
- increased understanding and expertise
- promoting engagement in science & research

Organisations/ NGOs

- enhanced access to research
- more effective advocacy/lobbying

National Governments

- evidence-informed policy
- promoting Human Rights and democracy

Source: OpenAire website

Università Vita-Salute San Raffaele

Open Science:
Funders' requirements

The EC has made a strong choice in favour of Open Science

Open Science: requirements in Horizon Europe

What?	How?	Mandatory in all calls/recommended
Open access to research outputs through deposition in trusted repositories	• Open access to publications • Open access to data • Open access to software, models, algorithms, workflows etc.	• **Mandatory** for peer-reviewed publications • **Mandatory** for research data **but** with exceptions ('as open as possible…') • Recommended for other research outputs
Research output management	Manage responsibly in line with FAIR; Data management plan (DMP)	**Mandatory**
Measures to ensure reproduciblity of research outputs	Information on outputs/tools/instruments and access to data/results for validation of publications	**Mandatory**
Early and open sharing of research	Preregistration, registered reports, preprints etc.	Recommended
Participation in open peer-review	Publishing in open peer-reviewed journals or platforms	Recommended
Involving all relevant knowledge actors	Involvement of citizens, civil society and end-users in co-creation of content (e.g. crowd-sourcing, etc.)	Recommended

Adapted from: Bottaro, S. (2021); https://zenodo.org/record/5844248

UniSR
Università Vita-Salute San Raffaele

Open Science: requirements in Horizon Europe

PUBLICATIONS

- Deposition + immediate open access (i.e., at the same time as the first publication) through a trusted repository using specific open licenses;

- Only publication fees in full open access venues are eligible for reimbursement

RESEARCH DATA

Before submitting, check that the journal's policies are compatible with the funders' requirements

- Establish + regularly update data management plan ('DMP'), by month 6

Source: Horizon Europe Annotated Model Grant Agreement – Article 17

Università Vita-Salute San Raffaele

Open Science: requirements in **Horizon Europe**

PUBLICATIONS

- Deposition + immediate open access (i.e., at the same time as the first publication) through a trusted repository using specific open licenses;

- Only publication fees in full open access venues are eligible for reimbursement

RESEARCH DATA

- Deposition in a trusted repository ensuring open access (as open as possible as closed as necessary), as soon as possible, using specific open licences;

- Establish + regularly update data management plan ('DMP'), by month 6

Source: Horizon Europe Annotated Model Grant Agreement – Article 17

Università Vita-Salute San Raffaele

Open Science:
Funders' requirements

… More are expected to follow, nationally and internationally

Open Science:
Publishers' requirements

CAREERS COMMENTARY JOURNALS | COVID-19 Science

Data and Code Deposition

As outlined in the TOP guidelines above, the *Science* Journals generally require all data underlying the results in published papers to be publicly and immediately available. Post-publication embargoes are not permitted, nor are stipulations for readers to contact the authors (rare exceptions involving third-party datasets must be discussed with the editor prior to publication and

nature portfolio View all journals Search

nature > nature portfolio > editorial policies > reporting standards and availability of data, materials, code and protocols

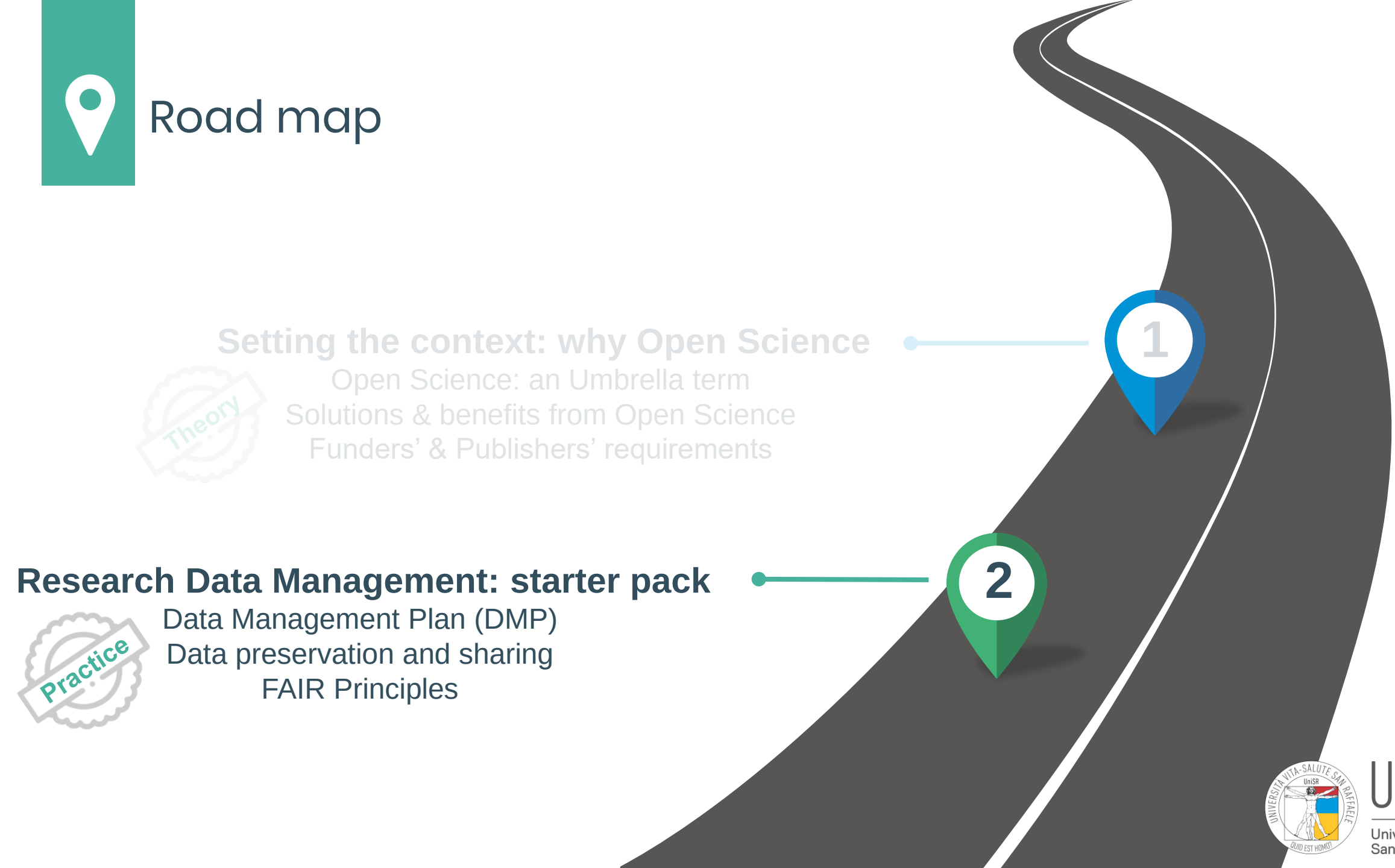
Editorial policies
Authorship
Competing interests
Confidentiality
Plagiarism and duplicate publication
Preprints & Conference Proceedings
Image integrity and standards
Peer Review

Reporting standards and availability of data, materials, code and protocols

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. A condition of publication in a Nature Portfolio journal is that **authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications.** Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission. Any restrictions must also be disclosed in the submitted manuscript.

UniSR
Università Vita-Salute San Raffaele

Road map



Setting the context: why Open Science

Open Science: an Umbrella term
Solutions & benefits from Open Science
Funders' & Publishers' requirements

Theory

1

Research Data Management: starter pack

Data Management Plan (DMP)
Data preservation and sharing
FAIR Principles

Practice

2

UniSR
Università Vita-Salute San Raffaele

Research Data Management (RDM): data **lifecycle**

"Research data management (RDM) concerns the organisation of data, from its entry to the research cycle through to the dissemination and archiving of valuable results. It aims to ensure reliable verification of results, and permits new and innovative research built on existing information".

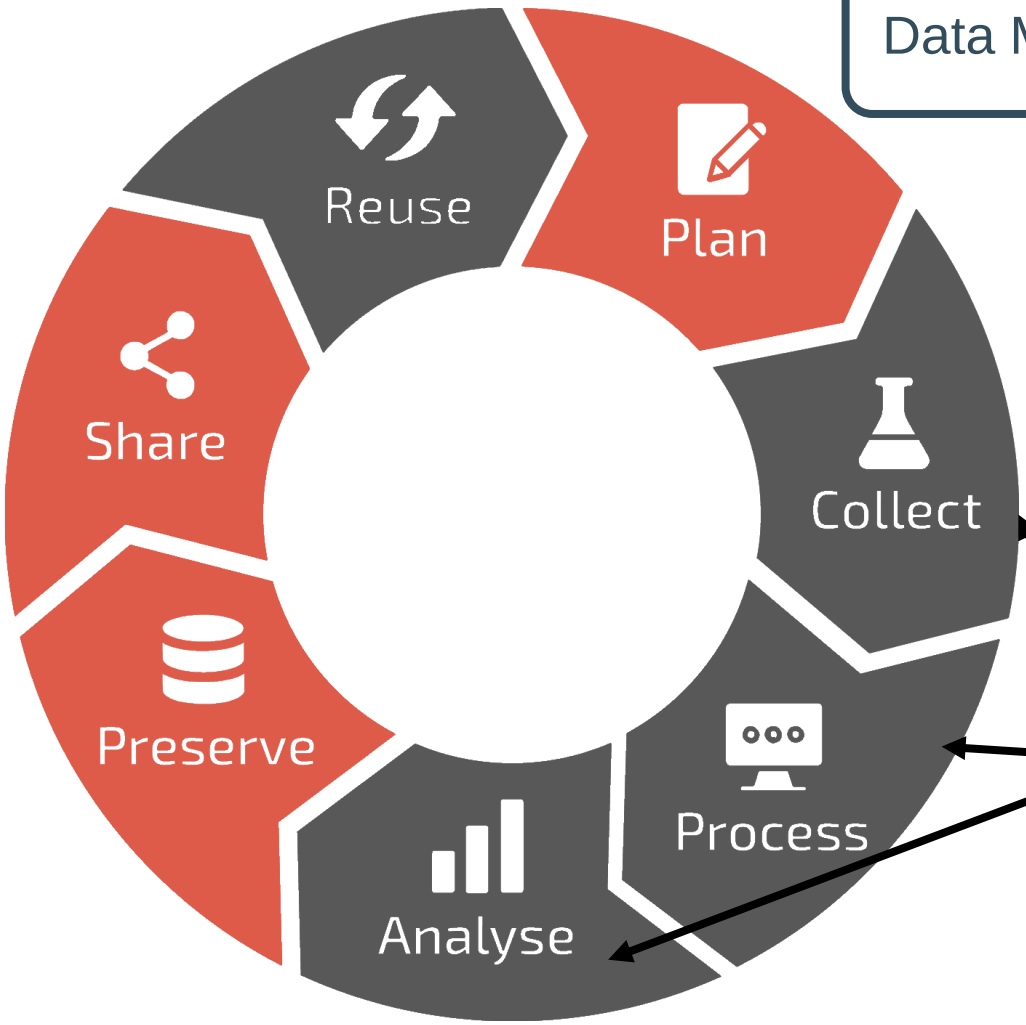
Whyte, A., Tedds, J. (2011); https://www.dcc.ac.uk/guidance/briefing-papers/making-case-rdm

Image by: ELIXIR (2021) Research Data Management Kit

UniSR
Università Vita-Salute San Raffaele

Research Data Management (RDM): data **lifecycle**

Data Management Plan (DMP)



Data preservation and sharing

Active data

Image modified from: ELIXIR (2021) Research Data Management Kit

Università Vita-Salute San Raffaele

Research Data Management (RDM): **Data management Plan** (DMP)

Image modified from: [ELIXIR (2021) Research Data Management Kit]

UniSR
Università Vita-Salute San Raffaele

Data management plan (DMP): **definition**

DMP is a **structured document** that outlines how you will **manage** research data both **during** a research project and **after** the project is completed

The goal is to help you consider the data lifecycle and their management BEFORE their generation

Source: Horizon Europe Programme Guide

Università Vita-Salute San Raffaele

Data management plan (DMP): **definition**

DMP is a **structured document** that outlines how you will **manage** research data both **during** a research project and **after** the project is completed

For example, have you thought about:

- The type and format of the data you will generate or re-use;
- How you will organize the data and the standards you will use;
- How to preserve data (e.g., backup; where data will be stored, etc);
- How to share data (including limitations due to privacy or IP issues);
- Whether you have allocated some resources/budget for data management.

Source: Horizon Europe Programme Guide

UniSR
Università Vita-Salute San Raffaele

Data management plan (DMP): **definition**

DMP is a **structured document** that outlines how you will **manage** research data both **during** a research project and **after** the project is completed

For example, have you thought about:

- The type and format of the data you will generate or re-use;
- How you will organize the data and the standards you will use;
- How to preserve data (e.g., backup; where data will be stored, etc);
- How to share data (including limitations due to privacy or IP issues);
- **Whether you have allocated some resources/budget for data management.**

> nature
>
> **Invest 5% of research funds in ensuring data are reusable**
>
> 
>
> It is irresponsible to support research but not data stewardship, says Barend Mons.
>
> https://doi.org/10.1038/d41586-020-00505-7

UniSR
Università Vita-Salute San Raffaele

Data management plan (DMP): **definition**

DMP is a **structured document** that outlines how you will **manage** research data both **during** a research project and **after** the project is completed;

DMP is a **roadmap** that you will use not to get lost in your own data (your project = journey).

As all plans, it may be changed. The DMP is a **living document** that will be revised during the project to reflect any change of direction.

Data management plan (DMP): the **reasons** why

- ✓ **It is now required by most funding bodies (e.g., Horizon Europe);**
- ✓ More efficient workflow: planning saves time, money and resources (less stress);
- ✓ Better management of all your collaborations;
- ✓ Ensure that data are preserved in the long term;
- ✓ Ensure, in advance, that you are compliant with personal data protection laws (e.g. GDPR).

UniSR
Università Vita-Salute San Raffaele

Data management plan (DMP)

What you do

Collect all relevant information

Coordinate with partners (in case of consortium projects)

Write the draft

What Open Science Team does

Raise awareness on the issues that must be described in the document

Provide clarifications and assistance

Assess conflicts and issues on IP and privacy

Identify weak points

Review the final draft

Research Data Management (RDM): Data **preservation** and **sharing**

Image modified from: [ELIXIR (2021) Research Data Management Kit]

UniSR
Università Vita-Salute San Raffaele

Data preservation and sharing

Some reactions to “Sharing” Requirements:

- It would take me 5 years to find all my data!
- The PhD/postdoc who had the data left the lab
- Nobody will understand my data
- People can just ask for my data when they need it.

BUT…

What if someone asks you for data supporting your publication, 5/10 years after publication?

BEST PRACTICE: DEPOSIT DATA IN A REPOSITORY

Adapted from: Kurapati, S. (2021); https://zenodo.org/record/4769844

UniSR
Università Vita-Salute San Raffaele

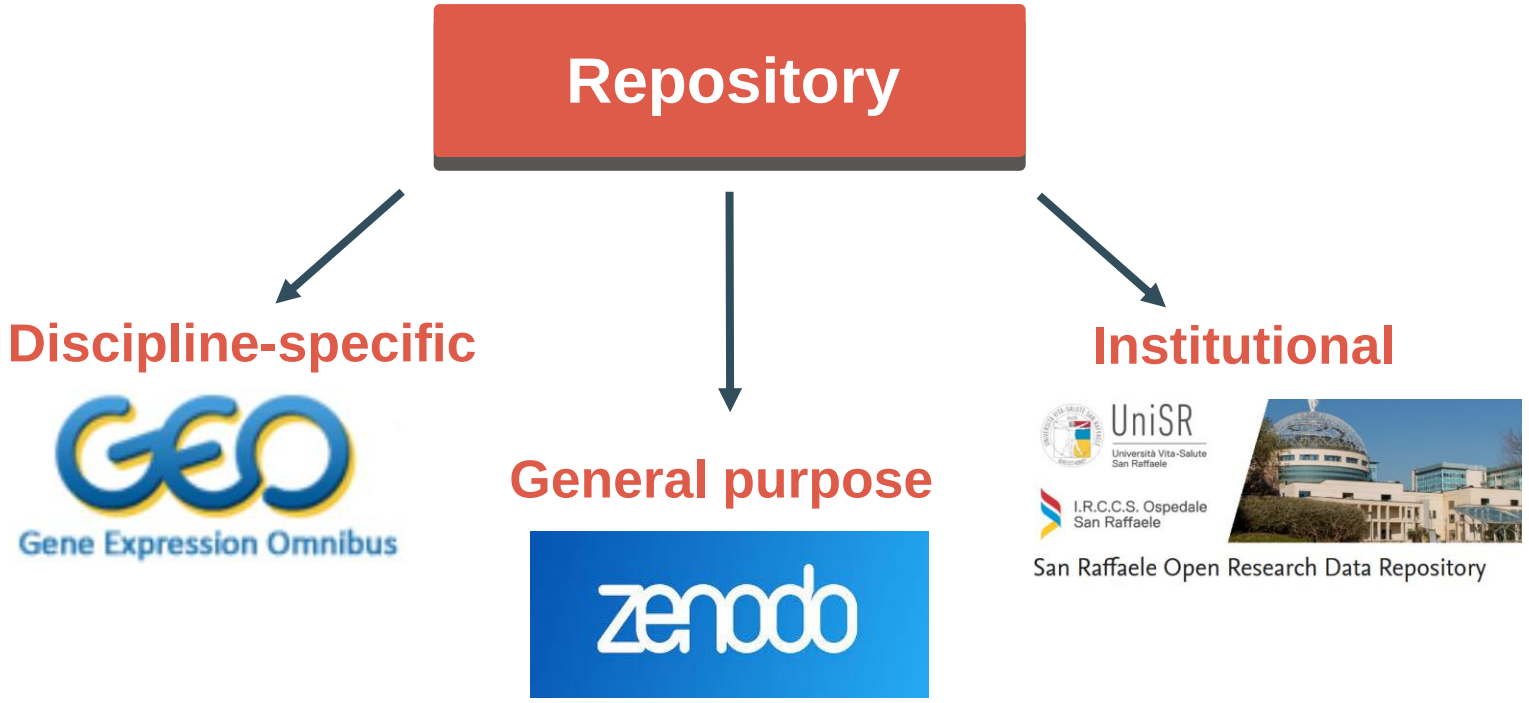
Data preservation and sharing: definition of **Repository**

An online **archive** where research outputs (paper, data, code, software) can be stored and shared:

- Persistent identifiers (**DOI**) to enable referencing and citation
- **Versioning** to keep track of every change to data over time
- **Open** or **restricted access** to contents
- **Long-term preservation** of deposited material

Source: [Horizon Europe Annotated Model Grant Agreement – Article 17]

Data preservation and sharing: Repository **types**



Data preservation and sharing: Repository **types**

Where do I deposit my (**open**) Data?

Adapted from: Pasquale V., Morelli M. J (2020); https://zenodo.org/record/3936442

UniSR
Università Vita-Salute San Raffaele

Data preservation and sharing

In a Repository, data become **Findable**

BUT…

this is only the **first step** to re(using) data

What are the other steps?

The **FAIR** principles

Findable Accessible Interoperable Reusable

A set of principles to enhance the value of all digital resources and its **reuse** by *humans* and *machines*

Wilkinson, M., et al. (2016); https://doi.org/10.1038/sdata.2016.18

UniSR
Università Vita-Salute San Raffaele

The FAIR principles

Findable

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 it describes

F4. (meta)data are registered or indexed in a searchable resource

Wilkinson, M., et al., (2016); https://doi.org/10.1038/sdata.2016.18

- ✓ Deposit data in a repository
- ✓ Machine-actionable comprehensive metadata

Data are easy to find

UniSR
Università Vita-Salute San Raffaele

The FAIR principles

Accessible

A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

A2. metadata are accessible, even when the data are no longer available

Wilkinson, M., et al., (2016); https://doi.org/10.1038/sdata.2016.18

- ✓ Describe exact conditions of accessibility: who, when, how
- ✓ Use standard protocols of access

Access conditions must be clear:
FAIR ≠ OPEN

The FAIR principles

Interoperable

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

I4. Data should be readable without proprietary formats

Modified from: Wilkinson, M., et al., (2016); https://doi.org/10.1038/sdata.2016.18

- ✓ Use international or community standard (e.g. Dublin Core)
- ✓ Convert proprietary formats (.xlsx → .csv)

Data can be read and integrated with other data

The FAIR principles

Reusable

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

Wilkinson, M., et al., (2016); https://doi.org/10.1038/sdata.2016.18

- ✓ Choose an appropriate license to distribute the data
- ✓ Metadata should be detailed and well-described
- ✓ Add methods/codes/SOPs/procedures to metadata

Data without Metadata are NOT reusable

UniSR
Università Vita-Salute San Raffaele

Take-home message

THIS IS HERE TO STAY

Adapted from: Kurapati, S. (2021); https://zenodo.org/record/5813531

UniSR
Università Vita-Salute San Raffaele

Need a hand?

The **Open Science Team** is here for you!

open.science@unisr.it

References

- **The reproducibility crisis in research:**
  - https://www.nature.com/articles/533452a: in this 2016 Nature paper, more than 1500 researchers were surveyed to ask their opinions on whether or not there is a reproducibility crisis in research;
  - https://elifesciences.org/articles/75830: in 2021, a US $2-million eight-year attempt to replicate influential cancer studies found that research in cancer biology is not as reproducible as it should be.
- **The fragility of research data:**
  - https://www.sciencedirect.com/science/article/pii/S0960982213014000:in this 2013 paper the authors requested data from a relatively homogenous set of 516 articles published between 2 and 22 years before, and found that availability of the data was strongly affected by article age;
  - https://www.nature.com/articles/nature.2013.14416: in this Nature editorial of 2013, the authors suggest that decline can mean 80% of data are unavailable after 20 years.
- **Open science practices in the Horizon Europe Funding Programme:**
  - Annotated Model Grant Agreement – Article 17: this document contains detailed annotations on all the Open science provisions in the grant agreement of the Horizon Europe (the relevant section starts at page 151).
- **Data Management Plan:**
  - https://www.nature.com/articles/d41586-020-00505-7: in this brief Nature editorial, Prof. Barend Mons suggests that, on average, 5% of overall research costs should go towards data stewardship.
- **The FAIR principles:**
  - https://www.nature.com/articles/sdata201618: the original article, published in 2016, which provides guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of research data.