

# All about FAIR principles

University of Pisa - Phd course

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

module 2 - 30 May 2022

A - Accessible

# 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

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

- E.g. http(s) or ftp.
- Protocols with limited implementations, poor documentation, and components involving manual human intervention should be avoided.
- There are exceptions: for example for highly sensitive data. In such cases, it is perfectly FAIR to provide an email or telephone number of a contact person who can discuss access to the data. This contact protocol must be clear and explicit in the metadata.

# In other words:

If one knows a data set's identifier and the location where it is archived, one can access at least the metadata. Furthermore, the user knows how to proceed to get access to the data (as long as the conditions of access are clearly detailed)





Are there any restrictions on access e.g. because of sensitive data? Conditions of access (e.g., who to contact and how) should be clearly specified

# A 1.1 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. (E.g. HTTP, FTP, SMTP, ...)

# A 1.2 The protocol allows for an authentication and authorisation where necessary

The exact conditions under which the data are accessible should be provided. Ideally, accessibility is specified in such a way that a machine can automatically understand the requirements, and then either automatically execute the requirements or alert the user to the requirements.

# In other words:

## A 1.1

Anyone with a computer and an internet connection can access at least the metadata.

ie. the repository should not rely on a proprietary or commercial communication protocol.

## A1.2

It often makes sense to request users to create a user account on a repository. This allows to authenticate the owner (or contributor) of each data set, and to potentially set user specific rights.

So repositories should provide a way for authentication and authorization of users, including machine-users.



# Some examples

## Good

HTTP, FTP, SMTP, ...

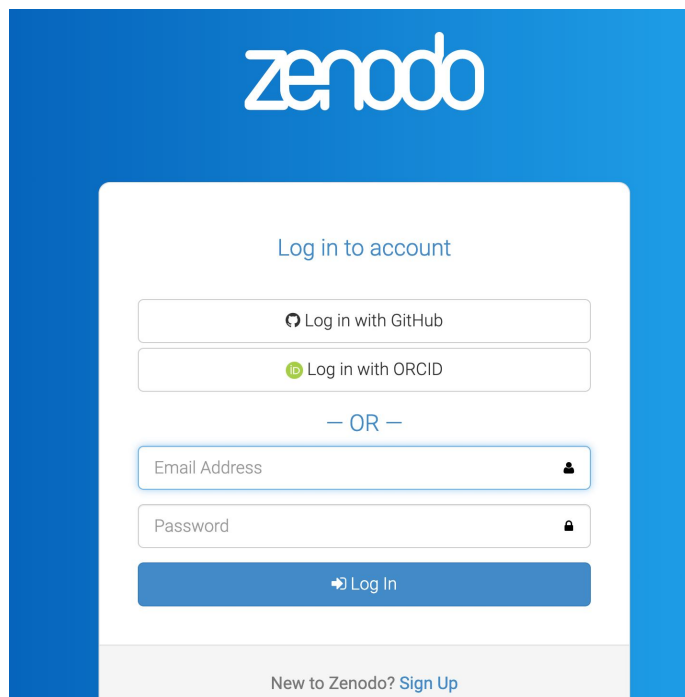
Telephone (arguably not  
universally-implementable, but  
close enough)

## Bad

Skype, as it is not  
universally-implementable  
because it is proprietary

Microsoft Exchange Server  
protocol is also proprietary

# E.g.: trying to facilitate authentication procedure



The Zenodo login interface features a blue header with the Zenodo logo. Below the header, the text "Log in to account" is centered. There are two buttons for social login: "Log in with GitHub" and "Log in with ORCID". Below these is a separator "– OR –". The main login form consists of three input fields: "Email Address" with a user icon, "Password" with a lock icon, and a blue "Log In" button with a right-pointing arrow. At the bottom, there is a link for "New to Zenodo? Sign Up".



Italiano v

## Accedi

Username o email

Password

Ricordami

[Password dimenticata?](#)

**Accedi**

	Academic / other
	LinkedIn
	Google
	Twitter
	GitHub

Nuovo utente? [Registrati](#)

# A.2 Metadata should be accessible even when the data is no longer available

Metadata should persist even when the data are no longer sustained

A2 is related to the registration and indexing issues described in F4.

**Open Data**  
and  
**FAIR Data**  
are different  
concepts



Photo by [Serhat Beyazkaya](#) on [Unsplash](#)

### **(FAIR) Open Data**

Data can be freely used, shared, enriched by anyone, anywhere for any purpose.

### **FAIR Data**





Data follow a series of good practices to allow data access, still respecting any ethical, legal and contractual restriction.

# Why do we need a distinction?

Photo by [Possessed Photography](#) on [Unsplash](#)



## Research data could:

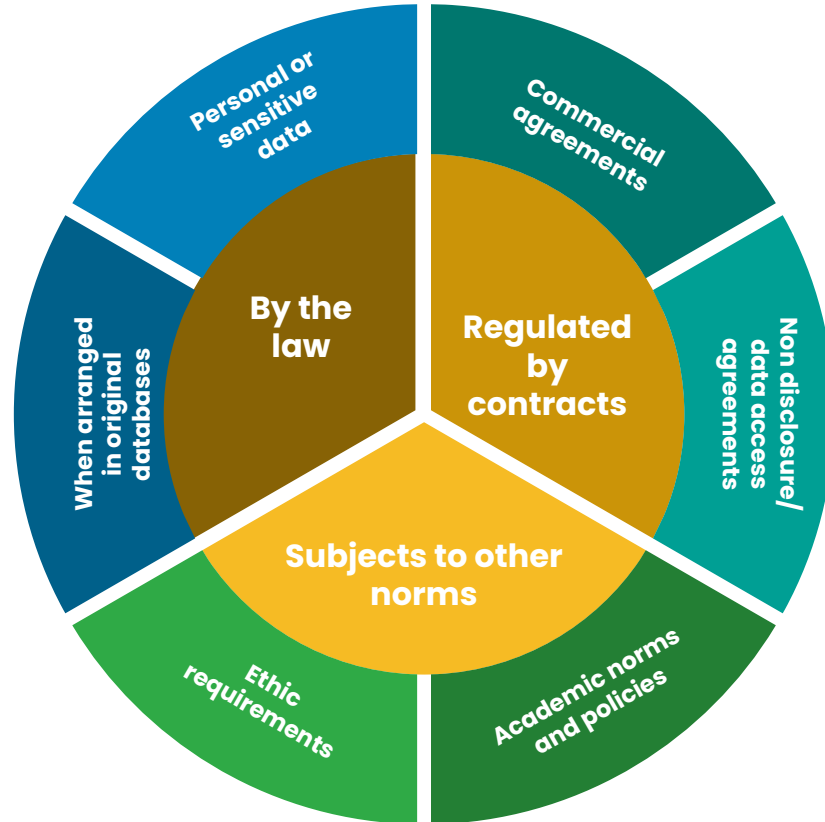
-  Contain personal information (privacy e GDPR)
-  Fall under copyright (in the case of a database with creative structure)
-  Fall under the Sui Generis right (database obtained thanks to a substantial investment)
-  Be protected by patent or industrial secret

**Data sharing needs to respect the specific law.**

**Data needs to be protected against non authorised access.**

# Data can be protected

Multiple types of protection might exist in research data, or there may be elements that have no legal protection



# Data to be handled with great care:

- Personal data: any information about an identified or identifiable natural person (directly or indirectly)
- Personal sensitive data (i.e. revealing racial or ethnic origin, political views, religious or philosophical beliefs, membership of a trade union, genetic data, biometric data, data about health or someone's sexual behavior or sexual orientation)
- Data protected by IPR (Intellectual Property Rights) agreements
- Confidential data (i.e. commercial agreements)

This means that access to the data must be managed and restricted.

They still can be FAIR



# An important difference



**Deposit:** upload a digital object (data, articles, ...) on a platform that allows to correctly describe the object through metadata and that implements long-term preservation.



**Give access:** once the object has been deposited, the authors can choose the type of access that can be granted (open, restricted, closed, embargoed, ...) and assigns a licence to reuse the contents (Creative Commons)



# Data are not yours



Data is **not** intellectual work, it is fact and information



Copyright protection covers expressions and not ideas, procedures, operating methods or mathematical concepts as such.



**Protection is on databases and not on data.** Data are protected only and especially when they are collected and organized in a database.



**The sui generis property right (only in Europe)** covers not only the reproduction and dissemination of the database, but also the extraction and reuse of substantial parts of the database.

# A valid resource (in Italian)

## banche dati: diversi livelli di tutela

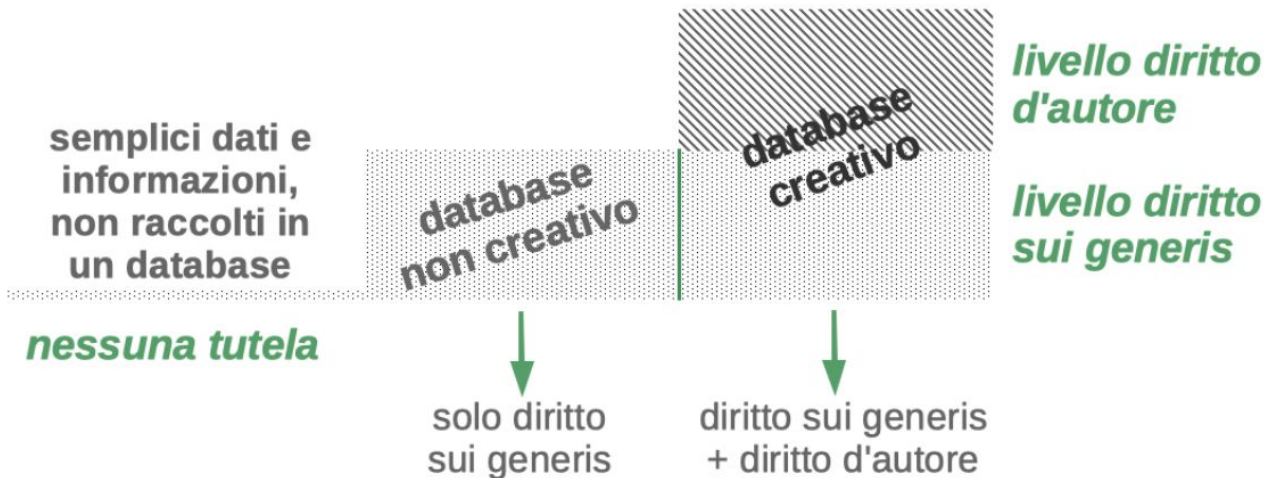


Figura 1: I due diversi livelli di tutela per le banche dati basati sul requisito del carattere creativo

Aliprandi, Simone. (2022). Aspetti legali degli open data: la guida definitiva (1.0 (maggio 2022)). Zenodo. <https://doi.org/10.5281/zenodo.6575822>

# Some consideration on data protection

- **Copyright** is a property right in certain types of original literary, artistic and scientific works.
- Copyright does not protect **ideas**.
- **Confidentiality** protects confidential information. This might be imposed by a contract or if the information is marked confidential. Use of confidential information might give rise to a claim for compensation if confidentiality is breached.
- Data Subject Rights arise in information that identifies individuals and are recognised by data protection laws in the EU.
- **Patents** are registered rights in novel inventions of products or processes.
- Some research data may not benefit from any legal protection, although **moral** and **ethical** considerations may apply.

# Data and law protection

- **Raw data** are not protected by copyright
- **Database** is defined as a collection of independent works, data or other materials arranged in a systematic or methodical way
- **Copyright** protects the structure, selection or arrangement of the database contents, not the data
- **Sui generis database right**: protects the substantial effort in obtaining data (not creating). Note: the right owner is often the institution.

# How can you adhere to FAIR principles if your data cannot be opened?







Photo by Jon Ivson on Unsplash

## Create and share a description of your data

- This way other researchers may **ask for permission to access** your data for reuse purposes, by giving a specific aim and following the rules defined by the law.
- **Restrict access** to the record payload (attachment, files,...)

# Access Right: Open Access

## Access right \*

-  Open Access
-  Embargoed Access
-  Restricted Access
-  Closed Access

Required. Open access uploads have considerably higher visibility on Zenodo.

## License \*

Creative Commons Attribution 4.0 International





Required. Selected license applies to all of your files displayed on the top of the form. If you want to upload some of your files under different licenses, please do so in separate uploads. If you cannot find the license you're looking for, include a relevant LICENSE file in your record and choose one of the *Other* licenses available (*Other (Open)*, *Other (Attribution)*, etc.). The supported licenses in the list are harvested from [opendefinition.org](https://opendefinition.org) and [spdx.org](https://spdx.org). If you think that a license is missing from the list, please [contact us](#).

This should be the default access right

Always assign a licence for reuse

# Access Right: Embargoed Access

Access right \*

-  Open Access
-  Embargoed Access
-  Restricted Access
-  Closed Access

Required. Open access uploads have considerably higher visibility on Zenodo.

 Embargo date

Required only for Embargoed Access uploads. Format: YYYY-MM-DD. The date your upload will be made publicly available in case it is under an embargo period from your publisher.

 License \*

Required. Selected license applies to all of your files displayed on the top of the form. If you want to upload some of your files under different licenses, please do so in separate uploads. If you cannot find the license you're looking for, include a relevant LICENSE file in your record and choose one of the *Other* licenses available (*Other (Open)*, *Other (Attribution)*, etc.). The supported licenses in the list are harvested from [opendefinition.org](https://opendefinition.org) and [spdx.org](https://spdx.org). If you think that a license is missing from the list, please [contact us](#).

Use it when you have a **valid reason** to delay access

Always assign a licence for reuse

Note: metadata is always accessible to everyone




# Access Right: Restricted Access

Access right \*

- Open Access
- Embargoed Access
- Restricted Access
- Closed Access

Required. Open access uploads have considerably higher visibility on Zenodo.

Conditions \*



Use it when you have a **valid reason** to restrict the access





Always specify conditions under which you grant access (who, how, why can get access to your payload)

Note: metadata is always accessible to everyone

Specify the conditions under which you grant users access to the files in your upload. User requesting access will be asked to justify how they fulfil the conditions. Based on the justification, you decide who to grant/deny access. You are not allowed to charge users for granting access to data hosted on Zenodo.

# Access Right: Closed Access

Access right \*

-  Open Access
-  Embargoed Access
-  Restricted Access
-  Closed Access

Required. Open access uploads have considerably higher visibility on Zenodo.

Are you really sure you need closed access?  
consider restricted or embargoed access instead!

I - Interoperable

# 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.

# An example

Making assumptions

Date	Temp
28/05/2022	200
29/05/2022	195
30/05/2022	197

Determining

Date (DD/MM/YYYY)	Temp (K)
28/05/2022	200
29/05/2022	195
30/05/2022	197

Machines do not make assumptions!

# I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

Data should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings.

Each computer system at least has knowledge of the other system's data exchange formats.

For this to happen and to ensure automatic findability and interoperability of datasets, it is critical to use commonly used controlled vocabularies, ontologies, thesauri and a good data model.

# In other words:

Each computer system has at least knowledge of the other system's formats in which data is exchanged. If (meta)data are to be searchable and if compatible data sources should be combinable in a (semi)automatic way, computer systems need to be able to decide if the content of data sets are comparable. Obvious issues arise when different languages are used to describe the data or when spelling errors make the comparison of descriptions and variable names more difficult. So provide machine readable data and metadata in an accessible language, using a well-established formalism.



# Two initiatives:

## Data Documentation Initiative

The Data Documentation Initiative (DDI) is an international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences.

<https://ddialliance.org/>

## NetCDF (Network Common Data Form)

It is a set of software libraries and **self-describing**, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.

<https://en.wikipedia.org/wiki/NetCDF>

<https://www.unidata.ucar.edu/software/netcdf/>



# Formats

To make your data understandable to others (humans and machines) you need to use adequate standards and formats

- Formats may refer to:
  - File format (.txt, .docx, .jpeg, etc)
  - Metadata (Dublin core, discipline specific standards)
  - Data organisation/visualisation
- Use specific ontologies and vocabularies to make your data easy to read
- Use your discipline specific standards: you will spend less time curating and interpreting data and more time to actually make science!

# Dublin Core metadata:

## Dublin Core Metadata Element Set [\[ edit \]](#)

The original DCMES Version 1.1 consists of 15 metadata elements, defined this way in the original specification:<sup>[6][14]</sup>

1. Contributor – "An entity responsible for making contributions to the resource".
2. Coverage – "The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant".
3. Creator – "An entity primarily responsible for making the resource".
4. Date – "A point or period of time associated with an event in the lifecycle of the resource".
5. Description – "An account of the resource".
6. Format – "The file format, physical medium, or dimensions of the resource".
7. Identifier – "An unambiguous reference to the resource within a given context".
8. Language – "A language of the resource".
9. Publisher – "An entity responsible for making the resource available".
10. Relation – "A related resource".
11. Rights – "Information about rights held in and over the resource".
12. Source – "A related resource from which the described resource is derived".
13. Subject – "The topic of the resource".
14. Title – "A name given to the resource".
15. Type – "The nature or genre of the resource".

Each Dublin Core element is optional and may be repeated. The DCMI has established standard ways to refine elements and encourage the use of encoding and vocabulary schemes. There is no prescribed order in Dublin Core for presenting or using the elements. The Dublin Core became a NISO standards, Z39.85, and IETF RFC 5013 in 2007, ISO 15836 standard in 2009 and is used as a base-level data element set for the description of learning resources in the [ISO/IEC 19788-2 Metadata for learning resources \(MLR\) – Part 2: Dublin Core elements](#), prepared by the [ISO/IEC JTC 1/SC 36](#).

Full information on element definitions and term relationships can be found in the [Dublin Core Metadata Registry](#).<sup>[15]</sup>

## Encoding examples [\[ edit \]](#)

```
<meta name="DC.Format" content="video/mpeg; 10 minutes" />
<meta name="DC.Language" content="en" />
<meta name="DC.Publisher" content="publisher-name" />
<meta name="DC.Title" content="HYP" />
```

Creating bespoke parsers, in all computer languages, for all data-types and all analytical tools that require those data-types, is not a sustainable activity. As such, the focus on assisting machines in their discovery and exploration of data through application of more generalized interoperability technologies and standards at the data/repository level, becomes a first-priority for good data stewardship.

wilkinson et al.

# I2: (Meta)data use vocabularies that follow the FAIR principles



# What is a controlled vocabulary

Controlled vocabularies are standardized and organized arrangements of words and phrases and provide a consistent way to describe data. Metadata creators assign terms from vocabularies to improve information retrieval.

<https://guides.lib.utexas.edu/metadata-basics/controlled-vocabs>

Controlled vocabulary schemes mandate the use of predefined, authorised terms that have been preselected by the designers of the schemes, in contrast to **natural language** vocabularies, which have no such restriction.

[https://en.wikipedia.org/wiki/Controlled\\_vocabulary](https://en.wikipedia.org/wiki/Controlled_vocabulary)

# F2 in other words:

The controlled vocabulary used to describe data sets needs to be documented. This documentation needs to be easily findable and accessible by anyone who uses the data set.



# DDI Controlled Vocabulary for Mode Of Collection

The procedure, technique, or mode of inquiry used to attain the data.

[https://ddialliance.org/Specification/DDI-CV/ModeOfCollection\\_3.0.html](https://ddialliance.org/Specification/DDI-CV/ModeOfCollection_3.0.html)

Value of the Code	Descriptive Term of the Code	Definition of the Code
<b>Interview</b>	Interview	A pre-planned communication between two (or more) people - the interviewer(s) and the interviewee(s) - in which information is obtained by the interviewer(s) from the interviewee(s). If group interaction is part of the method, use "Focus group".
<b>Interview.FaceToFace</b>	Face-to-face interview	Data collection method in which a live interviewer conducts a personal interview, presenting questions and entering the responses. Use this broader term if not CAPI or PAPI, or if not known whether CAPI/PAPI or not.
<b>Interview.FaceToFace.CAPIorCAMI</b>	Face-to-face interview: Computer-assisted (CAPI/CAMI)	Computer-assisted personal interviewing (CAPI), or computer-assisted mobile interviewing (CAMI). Data collection method in which the interviewer reads questions to the respondents from the screen of a computer, laptop, or a mobile device like tablet or smartphone, and enters the answers in the same device. The administration of the interview is managed by a specifically designed program/application.
<b>Interview.FaceToFace.PAPI</b>	Face-to-face interview: Paper-and-pencil (PAPI)	Paper-and-pencil interviewing (PAPI). The interviewer uses a traditional paper questionnaire to read the questions and enter the answers.
<b>Interview.Telephone</b>	Telephone interview	Interview administered on the telephone. Use this broader term if not CATI, or if not known whether CATI or not.
<b>Interview.Telephone.CATI</b>	Telephone interview: Computer-assisted (CATI)	Computer-assisted telephone interviewing (CATI). The interviewer asks questions as directed by a computer, responses are keyed directly into the computer and the administration of the interview is managed by a specifically designed program.
<b>Interview.Email</b>	E-mail interview	Interviews conducted via e-mail, usually consisting of several e-mail messages that allow the discussion to continue beyond the first set of questions and answers, or the first e-mail exchange.
<b>Interview.WebBased</b>	Web-based interview	An interview conducted via the Internet. For example, interviews conducted within online forums or using web-based audio-visual technology that enables the interviewer(s) and interviewee(s) to communicate in real time.
<b>SelfAdministeredQuestionnaire</b>	Self-administered questionnaire	Data collection method in which the respondent reads or listens to the questions, and enters the responses by him/herself; no live interviewer is present, or participates in the questionnaire administration. If possible, use a narrower term. Use this broader term if the method is not described by any of the narrower terms - for example, for PDF and diskette questionnaires.
<b>SelfAdministeredQuestionnaire.Email</b>	Self-administered questionnaire: E-mail	Self-administered survey in which questions are presented to the respondent in the text body of an e-mail or as an attachment to an e-mail, but not as a link to a web-based questionnaire. Responses are also sent back via e-mail, in the e-mail body or as an attachment.

# I2

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.



# I3: (Meta)data include qualified references to other (meta)data

The goal is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data (e.g. 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).

# In other words:

E.g. the controlled vocabulary used to describe data sets needs to be documented. This documentation needs to be easily findable and accessible by anyone who uses the data set.



Zenodo is integrated into reporting lines for research funded by the European Commission via [OpenAIRE](#). Specify grants which have funded your research, and we will let your funding agency know!

**Grants**

<input type="text" value="European Commission (EU)"/>	<input type="text" value="OpenAIRE-Advance 777541 OpenAIRE Advancing Open Scholarship"/>	✕
<input type="text" value="European Commission (EU)"/>	<input type="text" value="EOSCsecretariat.eu 831644 EOSCsecretariat.eu"/>	✕

Optional. OpenAIRE-supported projects only. For other funding acknowledgements, please use the **Additional Notes** field.  
Note: a human Zenodo curator will need to validate your upload - you may experience a delay before it is available in OpenAIRE.

[+ Add another grant](#)

Related/alternate identifiers

recommended ▾

Specify identifiers of related publications and datasets. Supported identifiers include: DOI, Handle, ARK, PURL, ISSN, ISBN, PubMed ID, PubMed Central ID, ADS Bibliographic Code, arXiv, Life Science Identifiers (LSID), EAN-13, ISTC, URNs and URLs.

**Related identifiers**

<input type="text" value="10.5281/zenodo.380176"/>	<input type="text" value="continues this upload"/>	<input type="text" value="Presentation"/>	⬆️ ✕
<small>Optional. Resource type of the related identifier.</small>			
<input type="text" value="10.5281/zenodo.382618"/>	<input type="text" value="continues this upload"/>	<input type="text" value="Presentation"/>	⬆️ ✕
<small>Optional. Resource type of the related identifier.</small>			
<input type="text" value="10.5281/zenodo.390163"/>	<input type="text" value="continues this upload"/>	<input type="text" value="Presentation"/>	⬆️ ✕
<small>Optional. Resource type of the related identifier.</small>			

[+ Add another related identifier](#)

funding agency know:

Grants

European Commission (EU)

OpenAIRE-Advance 777541 OpenAIRE Advancing

European Commission (EU)

EOSCsecretariat.eu 831644 EOSCsecretariat.eu

Optional. OpenAIRE-supported pro  
Note: a human Zenodo curator wil

+ Add another grant

Related/alternate identifiers

Specify identifiers of related publications and datasets. Supported identifiers i  
arXiv, Life Science Identifiers (LSID), EAN-13, IISTC, URNs and URLs.

Related identifiers

10.5281/zenodo.3801760

10.5281/zenodo.382618:

10.5281/zenodo.390163:

+ Add another related identifie

Contributors

References

- cites this upload
- is cited by this upload
- is supplemented by this upload
- is a supplement to this upload
- is referenced by this upload
- references this upload
- published this upload
- is previous version of this upload
- is new version of this upload
- continues this upload
- is continued by this upload
- describes this upload
- is described by this upload
- has this upload as part
- is part of this upload
- reviews this upload
- is reviewed by this upload
- documents this upload
- is documented by this upload
- is compiled/created by this upload
- compiled/created this upload
- is the source this upload is derived from
- has this upload as its source
- is required by this upload
- requires this upload
- replaces this upload
- is replaced by this upload

relationship

object

subject

PIDs

Publication date:

April 30, 2020

DOI:

DOI 10.5281/zenodo.3778807

Keyword(s):

Open Science, Open Access, OpenAIRE, funders mandates

Grants:

European Commission:

- OpenAIRE-Advance - OpenAIRE Advancing Open Scholarship (777541)
- EOSCsecretariat.eu - EOSCsecretariat.eu (831644)

Related identifiers:

Continued by

10.5281/zenodo.3801760 (Presentation)

10.5281/zenodo.3826183 (Presentation)

10.5281/zenodo.3901639 (Presentation)

Communities:

Open Science in Italy

License (for files):

Creative Commons Attribution 4.0 International

R - Reusable

# Reusable

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.

# R1: (Meta)data are richly described with a plurality of accurate and relevant attributes

It will be much easier to find and reuse data if there are many labels are attached to the data.

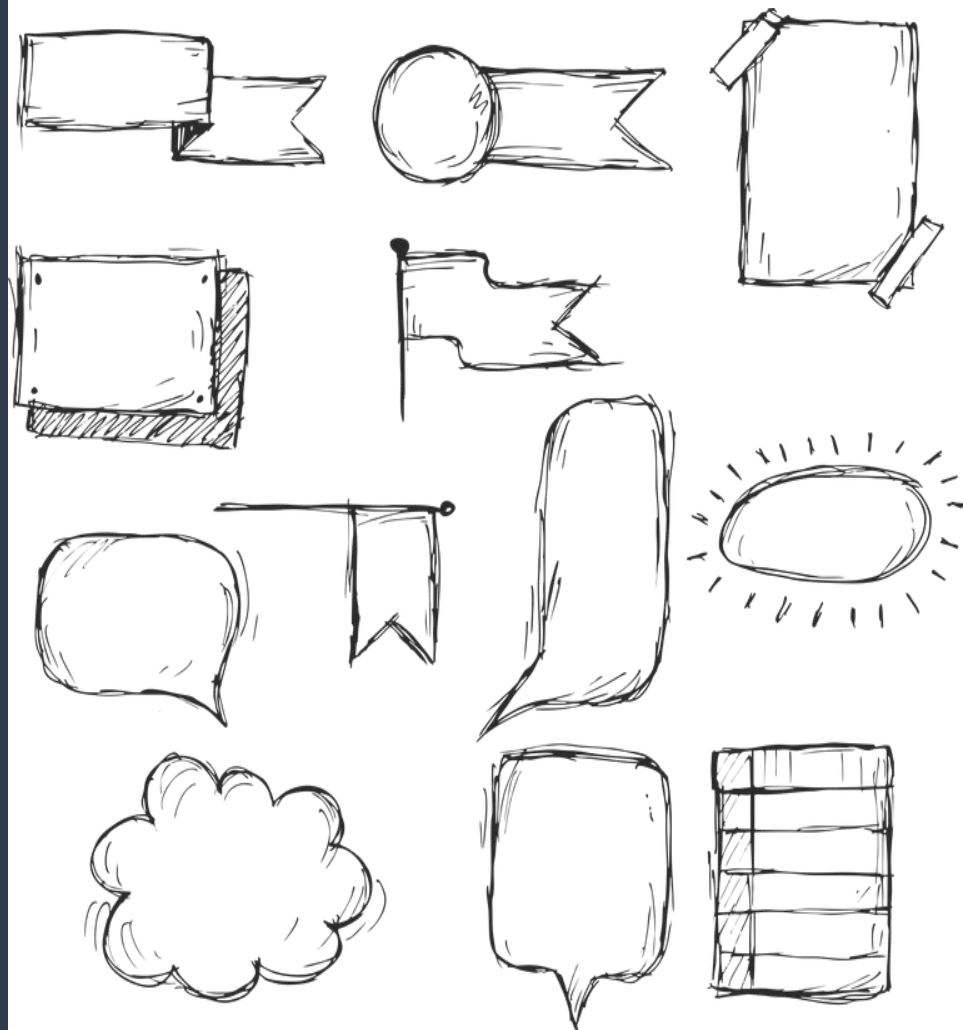
the data publisher should provide not just metadata that allows discovery, but also metadata that richly describes the context under which the data was generated. This may include the experimental protocols, the manufacturer and brand of the machine or sensor that created the data, the species used, the drug regime, etc.

'Plurality' indicates that the metadata author should be as generous as possible in providing metadata, even including information that may seem irrelevant.

# In other words:

Description of a data set is required at two different levels:

- (1) metadata describing the data set (intrinsic): what does the data set contain, how was the data generated, how has it been processed, how can it be reused ...
- (2) metadata describing the data (submitter-defined): any needed information to properly use the data, such as definitions of the variable names...





# Some hints for reusability

- Describe the scope of your data: for what purpose was it generated/collected?
- Mention any particularities or limitations about the data that other users should be aware of.
- Specify the date of generation/collection of the data, the lab conditions, who prepared the data, the parameter settings, the name and version of the software used.
- Is it raw or processed data?
- Ensure that all variable names are explained or self-explanatory (i.e., defined in the research field's controlled vocabulary).
- Clearly specify and document the version of the archived and/or reused data.

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

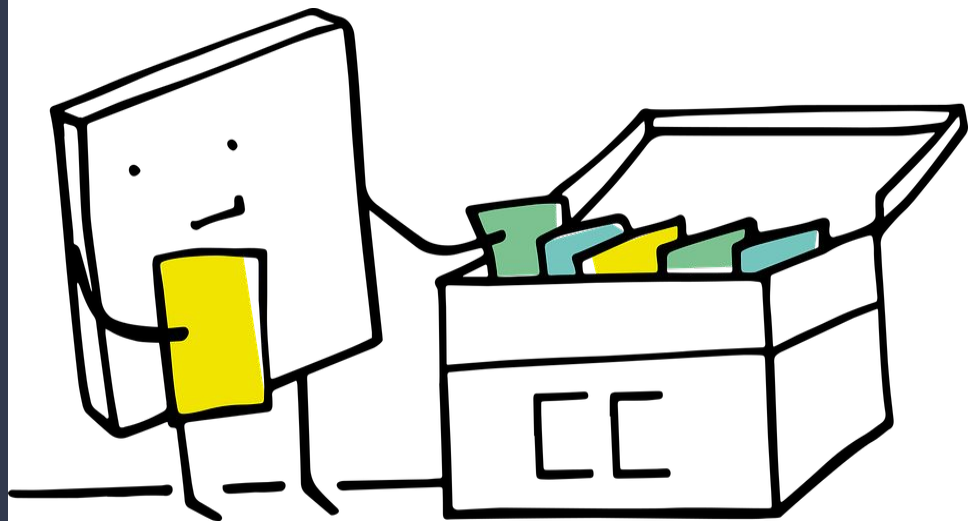
What usage rights do you attach to your data? This should be described clearly.

Ambiguity could severely limit the reuse of your data

# In other words:

The conditions under which the data can be used should be clear to machines and humans. This has to be specified in the metadata describing a data set.

Include information about the license in the metadata. If a particular license is needed, you have to provide it along with the data set. Where possible it is suggested to use common licenses, such as CC 0, CC BY, etc., which can be referred to by URL.



# Authors and rights owners

## Are you the author of the data you collected?

Yes, in case you can prove it (deposit with clear date, DOI, ... use a data repository!)

## Do you own any rights on the raw data you collected?

No, data is facts/information and none can own rights on it!












Licenses  
Tell other what they  
can do with your data


# Creative Commons


Not all of us are legal experts capable of writing proper licenses.


Creative Commons and Public Domain create legal certainty for everyone, who wants to use works, that are licensed respectively.


It is important to follow and understand the different meanings of the licenses and follow the rules for using them.


CREATIVE COMMONS LICENSES		COPY & PUBLISH	ATTRIBUTION REQUIRED	COMMERCIAL USE	MODIFY & ADAPT	CHANGE LICENSE
	PUBLIC DOMAIN	✓	✗	✗	✓	✓
	CC BY	✓	✓	✗	✓	✓
	CC BY-SA	✓	✓	✓	✓	✗
	CC BY-ND	✓	✓	✓	✗	✓
	CC BY-NC	✓	✓	✗	✓	✓
	CC BY-NC-SA	✓	✓	✗	✓	✗
	CC BY-NC-ND	✓	✓	✗	✗	✓

 You can redistribute (copy, publish, display, communicate, etc.)

 You have to attribute the original work

 You can use the work commercially

 You can modify and adapt the original work

 You can choose license type for your adaptations of the work.

# Types of CC Licenses

## **Public Domain**

Works are not covered by copyright

## **CC-0 (no rights reserved)**

Allows creators to give up their copyright and put their works into the worldwide public domain

## **CC-BY (Attribution)**

This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator

# Types of CC Licenses

## **CC-BY-SA (Attribution – ShareAlike)**

This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.

## **CC-BY-ND (Attribution – NonDerivative)**

This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, and only so long as attribution is given to the creator. The license allows for commercial use.

## **CC-BY-NC (Attribution – NonCommercial)**

This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator.



# Licensing your Research Data: Creative Commons

## FACT SHEET ON CREATIVE COMMONS & OPEN SCIENCE v.0.1



This information guide contains questions and responses to common concerns surrounding open science and the implications of licensing data under Creative Commons licences. It is intended to aid researchers, teachers, librarians, administrators and many others using and encountering Creative Commons licences in their work.

<https://doi.org/10.5281/zenodo.840651>

### What is Open Science?

**Open Science** is the movement to make scientific research and data accessible to all for knowledge dissemination and public reuse.

### How should I licence my data for the purposes of Open Science?

We recommend you use the [CC0 Public Domain Dedication](#), which is first and foremost a waiver, but [can act as a licence](#) when a waiver is not possible.

CC ZERO LICENCE, 'NO RIGHTS RESERVED' LOGO



By applying CC0 to your data you enable everyone to freely reuse your data as they see fit by waiving (giving up) your copyright and related rights in that data.

You should keep in mind that there are many situations in which data is not protected as a matter of law. Such data can include facts, names, numbers - things that are considered 'non-original' and part of the public domain thus not subject to copyright protections. Similarly, your database (which is a structured collection of data) might be considered 'non-original' and thus ineligible for copyright, and it might additionally be excluded

from other forms of protection (like the [EU sui generis database right](#), also known as the 'SGDR', for non-original databases).

In these cases, using a Creative Commons licence such as a CC BY could signal to users that you claim a copyright in the non-original data despite the law, and perhaps despite your real intention.

Finally, if your data is in the public domain worldwide, you might state simply and obviously on the material that no restrictions attach to the reuse of your data and apply a [Public Domain Mark](#).

PUBLIC DOMAIN MARK LOGO



When in doubt, consider which use may be appropriate according to the chart below:

CC0 & PUBLIC DOMAIN LICENCES WHICH LICENSE TO USE AND WHEN



'Creative arrangement' of data is original, but any copyright has been waived and content is made available copyright-free



'Creative arrangement' of data is not original; the author acknowledges this and communicates the data is in the public domain

# Licensing your Research Data: Creative Commons

- Use a CC0 or public domain, then ask for credit
- **Provide a citation** that researchers using your data can simply copy and paste to give you credit for your work
- Remember that it's bad science not to cite the source
- **CC0 does not mean academic unpoliteness**

<https://doi.org/10.5281/zenodo.840651>

**But I would like attribution when others use my dataset. In that case, shouldn't I use a CC BY licence?**

We recommend that you avoid using a CC BY licence. Here's why:

While attribution is a genuine, recognisable concern, not only might using a CC BY licence be legally unenforceable when no underlying copyright or SDR protects the work, but it may also communicate the wrong message to the world. A better solution is to use CC0 and [simply ask for credit](#) (rather than require attribution), and provide a citation for the dataset that others can copy and paste with ease. Such requests are consistent with scholarly norms for citing source materials.

Legally speaking, datasets that are not subject to copyright or related rights (and are thus in the public domain) cannot be the object of a copyright licence. Despite this, agreements based in contract law may be enforceable. Creative Commons licences, however, are copyright licences. Therefore, where the conditions for a copyright or related right are not triggered, copyright licences, such as the CC BY licence, are unenforceable.

In some cases, however, rights may exist (like the sui generis database right previously mentioned), and permission for others to use your dataset may be legally required. These rights are meant to protect the maker's investment, rather than originality. As such, database rights do not include the moral right of attribution. So by using a CC BY licence, you signal to users that you restrict access to your dataset beyond the protections provided by the law. We are not saying that this cannot be done, we are just saying that if you choose to do this, you should make sure you fully understand what it entails.

When you choose to do this, you should make sure you fully understand what it entails.

**I'm uncomfortable with others using my research for commercial purposes. Should I use a non-commercial licence for my dataset?**

We recommend you avoid using a non-commercial licence. Here's why:

For legal purposes, drawing a line between what is and is not 'commercial' can be tricky; it's not as black and white as you might think. For example, if you release a dataset under a non-commercial licence, it would clearly prohibit an organisation from selling your dataset to others for a profit. However, it might also prohibit someone using the dataset in their research if they intend to eventually publish that research. This is because most academic journals are commercial businesses that charge some sort of fee for access to their content, hence, such use could qualify as 'commercial'. Consequently, using a non-commercial licence prevents researchers from using your data in work destined for publication. This can subsequently affect the dissemination, recognition, and impact of your dataset.

**I'm uncomfortable permitting use of my research for any and all purposes. Should I use a 'No Derivatives' (ND) licence for my dataset?**

We recommend you avoid using a 'No Derivatives' licence. Here's why:

Similar to how a non-commercial licence might restrict meaningful reuse of your dataset, a ND licence can have the same effect: it may prevent someone from recombining and reusing your data for new research. For data to be truly Open Access, it must permit these important types of reuse.

**It sounds like you're really pushing for the use of CC0 for open science datasets.**

Exactly. Data is only open if anyone is free to use, reuse, and distribute it. This means it must be made available for both commercial and non-commercial purposes under non-discriminatory conditions that allow for it to be modified.

When data is made available for all reuse, others can create new knowledge from combining it. This leads to the enrichment of open datasets and further dissemination of knowledge. Accordingly, CC0 is ideal for open science as it both protects and promotes the unrestricted circulation of data.

And remember, it's bad science not to cite the source of data you use. To help others cite your data [include a citation](#) that users can copy and paste to give you credit for your hard work.

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

where the data came from (i.e., clear story of origin/history, see R1), who to cite and/or how you wish to be acknowledged. Include a description of the workflow that led to your data

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

If community standards or best practices for data archiving and sharing exist, they should be followed.

<https://www.go-fair.org/fair-principles/r1-2-metadata-associated-detailed-provenance/>

<https://www.go-fair.org/fair-principles/r1-3-metadata-meet-domain-relevant-community-standards/>

# In other words:

## R1.2

Detailed information about the provenance of data is necessary for reuse: this will, for example, allow researchers to understand how the data was generated, in which context it can be reused, and how **reliable** it is. Provenance is a central issue in scientific databases to validate data.

The metadata to thoroughly describe 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, potentially transformed or completed? Ideally the workflow is described in a machine-readable format. Criterion I3 is closely linked to this issue when reusing published data sets.

## R1.3

It is easier to reuse data sets if they are similar: same type of data, data organized in a standardized 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. Note that quality issues are not addressed by the FAIR principles. How reliable data is lies in the eye of the beholder and depends on the foreseen application.

Prepare your (meta)data according to community standards and best practices for data archiving and sharing in your research field. There might be situations where good practice exist for the type of data to be submitted but the submitter has valid and specified reasons to divert from the standard practice. This needs to be addressed in the metadata.

# To summarise...



# For next week:

1. Create an account on Zenodo sandbox:  
<https://sandbox.zenodo.org/>

2. Prepare a file to be uploaded there: a presentation, a dataset...whatever

# Thank you!

Ask questions and interact in the VRE:

[https://services.d4science.org/group/phdunipi\\_os21-22](https://services.d4science.org/group/phdunipi_os21-22)

gina.pavone@isti.cnr.it



Consiglio Nazionale  
delle Ricerche

