

# FAIR and Open Data

Silvio Peroni

[silvio.peroni@unibo.it](mailto:silvio.peroni@unibo.it) – <https://orcid.org/0000-0003-0530-4305> – [@essepuntato](https://twitter.com/essepuntato)

Open Science (A.Y. 2023/2024)

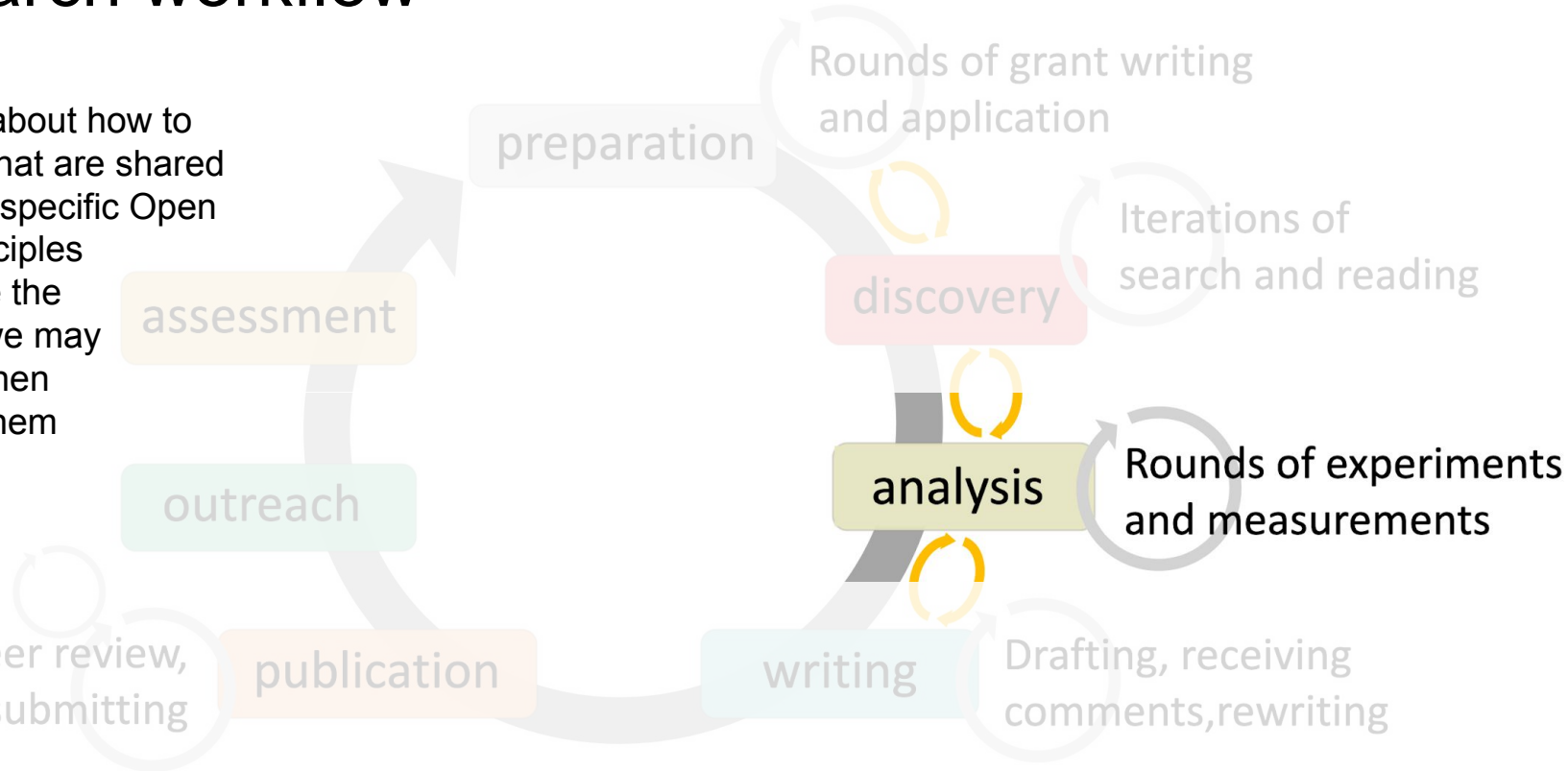
Second Cycle Degree in Digital Humanities and Digital Knowledge

Alma Mater Studiorum - Università di Bologna



# Research workflow

We discuss about how to create data that are shared according to specific Open Science principles and what are the Issues that we may encounter when processing them



# What are the data?

## Metadata about the publication

PeerJ

Submitted 9 August 2017  
Accepted: 25 January 2018  
Published: 13 February 2018

Corresponding authors:  
Heather Piowar, heather@piowar.org  
Jason Priem, jason@piowar.org  
Academic editor:  
Robert McDonald  
Additional Information and  
Declarations can be found on  
page 19  
DOI 10.7717/peerj.4375

© Copyright  
2018 Piowar et al.  
Distributed under  
Creative Commons CC-BY 4.0  
OPEN ACCESS

The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles

Heather Piowar<sup>1,2</sup>, Jason Priem<sup>1,3</sup>, Vincent Larivière<sup>4,5</sup>, Juan Pablo Alperin<sup>6,7</sup>, Lisa Matthias<sup>8</sup>, Bree Norlander<sup>9,10</sup>, Ashley Farley<sup>11</sup>, Jevin West<sup>12</sup> and Stefanie Haustein<sup>13</sup>

<sup>1</sup>Department, Sanford, NC, USA  
<sup>2</sup>École de Bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal, QC, Canada  
<sup>3</sup>Observatoire des Sciences et des Technologies (OST), Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Université de Québec à Montréal, Montréal, QC, Canada  
<sup>4</sup>Canadian Institutes for Studies in Publishing, Simon Fraser University, Vancouver, BC, Canada  
<sup>5</sup>Public Knowledge Project, Canada  
<sup>6</sup>Scholarly Communications Lab, Simon Fraser University, Vancouver, Canada  
<sup>7</sup>Information School, University of Washington, Seattle, USA  
<sup>8</sup>FloridaDA, USA  
<sup>9</sup>School of Information Studies, University of Ottawa, Ottawa, ON, Canada  
<sup>10</sup>These authors contributed equally to this work.

**ABSTRACT**  
Despite growing interest in Open Access (OA) to scholarly literature, there is an unmet need for large-scale, up-to-date, and reproducible studies assessing the prevalence and characteristics of OA. We address this need using oAid, an open online service that determines OA status for 67 million articles. We use three samples, each of 100,000 articles, to investigate OA in three populations: (1) all journal articles assigned a Creative Commons DOI, (2) recent journal articles indexed in Web of Science, and (3) articles viewed by users of Unpaywall, an open-source browser extension that lets users find OA articles using oAid. We estimate that at least 28% of the scholarly literature is OA (18% in total) and that this proportion is growing, driven particularly by growth in Gold and Hybrid. The most recent year analyzed (2015) also has the highest percentage of OA (48%). Because of this growth, and the fact that readers disproportionately access newer articles, we find that Unpaywall users encounter OA quite frequently: 47% of articles they view are OA. Notably, the most common mechanism for OA is not Gold, Green, or Hybrid OA, but rather an under-discussed category we dub *Browser articles* made free-to-read on the publisher website, without an explicit Open license. We also examine the citation impact of OA articles, corroborating the so-called open-access citation advantage: accounting for age and discipline, OA articles receive 18% more citations than average, an effect driven primarily by Green and Hybrid OA. We encourage further research using the free oAid service, as a way to inform OA policy and practice.

**Subjects** Legal Issues, Science Policy, Data Science  
**Keywords** Open access, Open science, Scientometrics, Publishing, Libraries, Scholarly communication, Bibliometrics, Science policy

To the interest of full disclosure, it should be noted that two of the authors of the paper are no members of the open peer-review organization that provided oAid.

1. What percentage of the scholarly literature is OA, and how does this percentage vary according to publisher, discipline, and publication year?

2. Are OA papers more highly-cited than their toll-access counterparts?

The next section provides a brief review of the background literature for this paper, followed by a description of the datasets and methods used, as well as details on the

PeerJ

Peerwar et al. (2018), PeerJ, DOI 10.7717/peerj.4375

32/39

## Metadata about the data

Random 100,000										
oa_color	Papers with		Percentage	Average relative citations						
closed	63,933		63.9%	0.90						
all open	36,067		36.1%	1.18						
bronze	12,939		12.9%	1.22						
hybrid	4,314		4.3%	1.31						
gold	7,351		7.4%	0.83						
green only	11,463		11.5%	1.33						
all papers	100,000		100.0%	1.00						
Access per year										
oa_color	2009	2010	2011	2012	2013	2014	2015	2009-2015		
closed	7,949	8,322	8,825	9,375	9,959	9,080	10,423	63,933		
all open	3,982	4,381	4,753	5,241	5,825	5,442	6,343	36,067		
bronze	1,757	1,792	1,886	2,092	1,916	1,686	1,810	12,939		
hybrid	417	483	541	539	660	759	915	4,314		
gold	381	527	705	1,031	1,455	1,370	1,882	7,351		
green only	1,427	1,579	1,621	1,679	1,794	1,627	1,736	11,463		
all papers	11,931	12,703	13,578	14,716	15,784	14,522	16,766	100,000		
oa_color	2009	2010	2011	2012	2013	2014	2015	2009-2015		
closed	66.6%	65.5%	65.0%	63.7%	63.1%	62.5%	62.2%	63.9%		
all open	33.4%	34.5%	35.0%	36.3%	36.9%	37.5%	37.8%	36.1%		
bronze	14.7%	14.1%	13.9%	14.2%	12.1%	11.6%	10.8%	12.9%		
hybrid	3.5%	3.8%	4.0%	3.7%	4.2%	5.2%	5.5%	4.3%		
gold	3.2%	4.1%	5.2%	7.0%	9.2%	9.4%	11.2%	7.4%		
green only	12.0%	12.4%	11.9%	11.4%	11.4%	11.2%	10.4%	11.5%		
all papers	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%		
Impact of access per year										
oa_color	2009	2010	2011	2012	2013	2014	2015	2009-2015		
closed	0.86	0.87	0.91	0.88	0.91	0.93	0.93	0.90		
all open	1.19	1.22	1.21	1.17	1.19	1.15	1.17	1.18		
bronze	1.12	1.20	1.20	1.16	1.26	1.12	1.46	1.22		
hybrid	1.18	1.23	1.32	1.32	1.30	1.33	1.39	1.31		
gold	1.04	0.90	0.95	0.86	0.89	0.82	0.66	0.83		
green only	1.32	1.33	1.30	1.34	1.33	1.37	1.32	1.33		
all papers	0.97	0.89	1.01	0.99	1.01	1.01	1.02	1.00		

Used in

A publication (<https://doi.org/10.7717/peerj.4375>)  
accepted in a journal

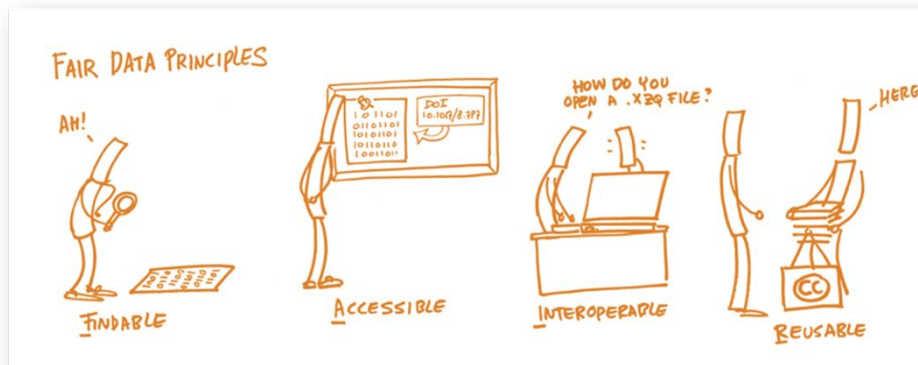
The related data on which the entire publication is  
based on

# FAIR data principles

**Findability, Accessibility, Interoperability, and Reusability:** these [four principles](#), proposed by the [FORCE11 community](#), serve to guide data producers and publishers for helping them to maximize the added-value gained by contemporary, formal scholarly digital publishing

They represent goals and desiderata of good data management and stewardship

Even if they have been devised originally for data, they have been proposed considering **all scholarly digital research objects** in mind, since all components of the research process must be available to ensure transparency, reproducibility, and reusability



# Findable

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (see also R1)

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

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

# Accessible

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

# 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

# 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



# Different types of data

The agnosticism of generic principles like FAIR requires the various communities' specificities (i.e. proper to Informatics, Humanities, Medicine, Statistics, etc.) to be taken into account

General rule: within the FAIR framework, **everything is data**

Data should be “intended to be as universal as possible, including datasets, publications, software, etc.”

# Some examples: what is data in Informatics?

A number of research communities and groups have been considering how to apply aspects of FAIR to **research software** since 2017

The adoption of FAIR principles enables the transparency, reproducibility, and reusability of research – and part of this story applies also to software that needs to be **well-described (i.e. metadata), inspectable, documented and appropriately structured** so that it can be executed, replicated, built-upon, combined, reinterpreted, reimplemented, and/or used in different settings

<b>F: Software, and its associated metadata, is easy for both humans and machines to find.</b>
F1. Software is assigned a globally unique and persistent identifier. <ul style="list-style-type: none"><li>• F1.1. Components of the software representing levels of granularity are assigned distinct identifiers.</li><li>• F1.2. Different versions of the software are assigned distinct identifiers.</li></ul> F2. Software is described with rich metadata. F3. Metadata clearly and explicitly include the identifier of the software they describe. F4. Metadata are FAIR, searchable and indexable.
<b>A: Software, and its metadata, is retrievable via standardized protocols.</b>
A1. Software is retrievable by its identifier using a standardized communications protocol. <ul style="list-style-type: none"><li>• A1.1. The protocol is open, free, and universally implementable.</li><li>• A1.2. The protocol allows for an authentication and authorization procedure, where necessary.</li></ul> A2. Metadata are accessible, even when the software is no longer available.
<b>I: Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.</b>
I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards. I2. Software includes qualified references to other objects.
<b>R: Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software).</b>
R1. Software is described with a plurality of accurate and relevant attributes. <ul style="list-style-type: none"><li>• R1.1. Software is given a clear and accessible license.</li><li>• R1.2. Software is associated with detailed provenance.</li></ul> R2. Software includes qualified references to other software. R3. Software meets domain-relevant community standards.

# Some examples: what is data in the Humanities?

Humanities (at least, those related to studies in philology, literary criticism, language, linguistics, history of art and archival and library studies) uses and/or produces several kinds of data

Probably, one of the most surprisingly data types is **events**, meaning “any one-off gathering of people organised as a result of a research project, to share ideas, offer training, or present something to the public”, including:

- Conferences
- Exhibitions
- Webinars
- Guided tours
- Teacher training

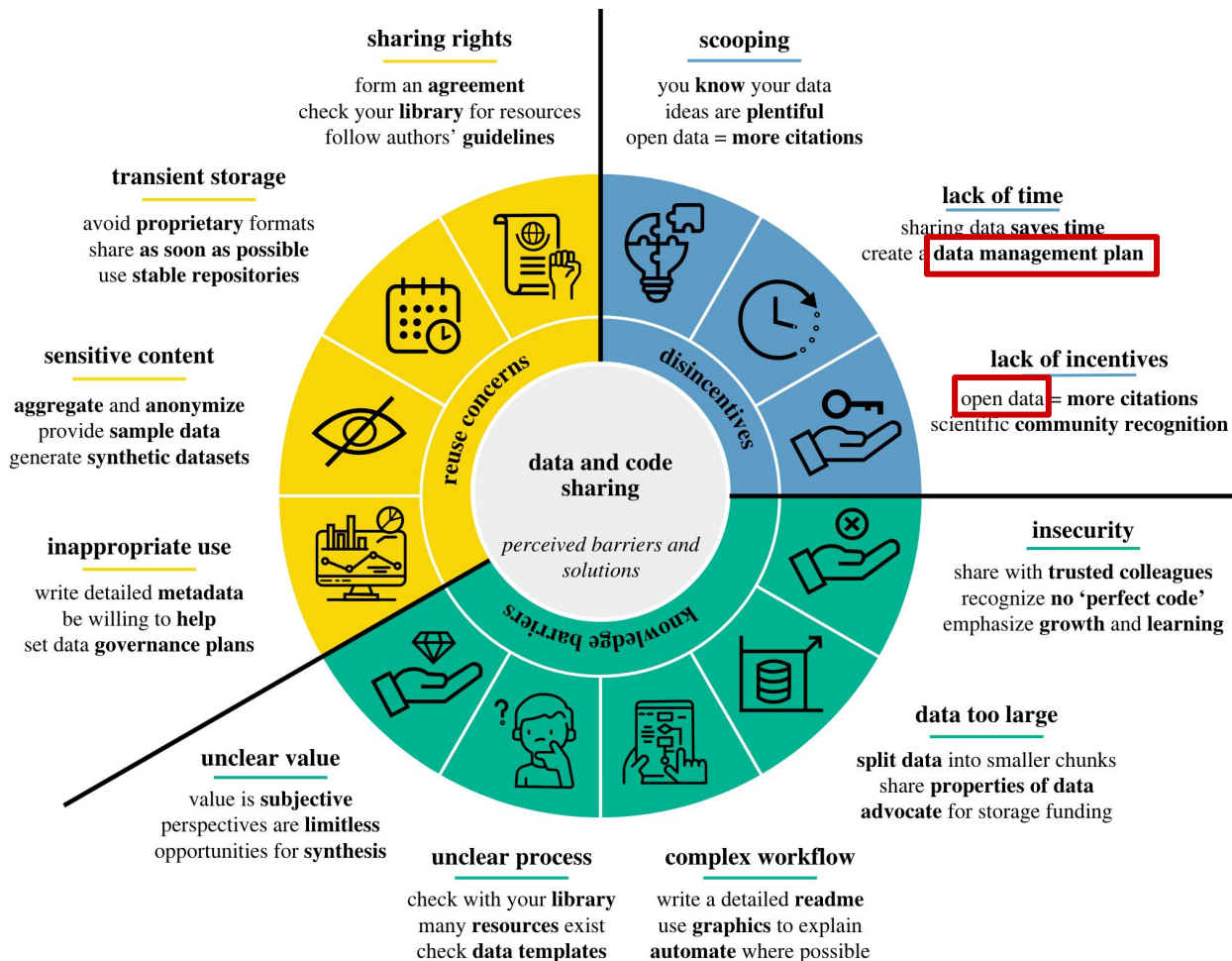
Data types	Produced by (tot. 19)	Used by (tot. 19)
publications	18	15
other primary sources (e.g., manuscripts, artworks)	0	18
digital representation of cultural objects or events (e.g., facsimiles, photo, video)	4	8
catalogues, databases and other search tools	2	9
events (e.g., conferences, exhibitions)	6	0
static websites	4	0
software	2	2
documentation	3	0
digital infrastructures (e.g., mobile apps, web platforms)	3	0
personal data	2	0
corpora	2	0
standards	0	2
born-digital artifacts (e.g., tags, associations, texts)	1	1

# Sharing: some guidelines

Many researchers do not share their data simply because they do not know **how**

Data should be published alongside **detailed metadata**

Data sharing will increase opportunities for **collaboration**



# What about Open?

Warning: FAIR data **does not imply** Open Data

Open as defined in the Open Definition:

“Open means anyone can **freely access, use, modify, and share for any purpose** (subject, at most, to requirements that preserve provenance and openness)”

License	Attribution (i.e. preserving provenance)	Forcing openness
<a href="#">Creative Commons Attribution (CC-BY)</a>	yes	no
<a href="#">Creative Commons Attribution - Share-Alike (CC-BY-SA)</a>	yes	yes
<a href="#">Creative Commons CCZero (CC0)</a>	no	no
<a href="#">Open Data Commons Public Domain Dedication and Licence (PDDL)</a>	no	no
<a href="#">Open Data Commons Attribution License (ODC-BY)</a>	yes	no
<a href="#">Open Data Commons Open Database License (ODbL)</a>	yes	yes

# Issues to the path towards the Open: personal data

GDPR is a regulation in **EU law** which concerns data protection and privacy, with particular regard to the gathering and processing of personal data of individuals

As stated in the Article 4 of the GDPR, **personal data** is any information that enable the identification of a natural person (i.e. the data subject), in particular by reference to an identifier such as a name, an identification number, etc.

Personal data can be gathered under specific conditions and **cannot be published** as Open Data – derogation to the this rule may exist, e.g. personal data in bibliographic information (authors' names, ORCIDs, etc.)

Note: the principles of data protection do not apply to **anonymous** information

# Anonymisation and pseudonymisation

It is **possible to publish** personal data if they are somehow changed in a way that do not allow the identification of a natural person anymore

Two ways:

1. **Anonymisation** – rendering personal data anonymous in such a manner that the data subject is not or no longer identifiable (there exists tools that try to semi-automate this activity, such as [OpenAIRE Amnesia](#))
2. **Pseudonymisation** – rendering personal data in such a manner that they can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person

Applying one of these approaches **enable one to publish Open Data** also when the original gathering of such data included personal data

# How to publish data

Generally speaking, data should be published

as open as possible, as closed as necessary

In choosing the license, in particular if considering or not the attribution clause, one has to think about the final goal to reach – that, within the Open Science ecosystem, should be fostering maximum reuse

Recently, the European Union has provided a [clear guideline](#) to follow for what it concerns the publication of data:

“raw data, metadata or other documents of comparable nature may alternatively be distributed under the provisions of the Creative Commons Universal Public Domain Dedication deed (CC0 1.0)”



# The (senseless) scholars' fear

Misconception: *I want to share my data, but it is important to me that I'm cited when people use it. I prefer CC BY to CC0 because this kind of attribution is what I care about most*

If the issue is to be cited and acknowledged for a work, problem solved: citation practices **are built around ethical norms**, not around legal requirements

Using and not citing and acknowledging an existing work relates to **plagiarism**, and this act has a legal consequence independently from the license used to protect the data

Thus, the general rule for sharing data compliantly with Open Science practices is to use CC0 license instead of CC BY to foster maximum reuse

# Trustworthy data

**Provenance** is a record that describes the **people, institutions, entities, and activities** involved in producing, influencing, or delivering data, and it is crucial in deciding whether information is to be **trusted**, how it should be **integrated with other** diverse information sources, and how to **give credit to its originators** when reusing it

Provenance can help to **make trust judgements** about a piece of data, since it , which can be used to form assessments about its **quality, reliability** or **trustworthiness**

Minimal provenance information:

- When a piece of data has been created and/or invalidated
- Who created a piece of data
- What is the source that has been used to create a piece of data

# Trustworthy long-term preservation of data

Guaranteeing that the data one publish are reusable by humans and machines, i.e. FAIR, is not enough, since a good part of the story concerns also to think ways to **trustworthily preserve** such data in the long term

You should take into consideration such digital repositories which are compliant with the **TRUST principles**:

- **Transparency** – repository services and data holdings are publicly verifiable
- **Responsibility** – authenticity+integrity of data and reliability+persistence of services
- **User Focus** – meet norms and expectations of target user communities
- **Sustainability** – preserve data and services for the long-term
- **Technology** – support secure, persistent, and reliable services

# Data Management Plan

A data management plan (DMP) is the tool one has to use to ensure that all the aspects introduced in the previous slides will be addressed appropriately – i.e. before one starts to gather the data

A DMP **is a document** that describes how you will treat your data during a project and what happens with the data after the project ends

What it is covered in a DMP:

- data discovery, collection and organization
- quality assurance/quality control
- documentation
- data preservation and sharing

# End

## FAIR and Open Data

Silvio Peroni

[silvio.peroni@unibo.it](mailto:silvio.peroni@unibo.it) – <https://orcid.org/0000-0003-0530-4305> – [@essepuntato](#)

Open Science (A.Y. 2023/2024)

Second Cycle Degree in Digital Humanities and Digital Knowledge

Alma Mater Studiorum - Università di Bologna

