

All about FAIR principles

University of Pisa - Phd course 2022

Gina Pavone, CNR-ISTI  0000-0003-0087-2151

module 1 - 30 May 2022

10.5281/zenodo.6597239



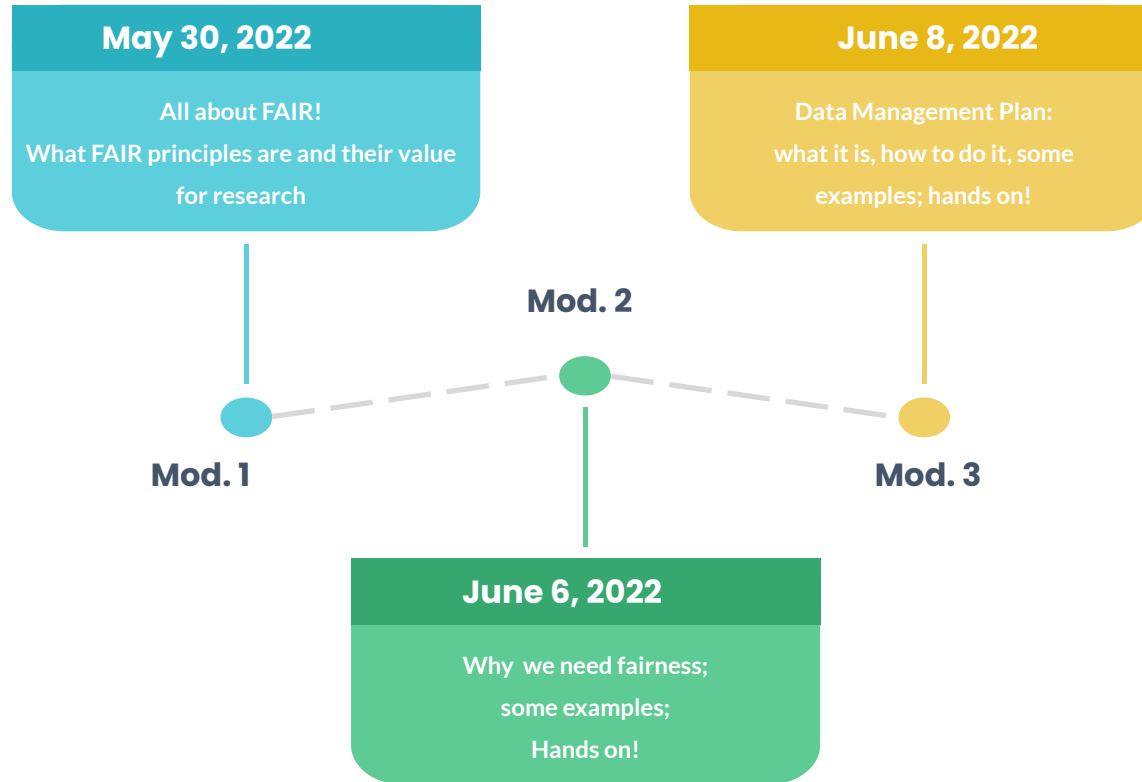
Gina Pavone

- Research fellow at the Institute of Information Science and Technologies of the Italian National Research Council in Pisa, Italy.
- Research focus: Open Science and Open Access; Research Data Management
- OpenAIRE National Open Access Desk (NOAD) for Italy
- Coordinator of the editorial board of open-science.it website
- My background: data journalism



FAIR data and DMP

COURSE OUTLINE



Today's agenda

University of Pisa
PhD students





Let's say that...a researcher needs to find datasets containing data about proteins that are activated in specific tissues and combine these data with information of which genes are involved in the production of such proteins.

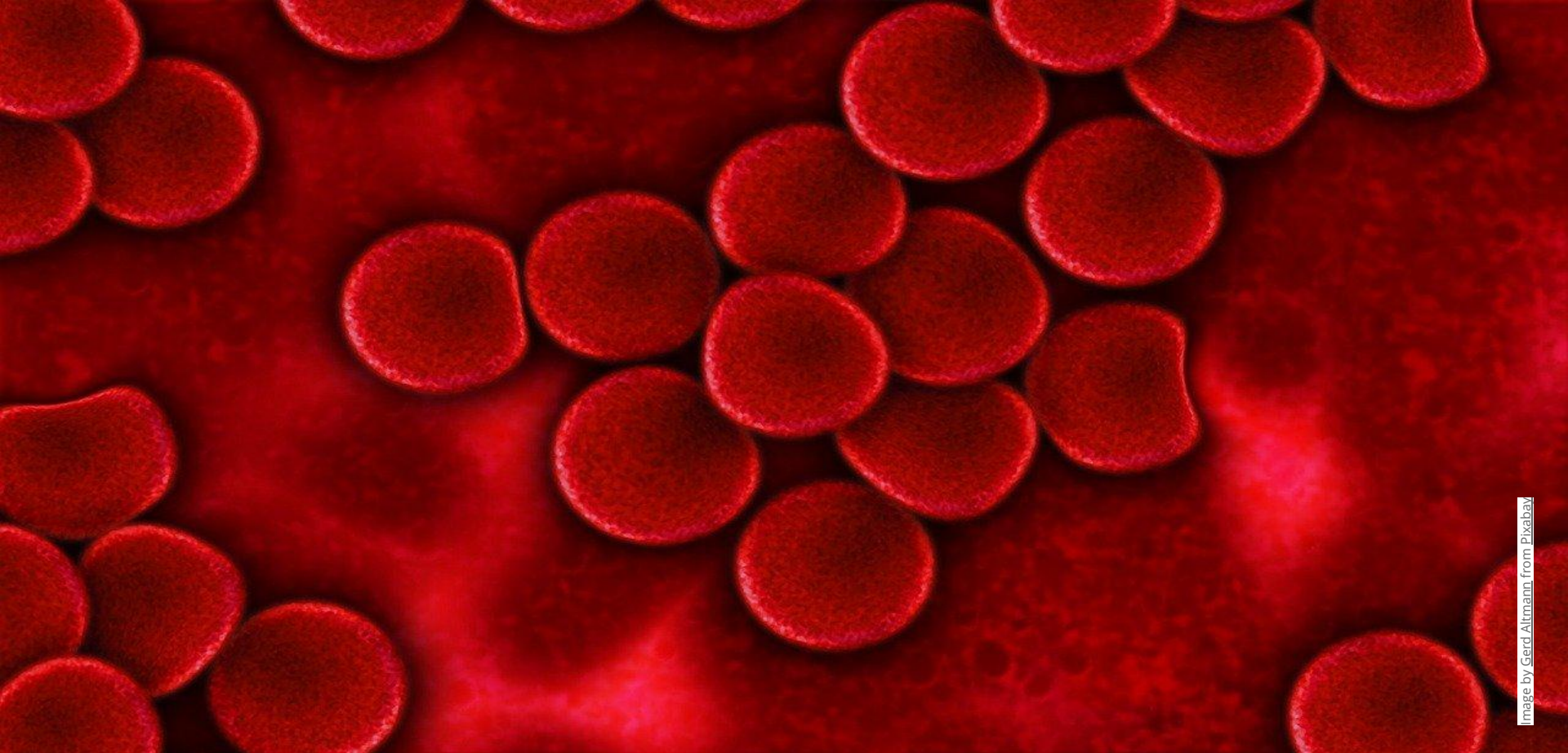
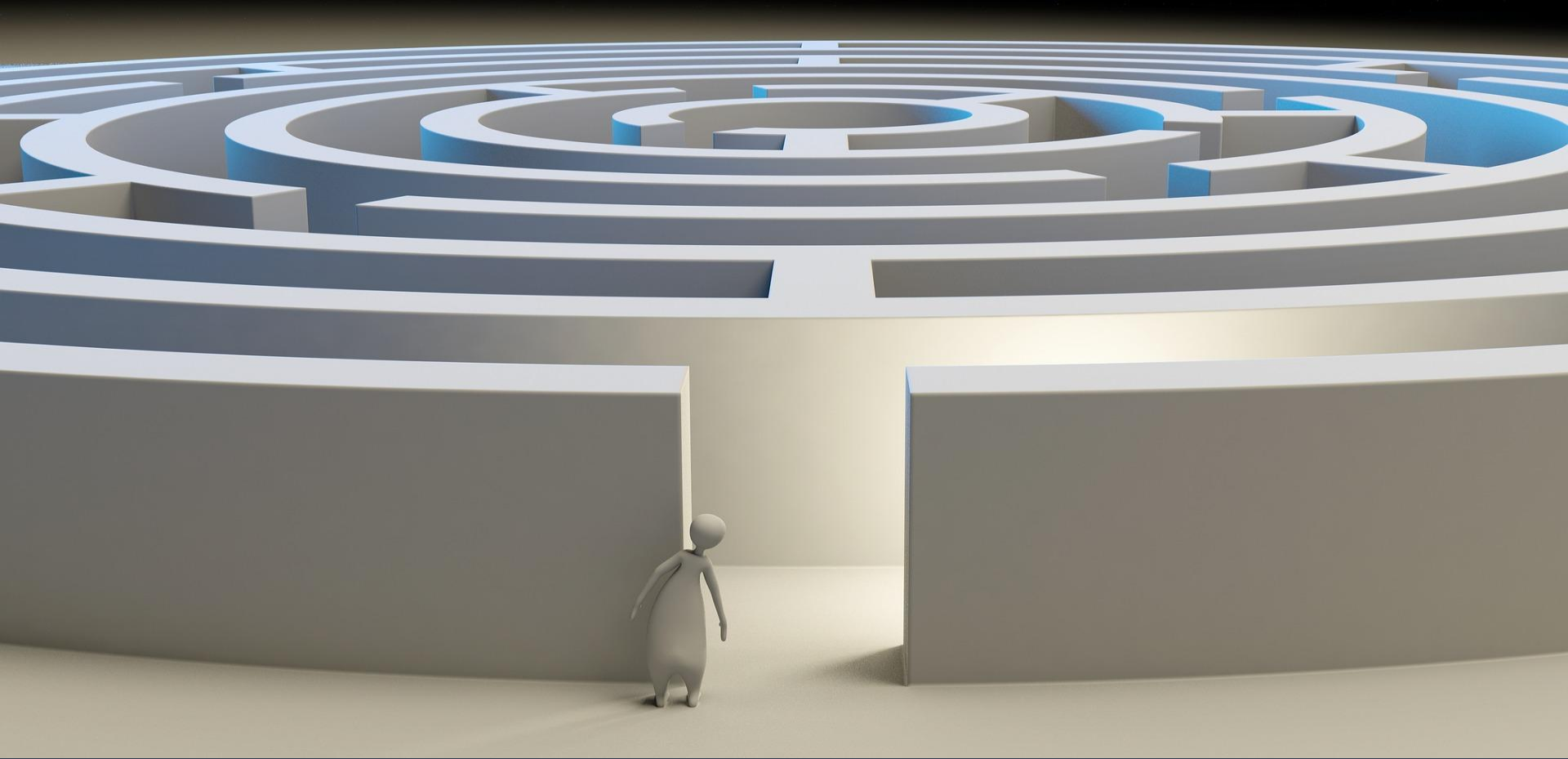


Image by Gerd Altmann from Pixabay

Or say....another researcher needs to know which biobanks carry a given type of biosample (e.g., blood samples) from patients possessing a specific phenotype (e.g., Alzheimer's disease) taken from a patient registry whose onset age was lower than 45 year-old

<https://github.com/FAIRDataTeam/FAIRDataPoint/wiki/FAIR-Data-Point-Specification#usage-scenarios>



These data users need to use a straightforward search application that allows them to find the required information.

<https://github.com/FAIRDataTeam/FAIRDataPoint/wiki/FAIR-Data-Point-Specification#usage-scenarios>

And it's still not over...

What if...you are that researcher and you do not know if you have the permission to reuse that information?

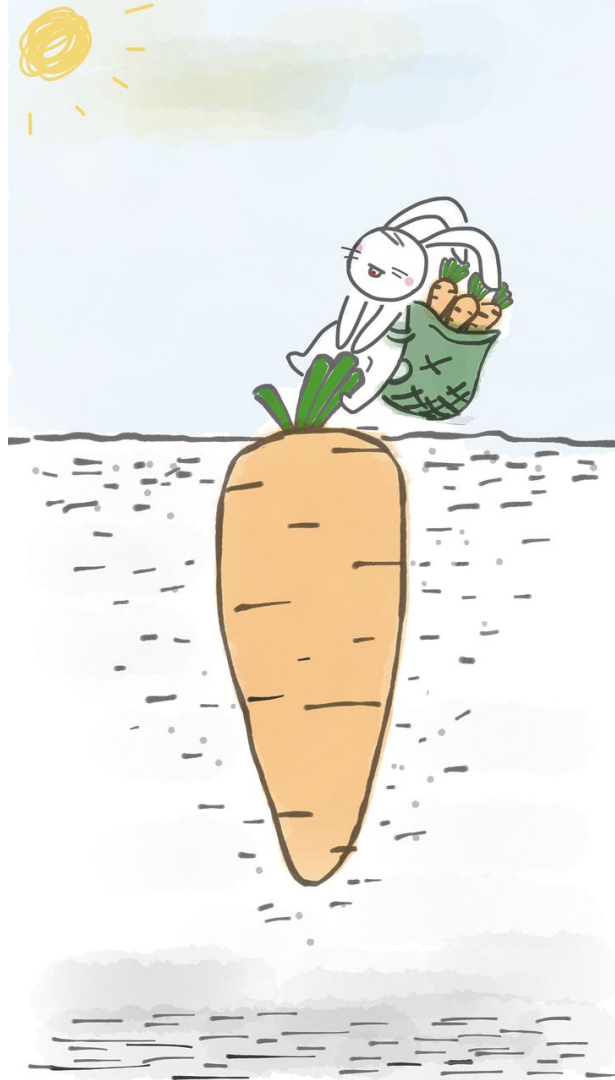
You will have to check licences.

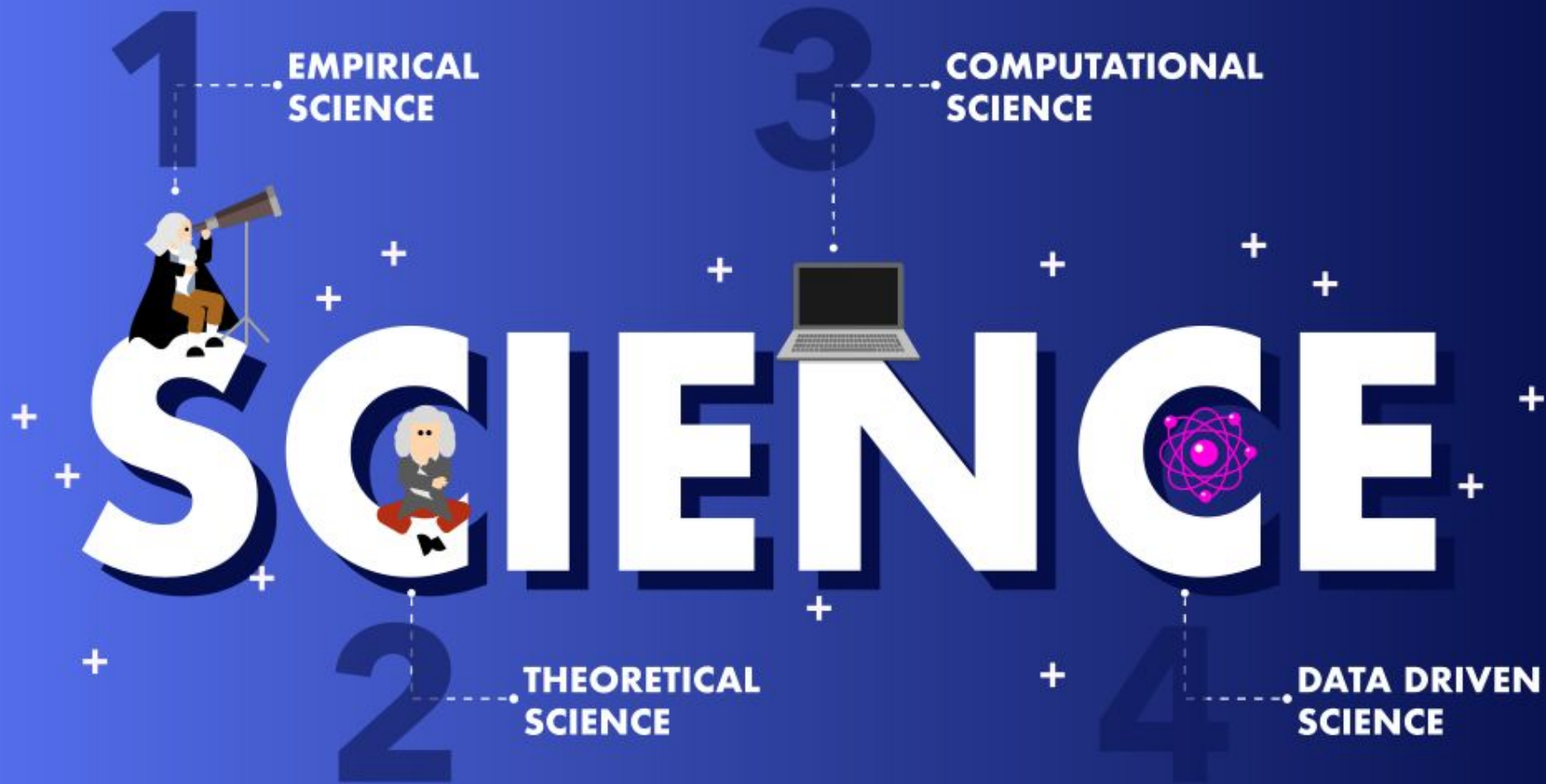
And then what if...the data you need is stored in a proprietary format and you do not have the software needed to access them?

You may not be able to use the data or you may have to find funds

And you will need to integrate them, combining different datasets.

And for this the data format from the datasets should be using a common representation technology that facilitates data integration.





What is data?

Data or it didn't happen!

Facts, observations or experiences on which an argument or theory is constructed or tested.

Data are information!
(in a variety of forms and formats)

UCL Research Data Policy

<https://www.ucl.ac.uk/library/research-support/research-data-management>

Types of research data

There is a huge variety of data types. Research data can be classified in different ways, for example based on their:

Content: numerical, textual, audiovisual, multimedia...

Format: spreadsheets, databases, images, maps, audio files, (un)structured text...

Mode of data collection: experimental, observational, simulation, derived/compiled from other sources

Digital (born-digital or digitized) or non-digital nature (e.g. paper surveys, notes...)

Primary (generated by the researcher for a particular research purpose or project) or **secondary** nature (originally created by someone else for another purpose)

Raw or **processed** nature

<https://www.ugent.be/en/research/datamanagement/why/rdm-explained.htm>



Image by [Gerd Altmann](#) from [Pixabay](#)

Data are first-class research objects

Check
Validation
Follow-ups
New research questions
Teaching
Business applications

...



PUBLICATIONS AND DATA

We need to Find the right match

We want that our contents will be read (and cited).

And we want to (easily) find/access and reuse scientific literature produced by others.

A sort of supply and demand matching!



We need to Access



We, and our
machines, need to
Interoperate with
the resource



We also want to
Reuse the resource
(or someone else
reuse ours)



The FAIR Principles



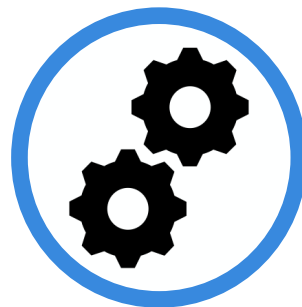
Findable

Others can easily discover your data



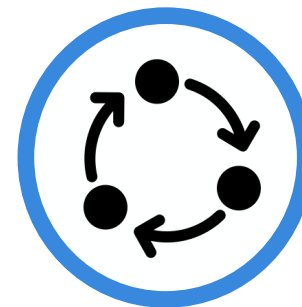
Accessible

It is clear who, when and how can access your data (does not mean open)



Interoperable

Your data can be integrated with other data and/or they can be easily used and read by machines.



Reusable

Your data can be reused by others in new research

FAIR is an evolution of the open data movement

More nuanced



More machines



OPEN

Comment: The FAIR Guiding Principles for scientific data management and stewardship

SUBJECT CATEGORIES

- » Research data
- » Publication characteristics

Mark D. Wilkinson et al.*

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those willing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of ‘long-term care’ of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes ‘good data management’ is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to ‘data’ in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other’s data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mom@qzhs.nl).

*A full list of authors and their affiliations appears at the end of the paper.

- FAIR indicate a list of principles that can help you in making your data ready for Open Science
- They are **principles**, not standards!
- They were designed to enable optimal use of research data and methods
- A group of different experts designed the FAIR principles between 2014 and 2016
- They identified a set of 15 principles

The experts behind FAIR principles:

- **researchers** wanting to share, get credit, and reuse each other's data and interpretations;
- **professional data publishers** offering their services;
- **software and tool-builders** providing data analysis and processing services such as reusable workflows;
- **funding agencies** (private and public) increasingly concerned with long-term data stewardship;
- **data science community** mining, integrating and analysing new and existing data to advance discovery.

By applying the FAIR principles

- It will be easier for you to produce high quality data
- You will maximise the impact of your research
- You will improve the recognition within and behind your research community
- You will be compliant with a growing number of funding agencies requirements (e.g. European Commission)

Importantly, it is our intent that the principles apply not only to ‘data’ in the conventional sense, but also to the algorithms, tools, and workflows that led to that data.

All scholarly digital research objects—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

Wilkinson et al.

No one size fits all

The application of the FAIR principles strongly depends on the specific **discipline** and on the way the **single researcher** works



Elements you need to make your data FAIR

F - Findable

Findable

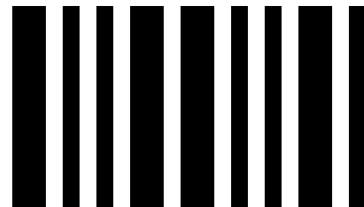
- The first step in (re)using data is to find them.
- Metadata and data should be easy to find for both humans and computers.
- Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the [FAIRification process](#).

F1: (Meta) data are assigned globally unique and persistent identifiers



Persistent Identifiers

- A **persistent identifier** (PI or PID) is a long-lasting reference to a document, file, web page, or other object.
- The term persistent identifier is usually used in the context of **digital objects** that are accessible over the Internet.
- Typically, such an identifier is not only persistent but **actionable**: you can plug it into a web browser and be taken to the identified source.
- It is like the barcode used on products...



Many types of PIDs

People - ORCID

Projects - RAiD www.raid.org.au

Digital objects - DOI

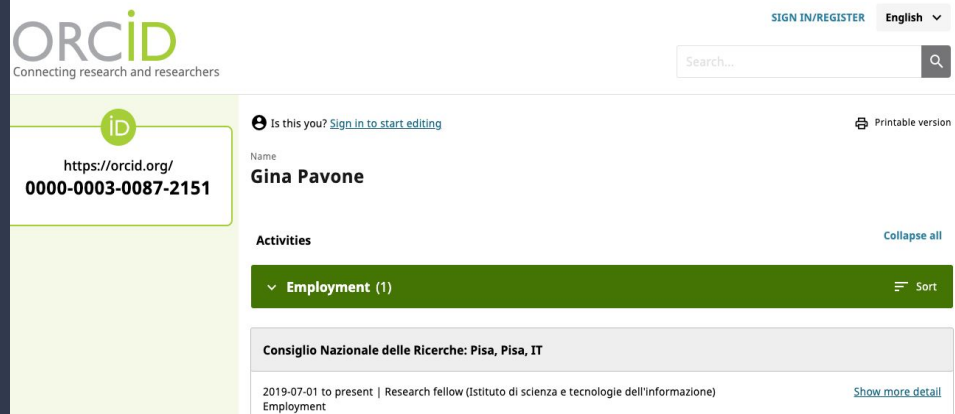
Physical samples IGSN - <https://www.igsn.org/>

Example services that supply globally unique and persistent identifiers

- Identifiers.org provides resolvable identifiers in the form of URIs and CURIEs: <http://identifiers.org>
- Universally unique identifier: https://en.wikipedia.org/wiki/Universally_unique_identifier
- Persistent URLs: <http://www.purlz.org>
- Digital Object Identifier: <http://www.doi.org>
- Archival Resource Key: <https://escholarship.org/uc/item/9p9863nc>
- Research Resource Identifiers: <https://scicrunch.org/resources>
- Identifiers for funding organisations (see F3 & R1): <https://www.crossref.org/services/funder-registry/>
- Identifiers for the world's research organisations (see F3 & R1): <https://www.grid.ac>

ORCID: do you have one? you should...

Open Researcher and Contributor ID is a nonproprietary alphanumeric code to uniquely identify scientific and other academic authors and contributors



The screenshot shows the ORCID profile page for Gina Pavone. At the top left is the ORCID logo with the tagline "Connecting research and researchers". To the right are links for "SIGN IN/REGISTER" and a language dropdown set to "English". A search bar is also present. The profile header includes a green "id" icon, the URL "https://orcid.org/0000-0003-0087-2151", and a question "Is this you? Sign in to start editing" with a "Printable version" link. The "Name" field displays "Gina Pavone". Under the "Activities" section, there is a green bar for "Employment (1)" with a "Sort" option. Below this, a specific employment entry is shown for "Consiglio Nazionale delle Ricerche: Pisa, Pisa, IT" from "2019-07-01 to present" as a "Research fellow (Istituto di scienza e tecnologie dell'informazione)", with a "Show more detail" link.

ORCID
Connecting research and researchers

SIGN IN/REGISTER English

Search...

id
https://orcid.org/
0000-0003-0087-2151

Is this you? [Sign in to start editing](#) Printable version

Name
Gina Pavone

Activities [Collapse all](#)

▼ Employment (1) [Sort](#)

Consiglio Nazionale delle Ricerche: Pisa, Pisa, IT

2019-07-01 to present | Research fellow (Istituto di scienza e tecnologie dell'informazione) [Show more detail](#)
Employment

DOI – Digital Object Identifier

- In computing, a **digital object identifier** (DOI) is a persistent identifier or handle used to identify objects uniquely, standardized by the International Organization for Standardization (ISO).
- A DOI aims to be **resolvable**, usually to some form of access to the information object to which the DOI refers.
- This is achieved by **binding the DOI to metadata** about the object, such as a URL, **indicating where** the object can be found
- a DOI differs from identifiers such as ISBNs and ISRCs which aim only to identify their referents uniquely

F1

Persistent identifiers remove ambiguity. They allow to find, cite and track (meta)data.

Globally unique (i.e., someone else could not reuse/reassign the same identifier without referring to your data)

It must be persistent. Registry services guarantee resolvability of that link into the future, at least to some degree.

F2: Data are described with rich metadata



What are metadata?

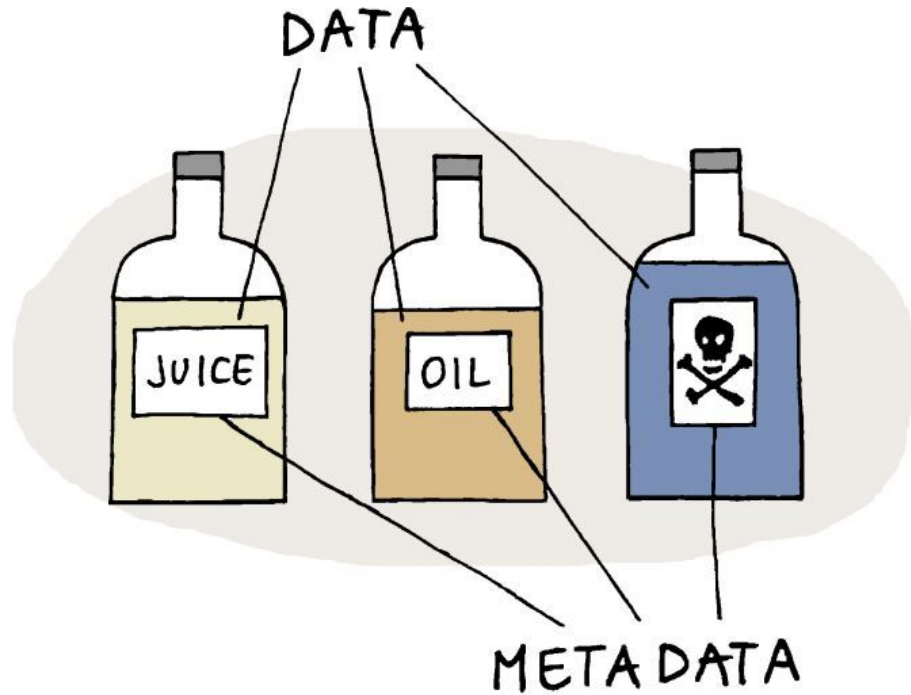
Literally “data about data.”

The information we create, store, and share to describe things

ie. information about the item’s creation, name, topic, features, and the like

They allows us to interact with these things to obtain the knowledge we need

Metadata is key to the functionality of the systems holding the content, enabling users to find items of interest, record essential information about them, and share that information with others.



PioloDataedo

The difference between data and metadata

Piotr Kononow through <https://twitter.com/aabella/status/1527533226574680064/photo/1>

Data and metadata

Usually there is a fuzzy boundary between metadata and the information it describes

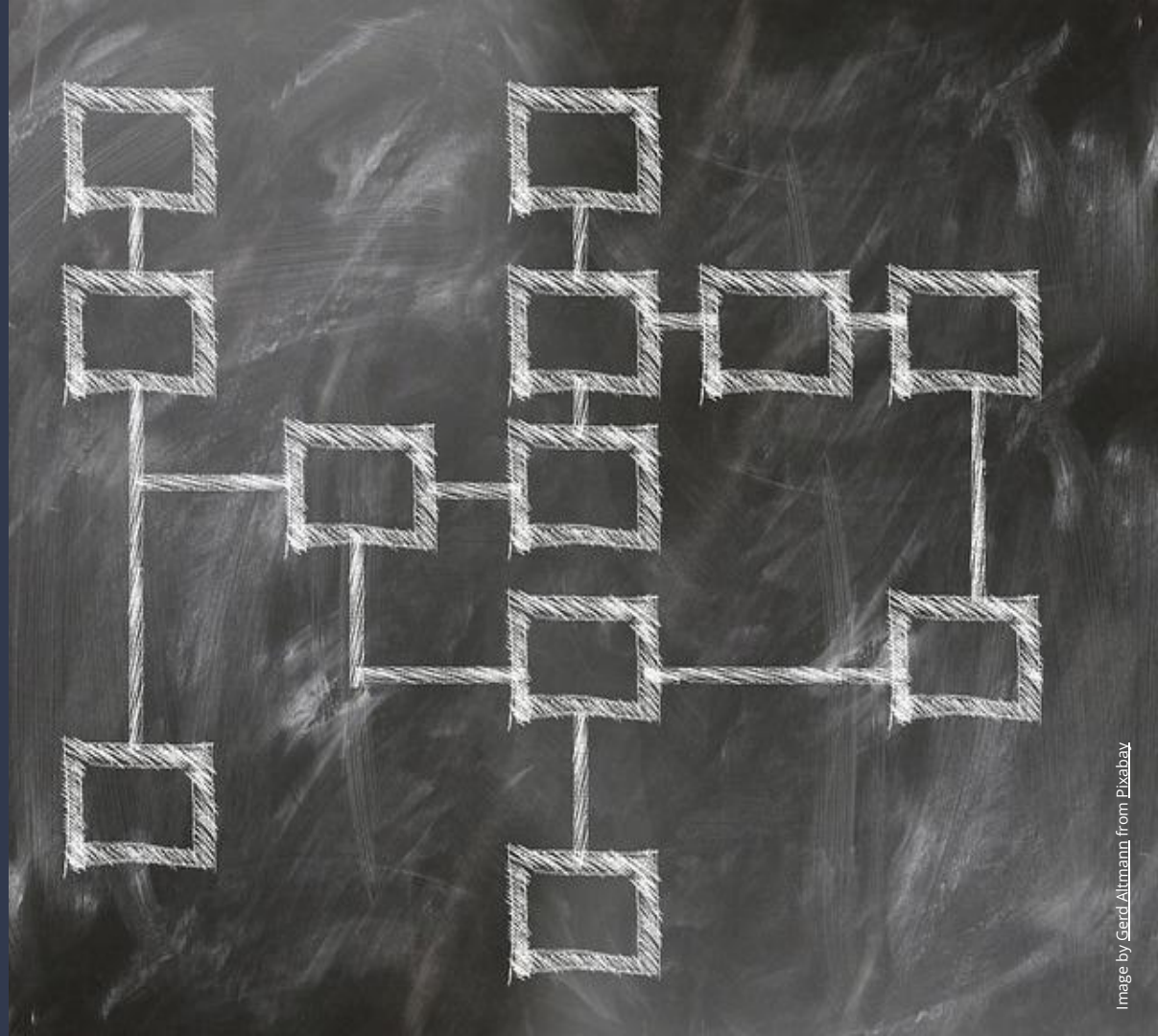
They do not necessarily are co-located



Structured information

Metadata has to be structured to some degree. The metadata is collected so that it can fulfill a useful purpose, and sorted into known categories. It is this notion of structure that turns raw information into actionable metadata

Metadata are collected and stored explanatory labels. Properties or elements are common terms for these labels



'Rich' information

'Intrinsic' metadata

e.g., the data captured automatically by machines)

'Contextual' metadata

e.g., the protocol used, with both keywords and links to a formal protocol document,

The measurement devices used (with both keywords and links to manufacturers)

The units of the captured data, the species involved

The 'object' that is the focus of the study (genes/proteins/whatever....)

Any other details about the experiment.



Main types of metadata

Descriptive metadata	For finding or understanding a resource
Administrative metadata: <ul style="list-style-type: none">- Technical metadata- Preservation metadata- Rights metadata	<ul style="list-style-type: none">- For decoding and rendering files- Long-term management of files- Intellectual property rights attached to content
Structural metadata	Relationships of parts of resources to one another

Types of metadata: some examples

Descriptive metadata	<ul style="list-style-type: none">- Title- Author- Subject- Genre- Publication date	Discovery Display Interoperability
Technical metadata	File type File size Creation date/time Compression scheme	Interoperability Digital object management Preservation
Preservation metadata	Checksum Preservation event	Interoperability Digital object management Preservation
Rights metadata	Copyright status License terms Rights holder	Interoperability Digital object management

Examples of metadata schema

- Dublin core - describing resources on the web
- Schema.org - commercial applications
- Crossref - research outputs
- Datacite metadata schema - describing research data
- Disciplinari metadata
 - Data Documentation Initiative
 - Darwin Core - biological sciences

keep in mind:

Other researchers in any field, or their computer, should be able to properly understand the nature of your dataset.

Be as generous as possible with your metadata!



Use your
discipline specific
standard!

You will spend less time
curating and interpreting
data and more time to
actually make science!

<https://rd-alliance.github.io/metadata-directory/>



F2

Generous and extensive information on the resource:

- its characteristics
- descriptive information about the context, quality and condition

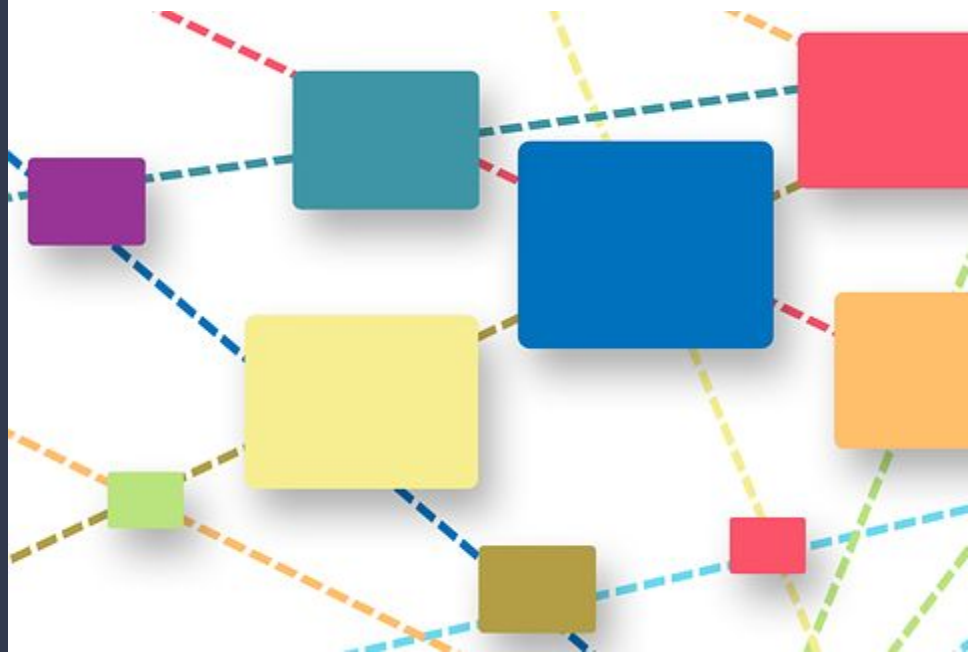
Someone should be able to find data based on the information provided by their metadata (even without the data's identifier)

F3: Metadata clearly and explicitly include the identifier of the data they describe



In other words:

The metadata and the data set they describe are separate files. The association between a metadata file and the data set is obvious thanks to the mention of the data set's PID in the metadata.



Not necessarily co-located



Resolvable



F3

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.

F4: (Meta)data are registered or indexed
in a searchable resource



In other words:

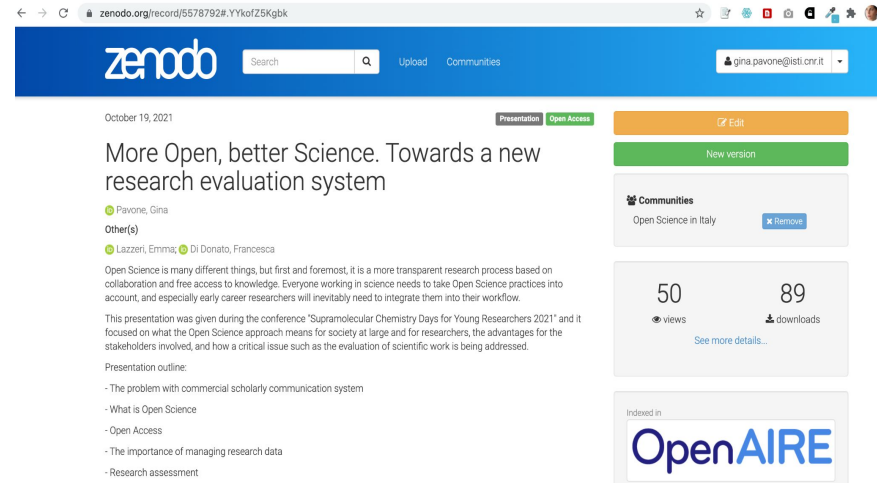
Metadata are used to build easily searchable indexes of data sets. These resources will allow to search for existing data sets similarly to searching for a book in a library.



Repository

An **open-access repository** or **open archive** is a digital platform that holds research output and provides free, immediate and permanent access to research results for anyone to use, download and distribute.

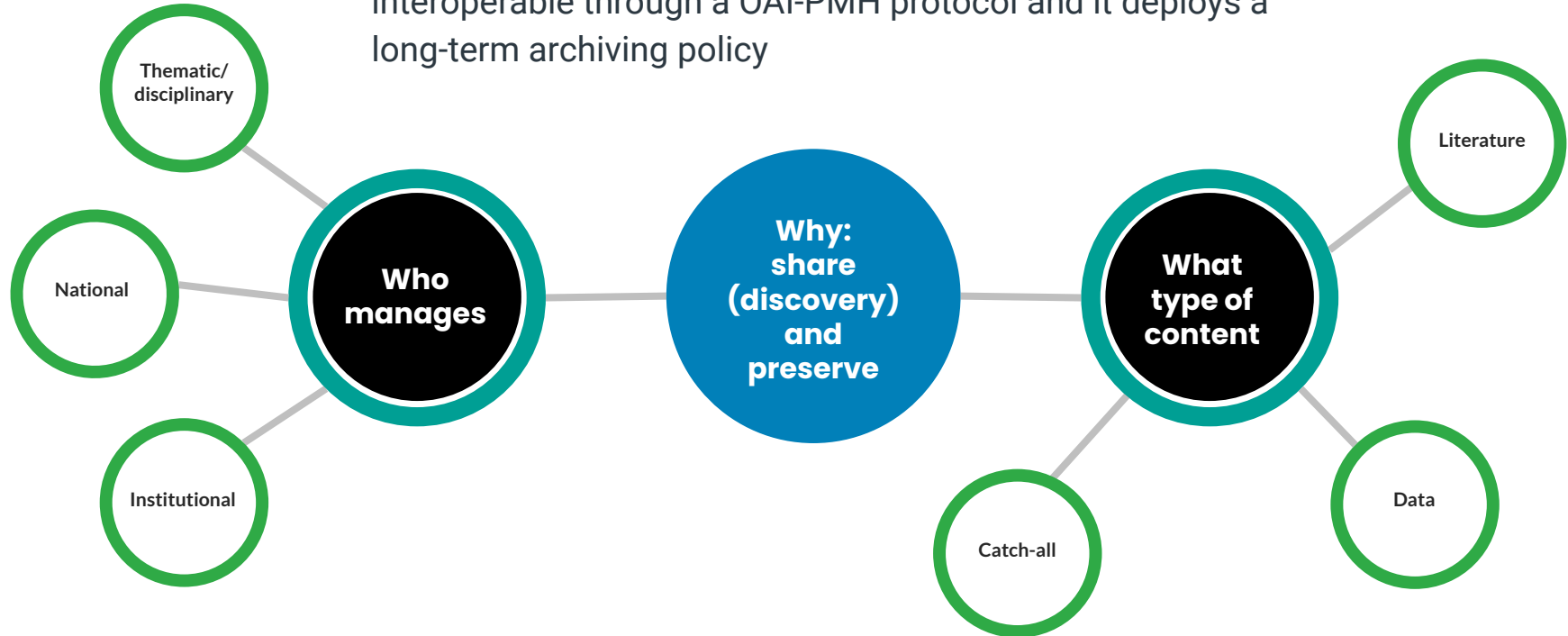
To facilitate **open access** such repositories must be **interoperable** according to the **Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)**.



The screenshot shows a Zenodo record page. The browser address bar displays the URL: zenodo.org/record/5578792#.YYkofZ5Kgbk. The Zenodo logo is visible in the top left of the page header, along with a search bar and navigation links for 'Upload' and 'Communities'. The user profile 'gina.pavone@isti.cnr.it' is shown in the top right. The record details include the date 'October 19, 2021', the title 'More Open, better Science. Towards a new research evaluation system', and the author 'Favone, Gina'. There are buttons for 'Presentation', 'Open Access', 'Edit', and 'New version'. The record has 50 views and 89 downloads. A 'Communities' section shows 'Open Science in Italy' with a 'Remove' button. The 'Indexed in' section shows 'OpenAIRE'.

Open Access repositories

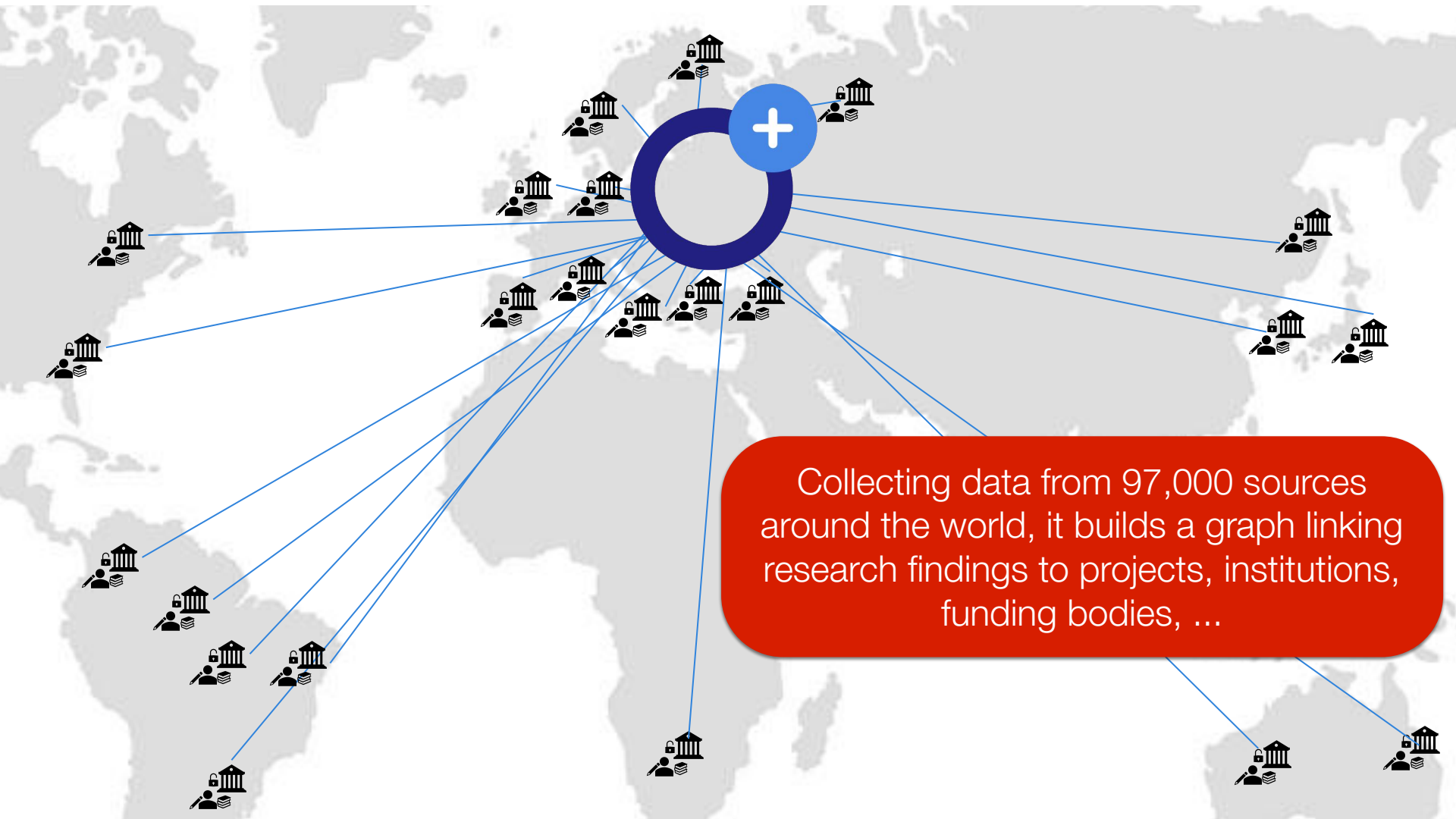
A repository stores Open Access digital objects and makes them available and downloadable. It's accessible and interoperable through a OAI-PMH protocol and it deploys a long-term archiving policy



OpenAIRE



OpenAIRE
L'infrastruttura
Europea per
l'Open Access



Collecting data from 97,000 sources around the world, it builds a graph linking research findings to projects, institutions, funding bodies, ...

OpenAIRE in numbers

Our growing Community

24

FUNDERS

97K

CONTENT PROVIDERS

3M

PROJECTS

138,849,595

 PUBLICATIONS

15,628,503

 RESEARCH DATA

278,357

 SOFTWARE

6,721,123

 OTHER RESEARCH PRODUCTS

Many components of the graph

Organization



OpenAIRE | EXPLORE

UNIGE

Università degli Studi di Genova

Organization Italy

Publications (0) +

Projects (154) +

Content Providers (1) +

Repository



OpenAIRE | EXPLORE

Archivio istituzionale della ricerca - Università di Genova

Institutional Repository OpenAIRE 3.0 (OA, funding)

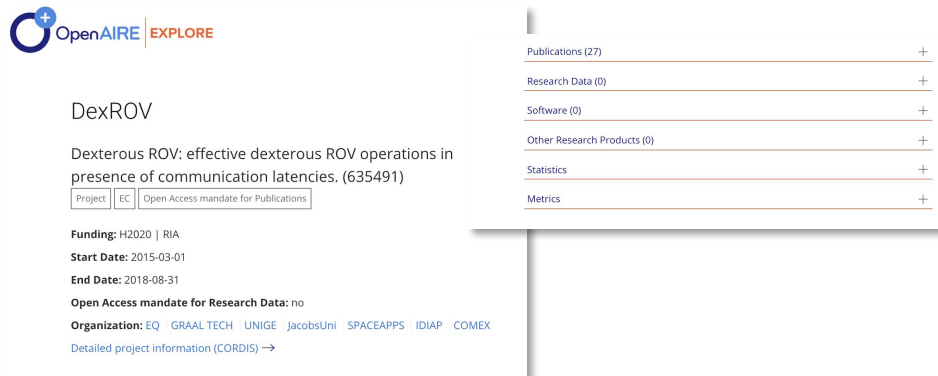
OAI-PMH: <https://iris.unige.it/oai/request?> →

Detailed content provider information (OpenDOAR) →

Countries: Italy

Publications (262)	+
Research Data (0)	+
Software (0)	+
Other Research Products (1)	+
Organizations (1)	+
Statistics	+
Metrics	+

Project



OpenAIRE | EXPLORE

DexROV

Dexterous ROV: effective dexterous ROV operations in presence of communication latencies. (635491)

Project EC Open Access mandate for Publications

Funding: H2020 | RIA

Start Date: 2015-03-01

End Date: 2018-08-31

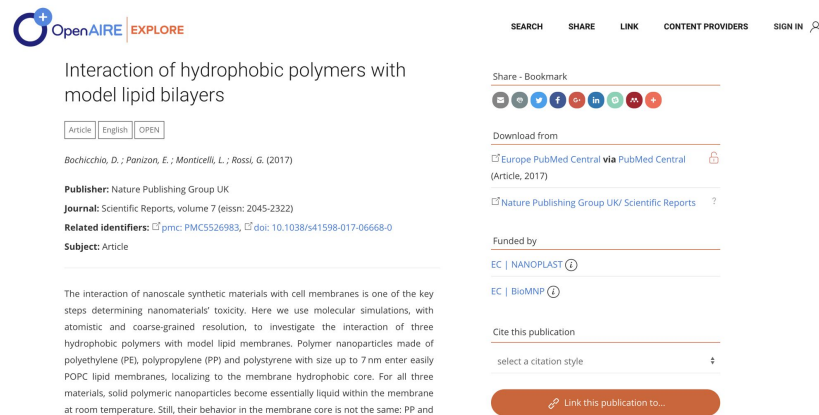
Open Access mandate for Research Data: no

Organization: EQ GRAAL TECH UNIGE JacobsUni SPACEAPPS IDIAP COMEX

Detailed project information (CORDIS) →

Publications (27)	+
Research Data (0)	+
Software (0)	+
Other Research Products (0)	+
Statistics	+
Metrics	+

Research output



OpenAIRE | EXPLORE

Interaction of hydrophobic polymers with model lipid bilayers

Article English OPEN

Boichichio, D.; Panizon, E.; Monticelli, L.; Rossi, G. (2017)

Publisher: Nature Publishing Group UK

Journal: Scientific Reports, volume 7 (eissn: 2045-2322)

Related identifiers: [PMCID: PMC5526983](https://doi.org/10.1038/s41598-017-06668-0), [doi: 10.1038/s41598-017-06668-0](https://doi.org/10.1038/s41598-017-06668-0)

Subject: Article

The interaction of nanoscale synthetic materials with cell membranes is one of the key steps determining nanomaterials' toxicity. Here we use molecular simulations, with atomistic and coarse-grained resolution, to investigate the interaction of three hydrophobic polymers with model lipid membranes. Polymer nanoparticles made of polyethylene (PE), polypropylene (PP) and polystyrene with size up to 7 nm enter easily POPC lipid membranes, localizing to the membrane hydrophobic core. For all three materials, solid polymeric nanoparticles become essentially liquid within the membrane at room temperature. Still, their behavior in the membrane core is not the same: PP and

SEARCH SHARE LINK CONTENT PROVIDERS SIGN IN

Share - Bookmark

Download from

Europe PubMed Central via PubMed Central (Article, 2017)

Nature Publishing Group UK/ Scientific Reports ?

Funded by

EC | NANOPLAST

EC | BioMNP

Cite this publication

select a citation style

Link this publication to...

Registries and other tools for findability

- OpenDOAR <https://v2.sherpa.ac.uk/opendoar/>
A quality-assured, global Directory of Open Access Repositories
- FAIRsharing: <https://fairsharing.org/>
A curated, informative and educational resource on data and metadata standards, inter-related to databases and data policies.
- Re3data <https://www.re3data.org/>
Registry of research data repositories
- Roar <http://roar.eprints.org/>
Registry of Open Access repositories
- Google search/scholar + Unpaywall <https://unpaywall.org/>
Unpaywall is database of scholarly articles and a browser extension skip the paywall on millions of peer-reviewed journal articles: it free, and legal!

F4

Use appropriate resources to improve discoverability

HAVE

A

BREAK

