Digital Object, FAIR principles for research products, Scientific Data Repositories

Massimiliano Assante, PhD

Institute of Information Science and Technologies (ISTI) - CNR

https://orcid.org/0000-0002-3761-1492

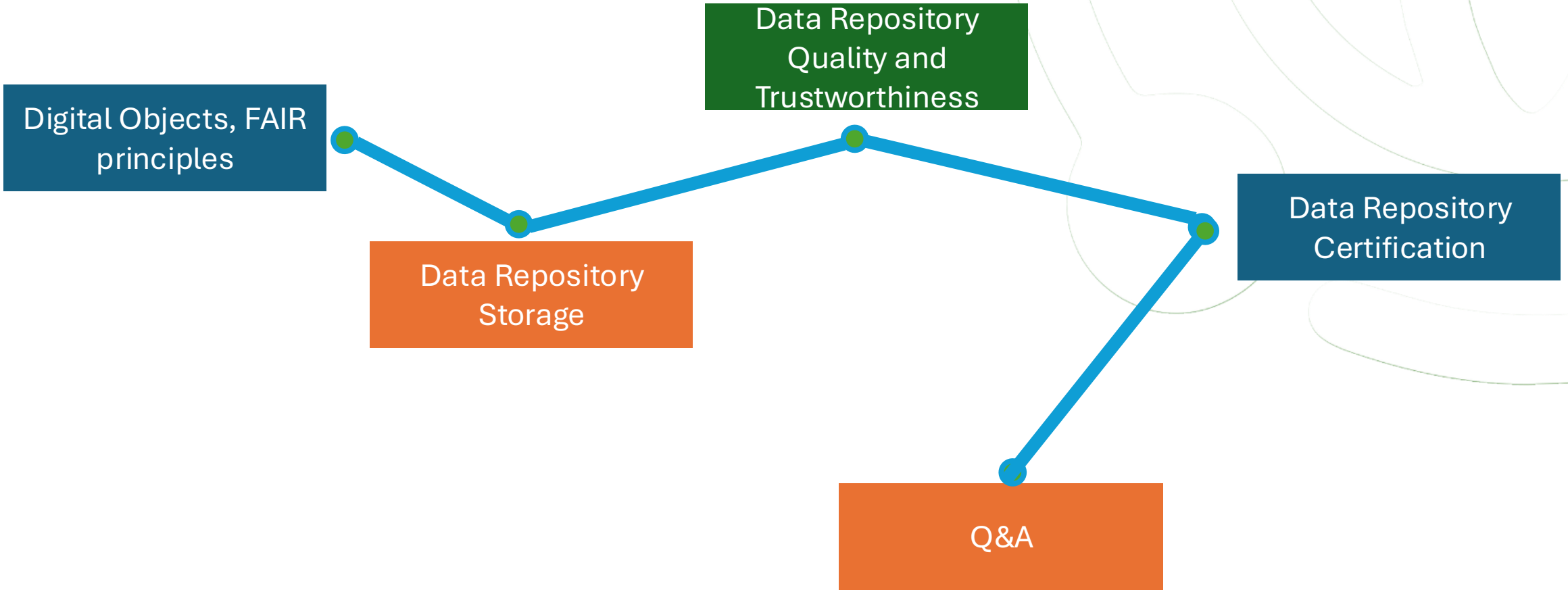
IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System (D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

Today's Journey



Digital Objects, what are they?

A **Digital Object** is any research result in its digital form that can be stored (and possibly shared).

Examples: articles, dataset, software, images, videos, reports, conference poster or presentations, lectures, etc

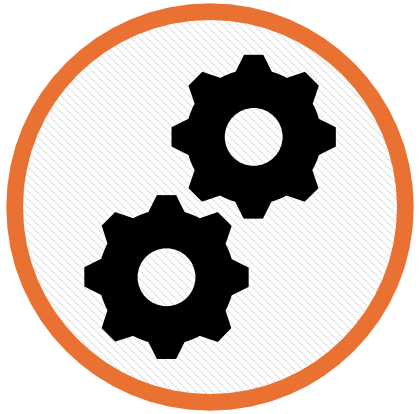
FAIR principles in a nutshell

Findable

Other can find your data

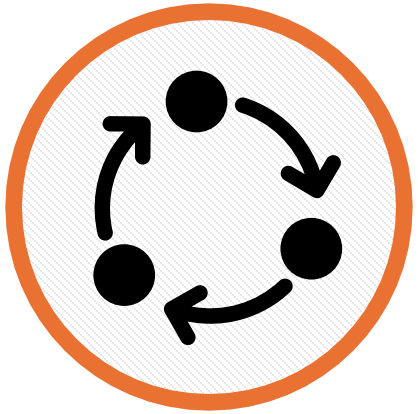
Accessible

Your data is accessible to others



Interoperable

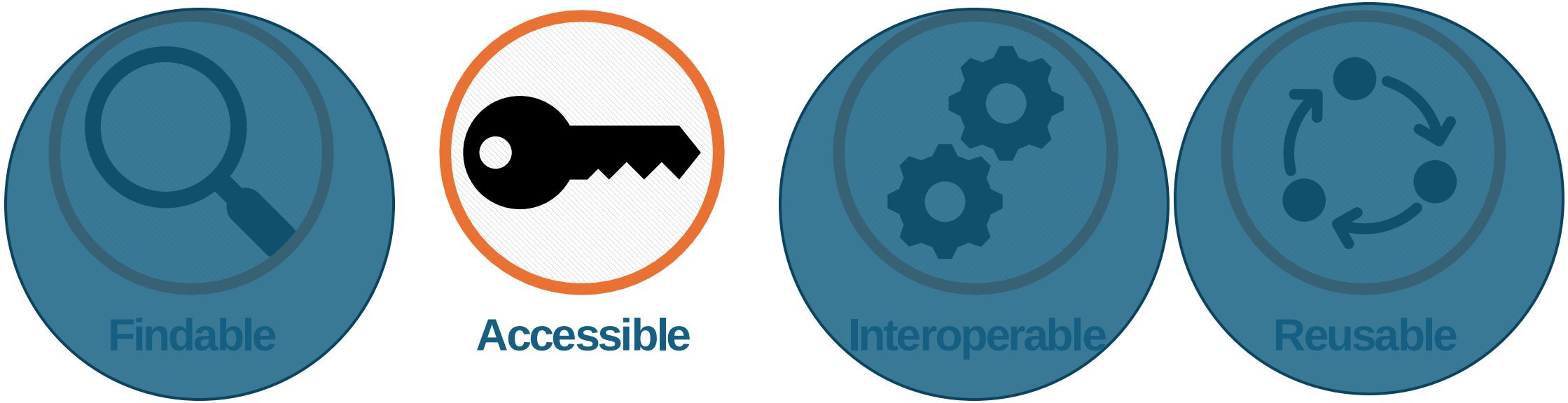
Your data can be integrated with other data and/or they can be easily used and read by machines.



Reusable

Your data can be reused by others in new research

Accessible



Your data is accessible to others

Accessible

- Can I make my data accessible to others?
- Who will be granted access?
- How?

How do you give acce s to ou data?

Through a Data Reposi r

Data Repositories

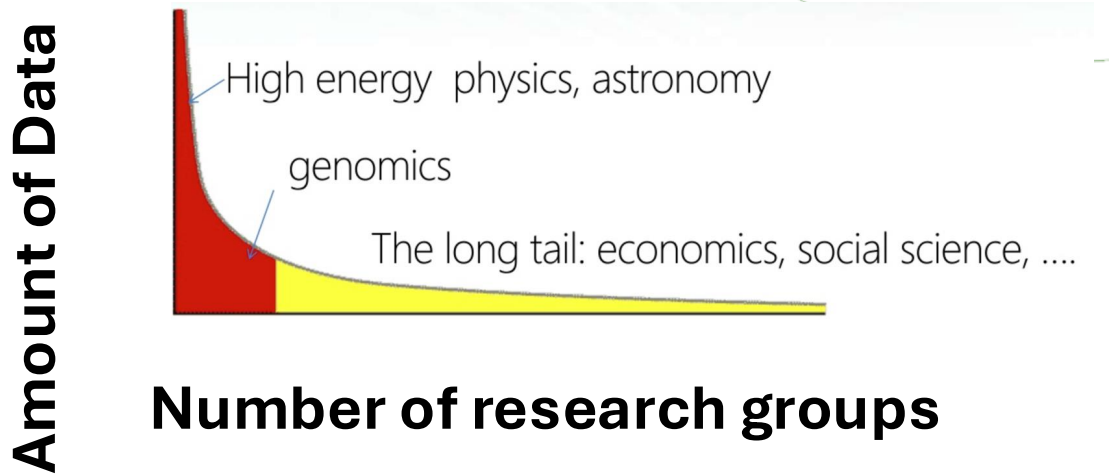
Defined as "*a subtype of a sustainable information infrastructure which provides long-term storage and access to research data*" (re3data.org) .

Provide sustainable preservation to the data created by researchers.

Ensure research data is accessible beyond the life of a grant, research project, or individual careers.

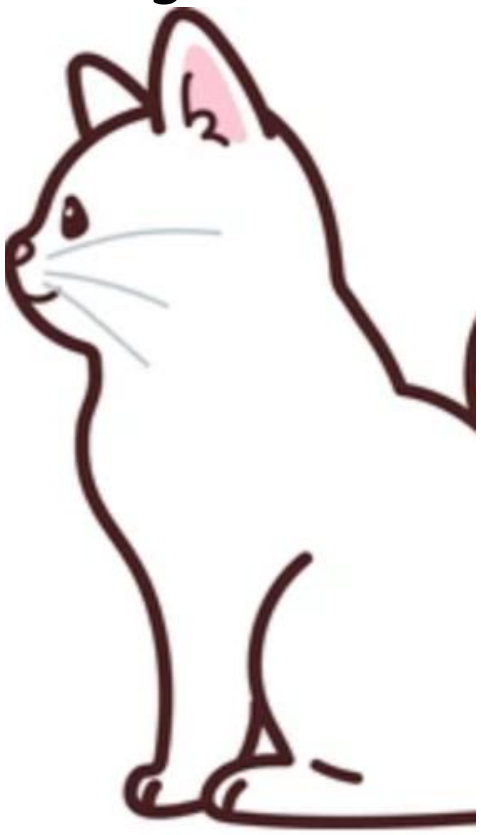
Data Repository, scientific communities

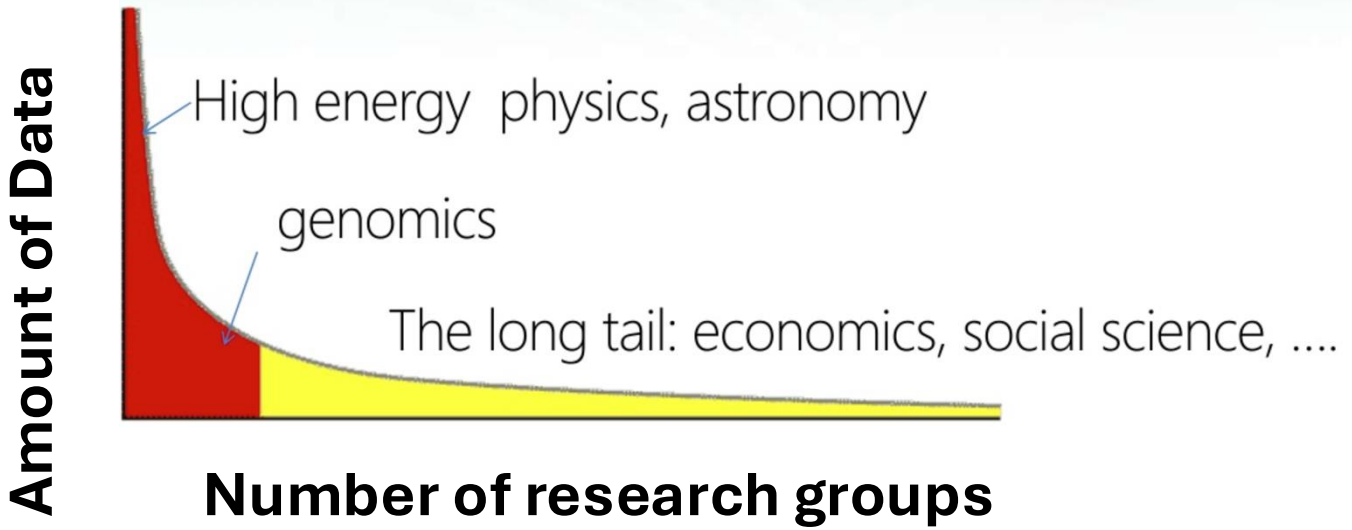
- Largely diffused within communities that produce a huge amount of data, such as physics (e.g., LHC at CERN), genetics (e.g., GenBank at NCBI Data), or environmental sciences (e.g., British Atmospheric Data Centre).
- However, the demand for data repositories is emerging also within the so-called "*long-tail*" of science.



Data Repository, scientific communities: long-tail

- However, the demand for data repositories is emerging also within the so-called "*long-tail*" of science.

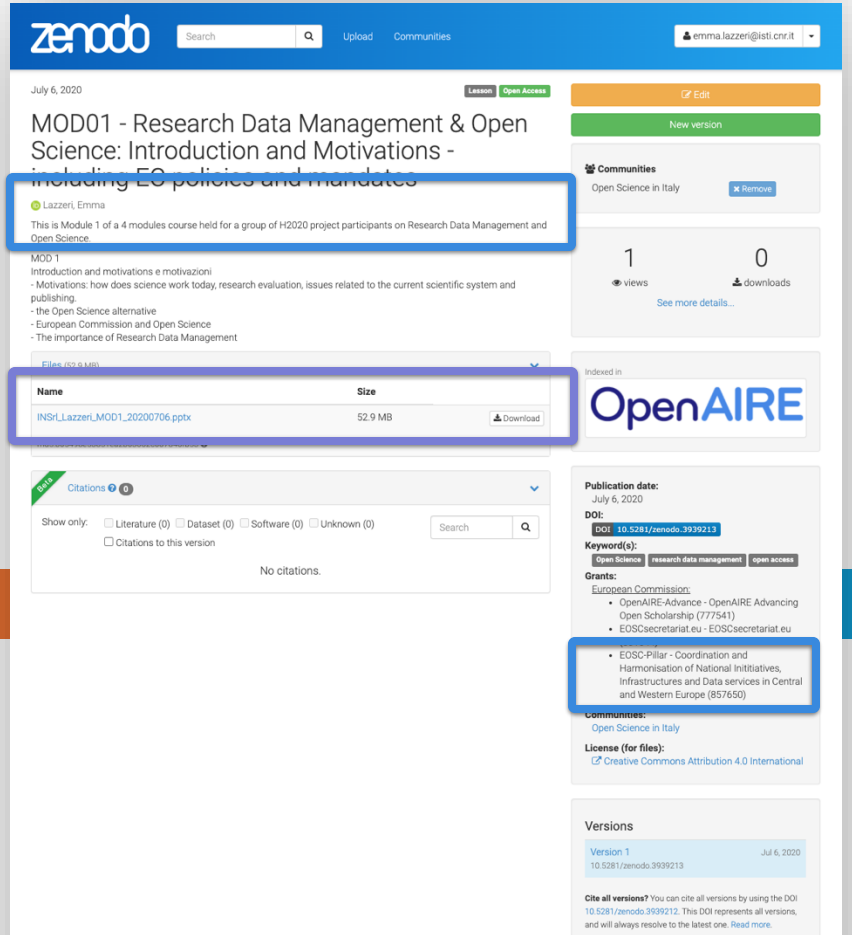




Data Repositories Storage: Records deposited

Metadata

A set of data describing the digital object(s) you are depositing



Payload

The digital object(s) you are uploading to be stored. Payload includes attached files such as the file containing the data and the accompanying material(s), readmefile, etc

Data Repository Storage: Metadata and Payloads

Metadata

- Generally stored on Relational Database Management Systems;
- Daily or hourly backup cycles;
- Some kind of indexing and search feature is generally supported, e.g. Lucene, ElasticSearch, Solr etc.

Payload

- Dedicated services offering petabytes disk clusters (Data storage services);
- Digital object replicas located on different disk servers;
- MD5 checksums for data integrity;
- Different type of data persistence (e.g. SSD, Magnetic Disk, Tapes) for long-term storage cost optimisation.

An example of data storage technology

EOS Open Storage: CERN storage technology used at the LHC

EOS provides a service for storing large amounts of physics data and user files, with a focus on interactive and batch analysis.

Flexible

EOS is a storage solution for central data recording, analysis and processing++

Adaptable and Scalable

EOS supports thousands of clients with random remote I/O patterns with multi protocol support
WebDAV, CIFS, FUSE, XRootd, GRPC.

Over 350 PB at CERN

Designed for high capacity and low latency.

Security

EOS offers a variety of authentication methods:KRB5, X509, OIDC, shared secret, and JWT and proprietary token authorisation.

Sync & Share

EOS provides Sync&Share functionality for the CERNBox front-end services.

Tape Storage

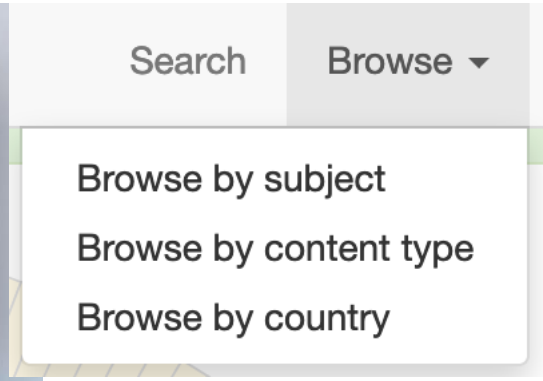
EOS includes tape storage in combination with the CTA Cern Tape Archive software

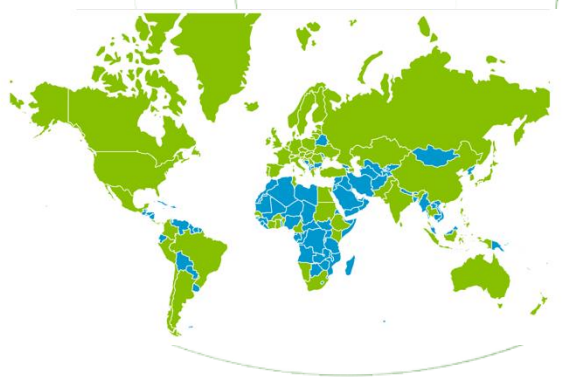
Which repository to use?

1. use a disciplinary repository if there is one;
2. use the institutional repository, if you have one where the data will also be available for the long term;
3. search the global re3data.org for a suitable repository;
4. Can’t find one? use the catch-all repository Zenodo

3. Where to find a suitable research data repository

www.re3data.org





re3data.org +2700 repositories

Search Browse Suggest Resources Contact DataCite

Filter

- Subjects
- Content Types
- Countries
- AID systems
- API
- Certificates
- Data access
- Data access restrictions
- Database access
- Database access restrictions
- Database licenses

social sciences and humanities — Search

Toogle short help

← Previous 1 2 3 4 5 6 7 … 35 Next → Sort by

Found 857 result(s)

TESS

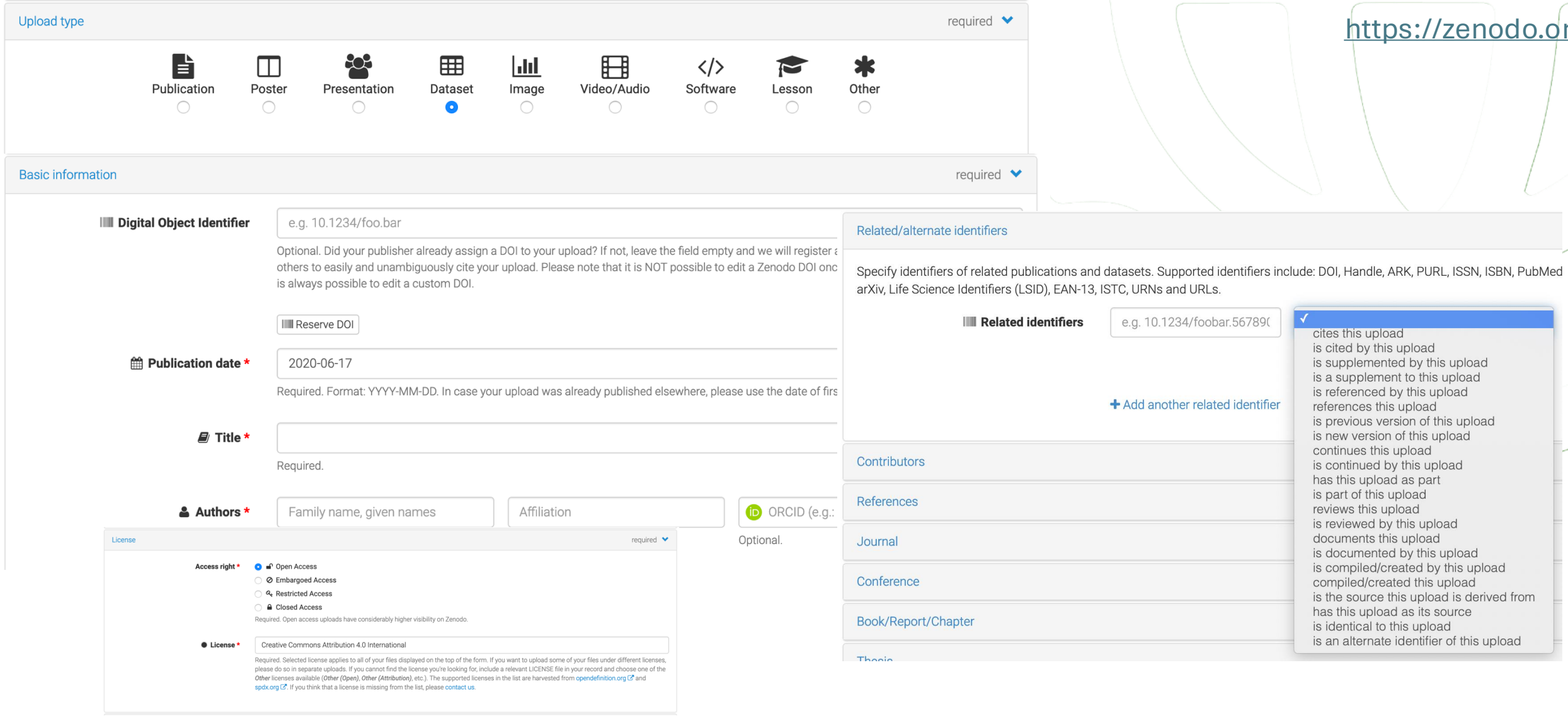
Time-sharing Experiments for the Social Sciences

Subject(s): Social Sciences | Social and Behavioural Sciences | Humanities and Social Sciences

Content type(s): Standard office documents | Plain text

Can’t find one? Try Zenodo

https://zenodo.org



Quality of Data Repositories

- Not easy to evaluate;
- Influenced by many external factors, starting with the mission of the repository:
  - does it explicitly aim for long-term preservation - with the appropriate expertise and budget - or not?
  - Is it dedicated to a specific research community and familiar with their data formats, or is it generic?

A Data Repository of quality should adheres to majority of the FAIR principles, it should at least support:

Metadata Standards;

Persistent and Global Unique Identifiers;

Access rights and Licenses;

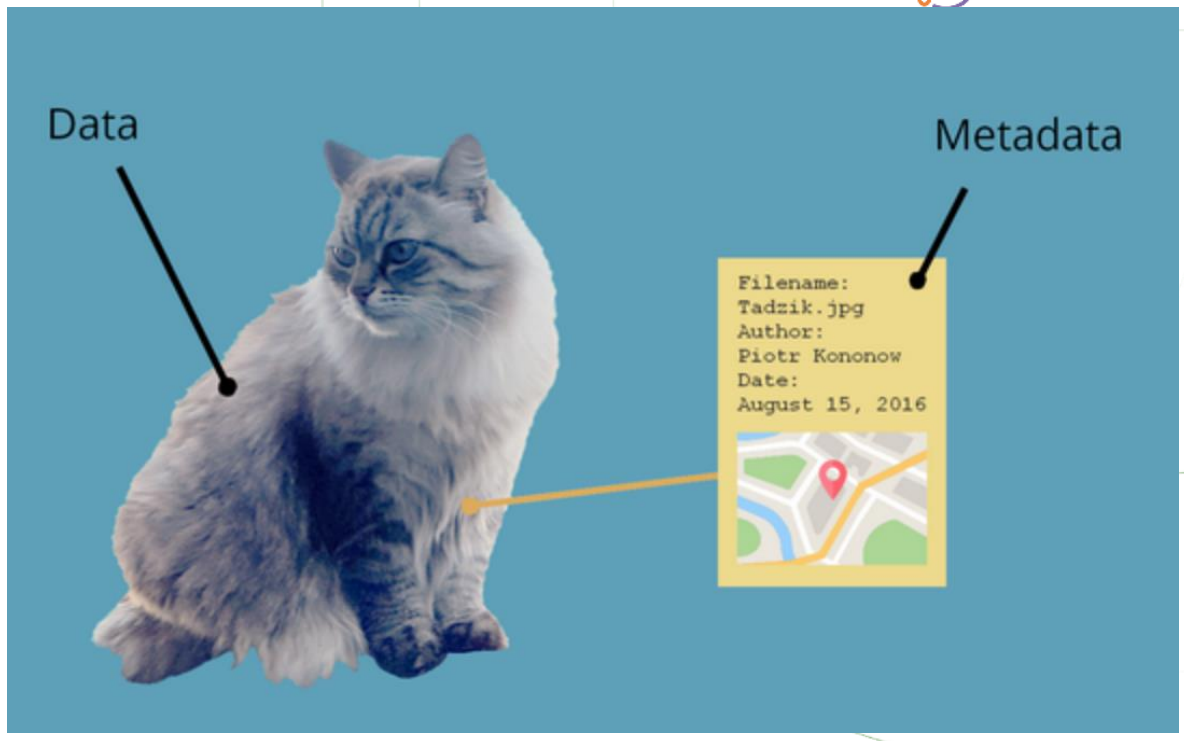
Data citation;

*Be possibly C**ertified. (if certified*** the quality assessment has already been done for you)

Data Repository quality

Support ***Metadata Standards;***

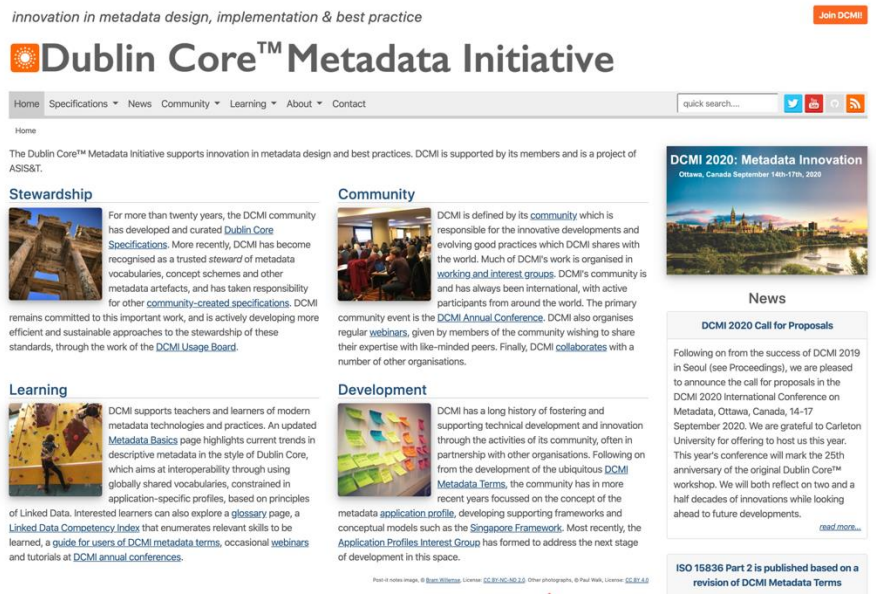
Metadata



- Data describing data
- Very important for:
  - Access
  - Comprehension
  - Process
- Use your discipline specific standards: you will spend less time interpreting data and more time to actually make science!

Dublin Core

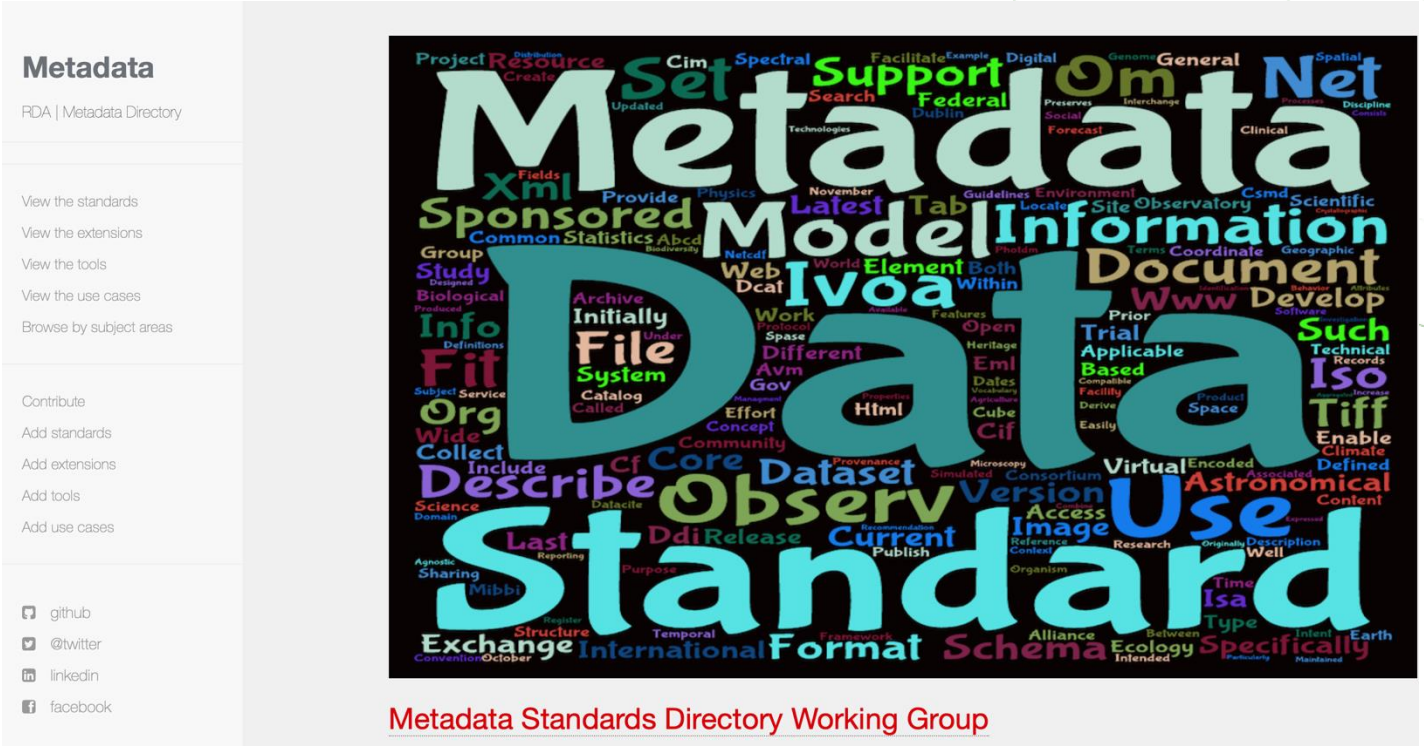
- **The Dublin Core**, a set of fifteen generic, widely used elements:
- *Creator, Contributor, Publisher, Title, Date, Language, Format, Subject, Description, Identifier, Relation, Source, Type, Coverage, and Right.*
- It was first drafted at a 1995 meeting in Dublin, Ohio (USA).



Where to find Metadata?

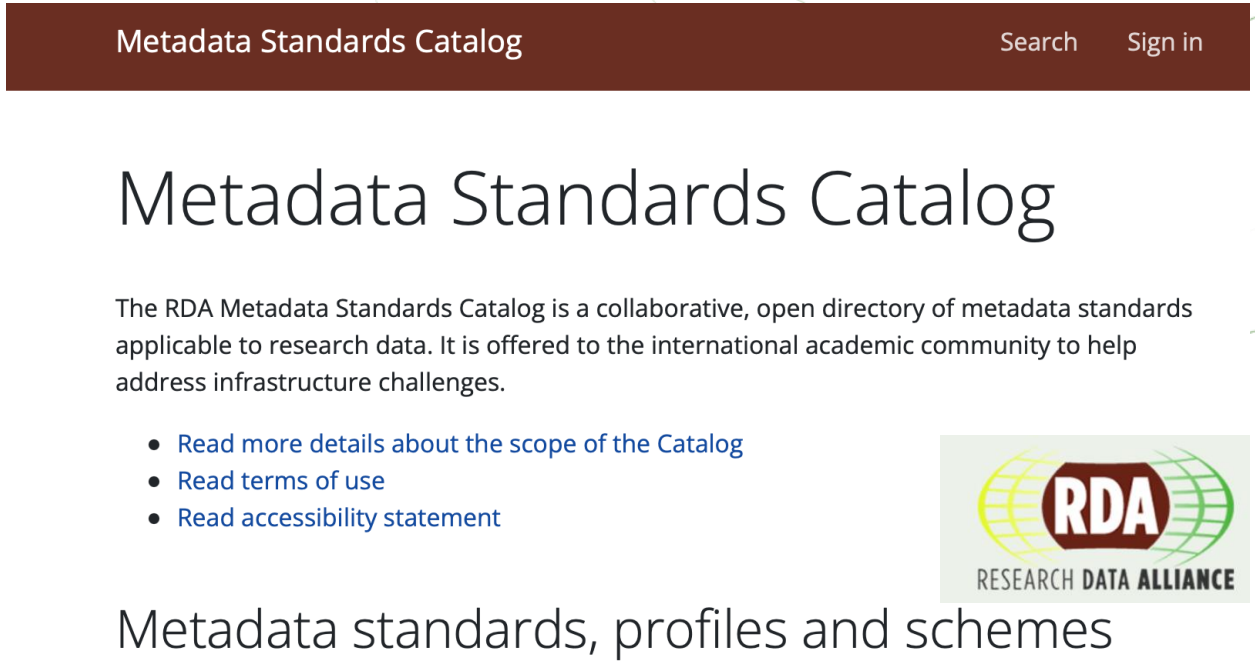
Metadata Directory:

http://rd-alliance.github.io/metadata-directory/



Where to find Metadata?

- Metadata Standards Catalogue: https://rdamsc.bath.ac.uk
  - Contains metadata standards that may be used to document research data



Data Repository quality

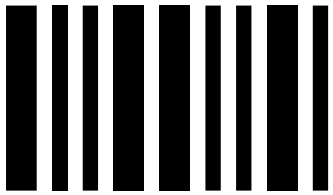
Provide Persistent and Global Unique Identifiers;

Persistent identifiers make your data

- Findable
- Accessible

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 Objec** 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 bar code used on products …



Persistent Identifiers: Digital Object Identifier (DOI)

- 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 which aims only to identify their referents uniquely;

DOI: an example

Are Scientific Data Repositories Coping with Research Data Publishing

Authors: Assante M. Candela L. see more

DOI: 10.5334/DSJ-2016-006

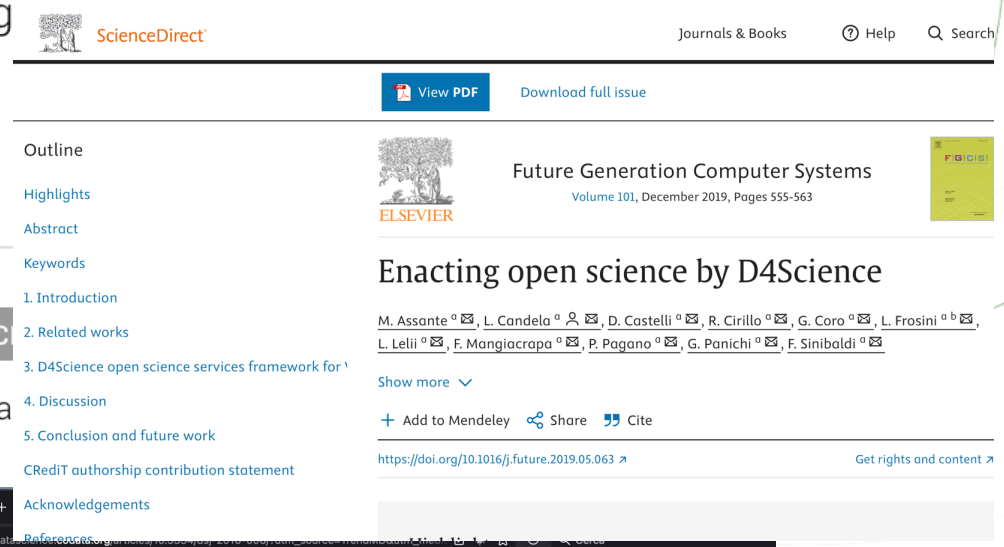
Repositories for Open S ce: The SciRepo Reference Model WEB OF SC

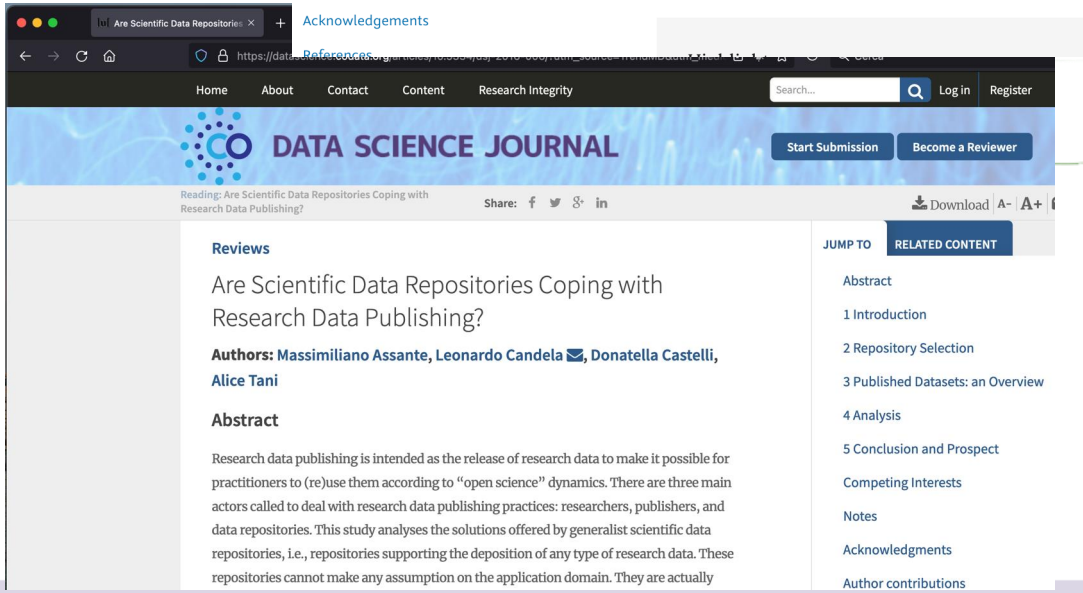
Authors: Assante, Mas ano; Candela, Leonardo; Castelli, Donatella; ... Pagano, Pa

Published: Nov 2015 in munications in Computer and Information Science

DOI: 10.1007/978-3-31 29-6_26

https://doi.org/10.5334/DSJ-2016-006





How can you assign a Persistent Identifier to your digital object?

- Persistent identifiers need to be assigned by an entity that can ensure the persistency of the link to the object;
- Zenodo assigns DOIs to digital objects that do not already have one

DOI Registration Agencies

DataCite Crossref mEDRA 中文DOI Japan Link Center

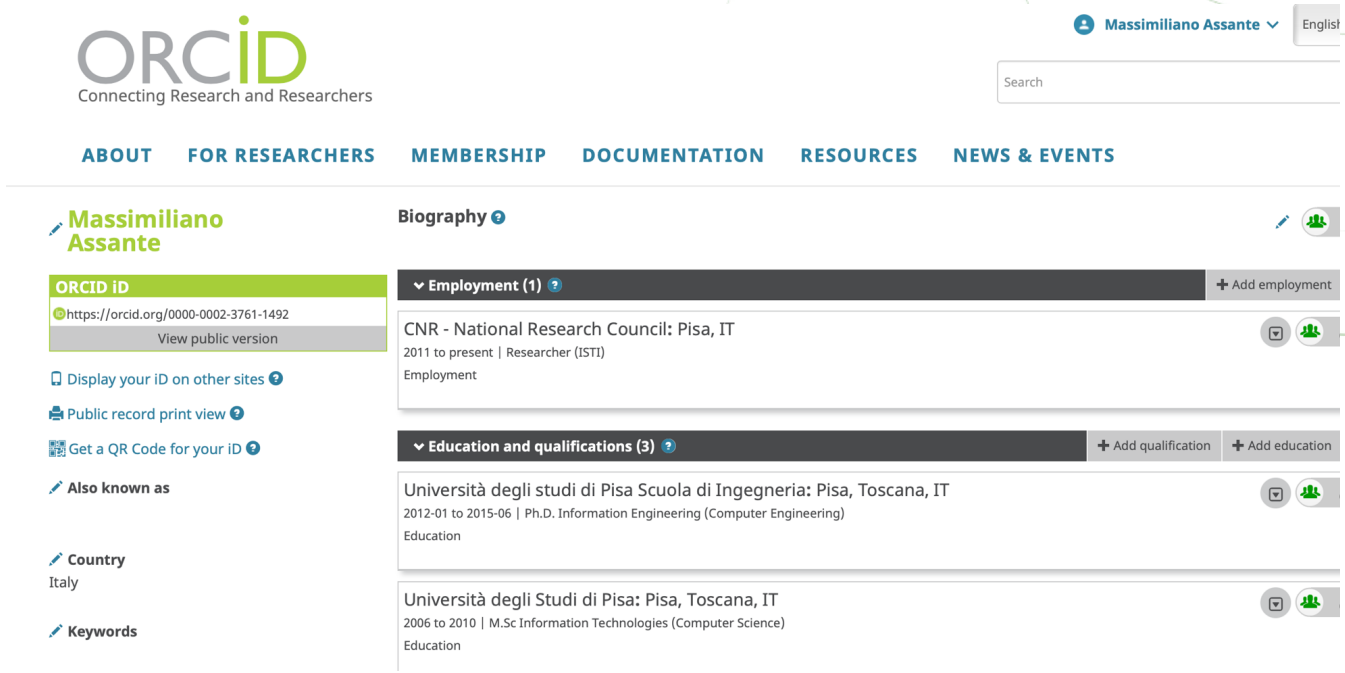
KiSTi www.kisti.re.kr airiti DOI

ORCID : Open Researcher and Contributor ID

It is a non-proprietary alphanumeric code to uniquely identify scientific and other academic authors and contributors.

e.g.: https://orcid.org/0000-0002-3761-1492

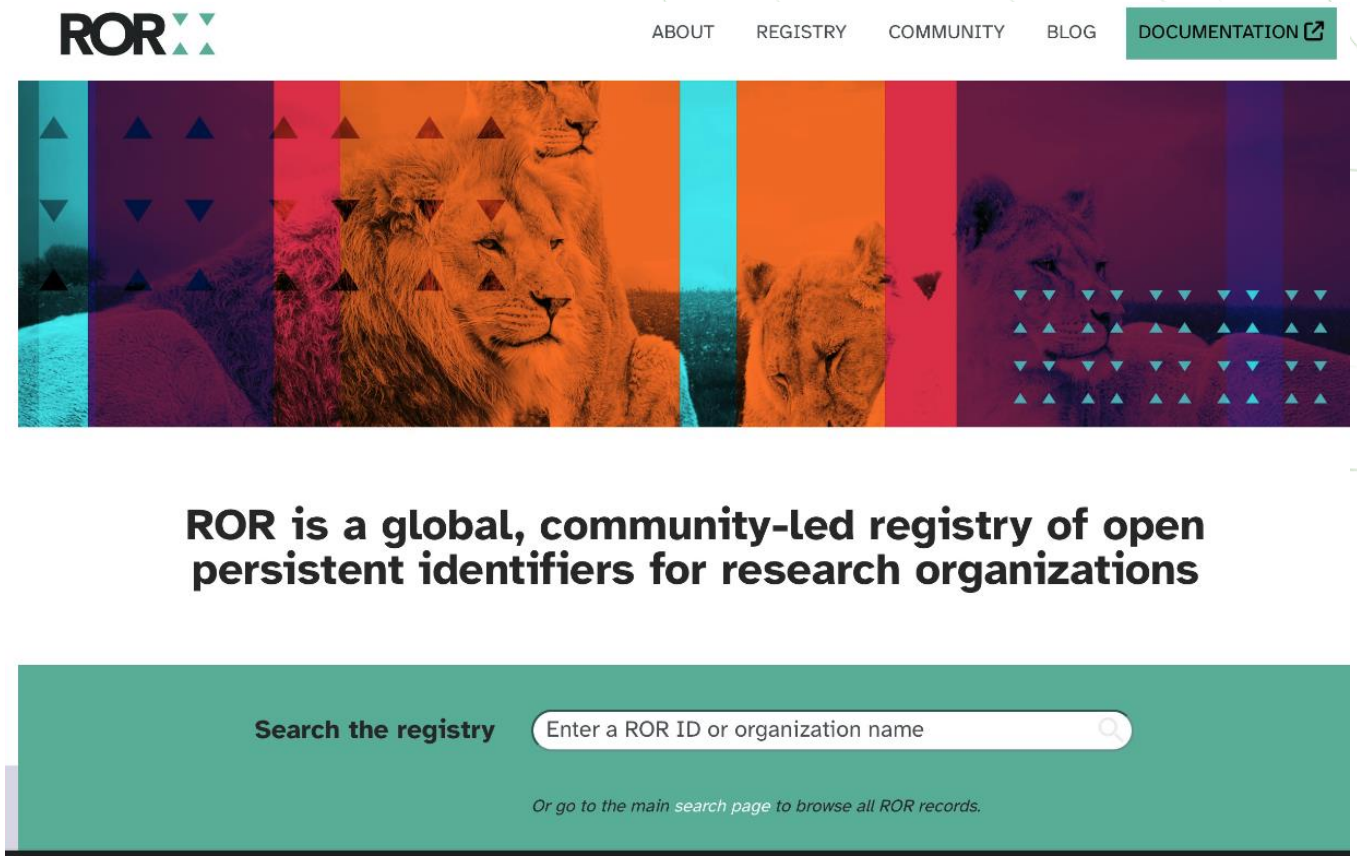
Do you have one?



ROR : Research Organization Registry

For Research organization identifiers (founded Jan 2019)

https://ror.org



Data Repository quality: access rights, licenses

Support ***access rights, licenses;***

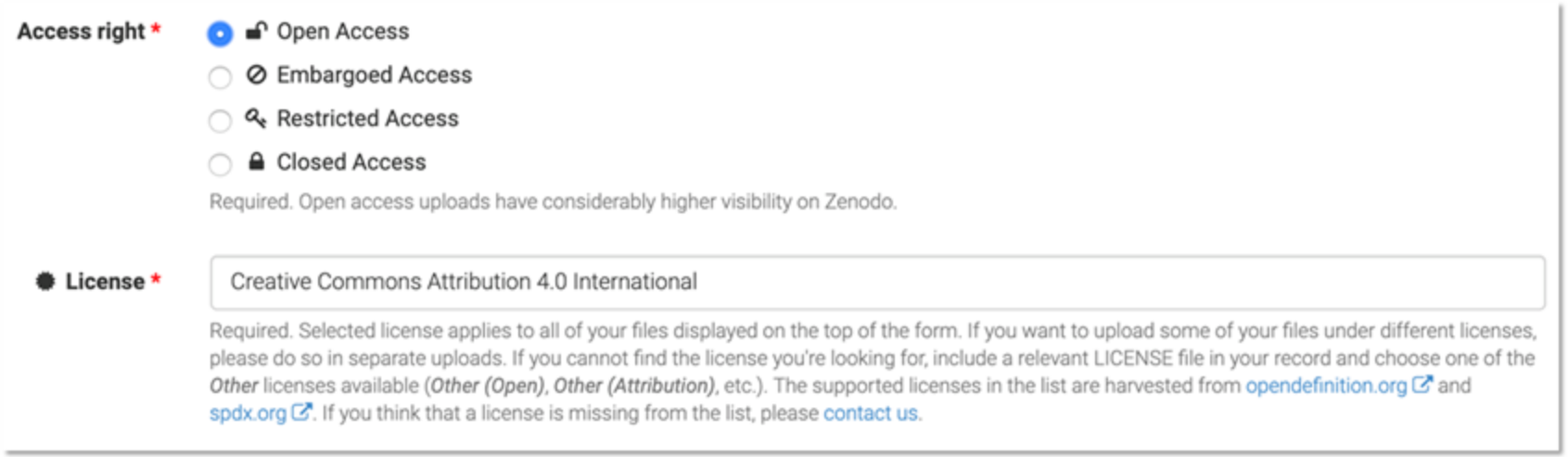
Deposit vs. Access: an imp rtant iffe nce

Deposit: upload a digital object (data, icles. …) on a platform that allows to correct describe the object through medatada d that implements long-term preservati

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)

Photo by Jason Dent on U

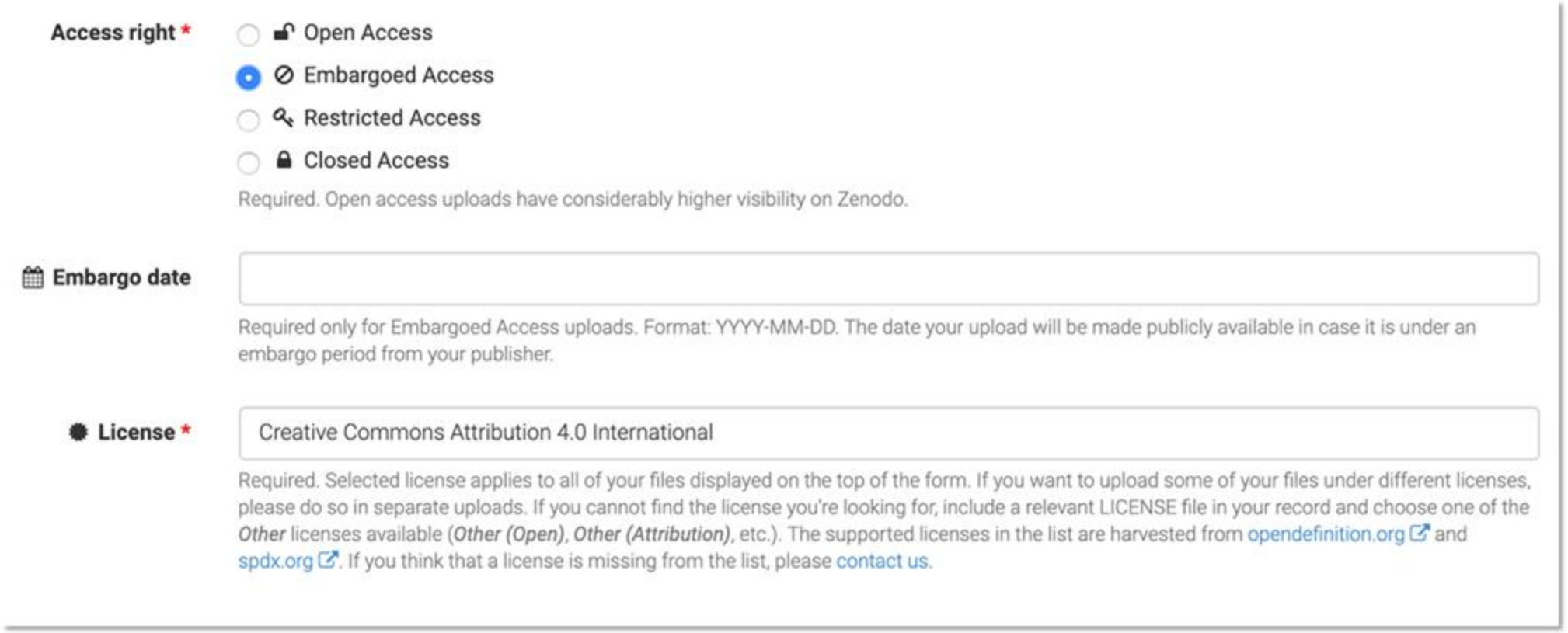
Access Right: Open Access



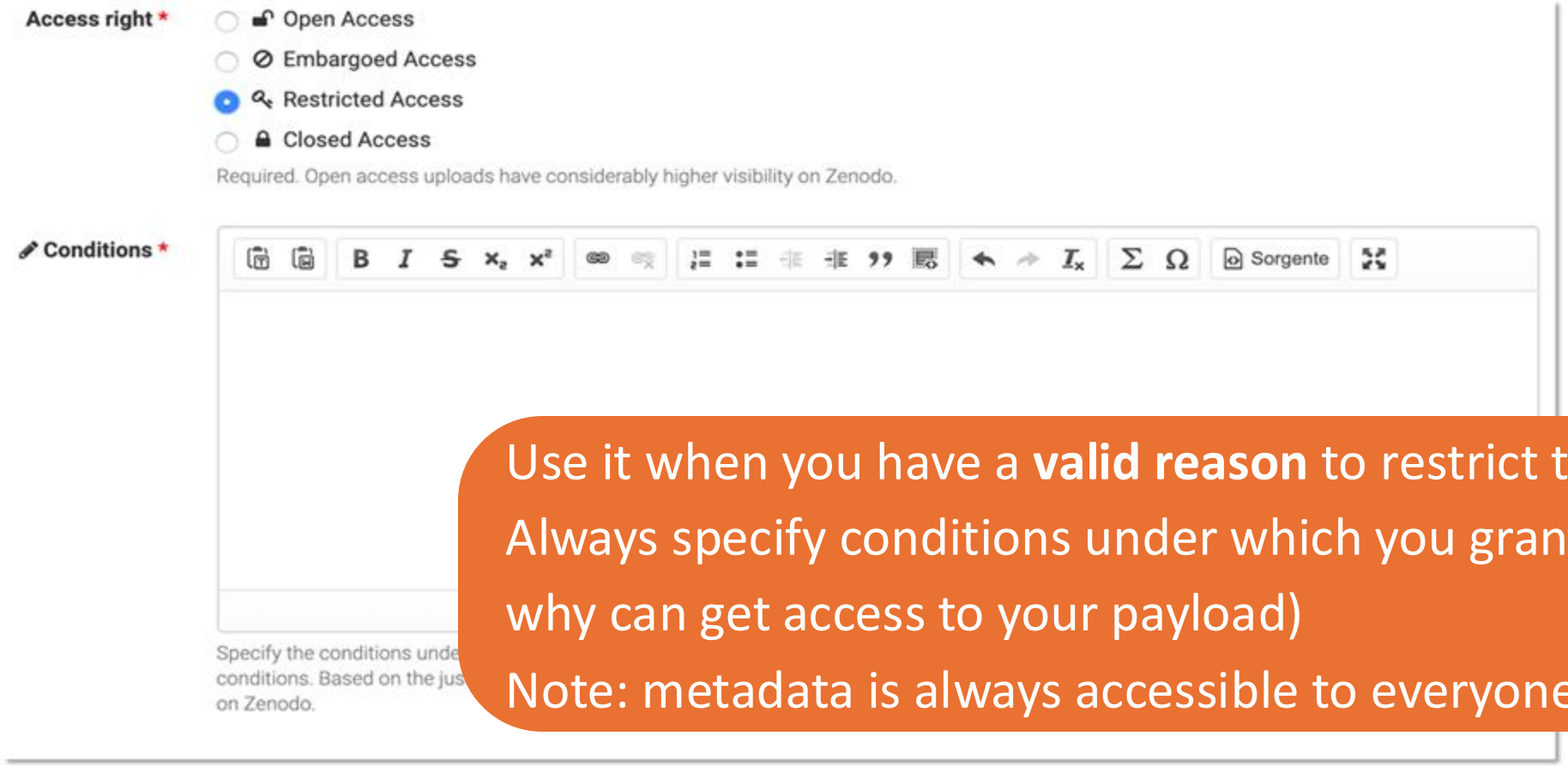
This should be the default access right

Always assign a licence for reuse

Access Right: Embargoed Access



Access Right: Restricted Access

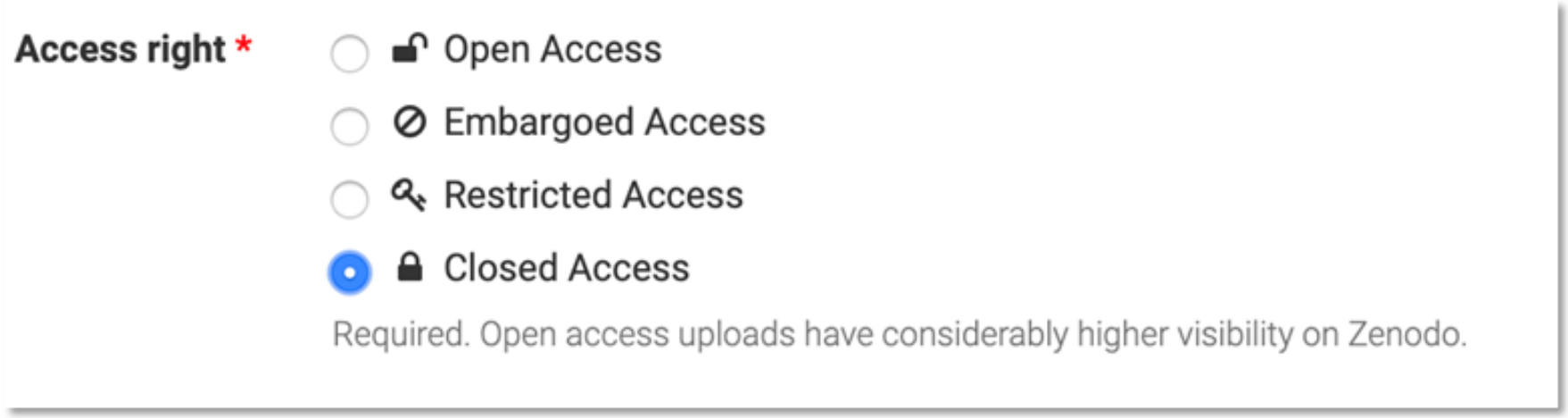


Use it when you have a **valid reason** to restrict the access

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

Note: metadata is always accessible to everyone

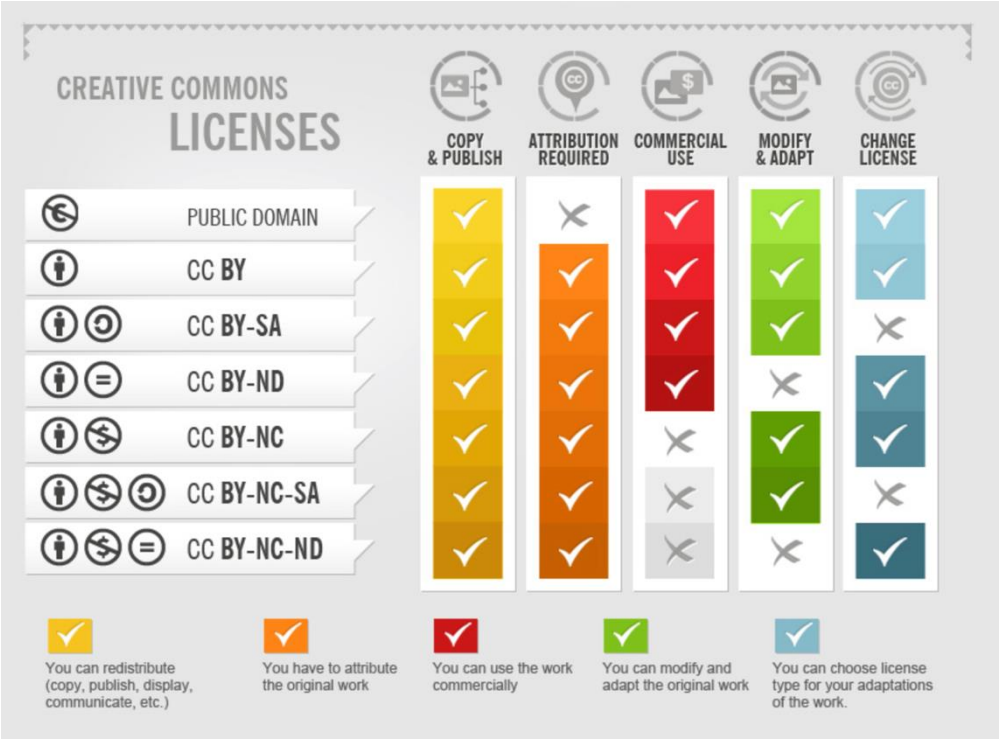
Access Right: Closed Access



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

Licenses: 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.



Licenses: Creative Commons

- **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

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

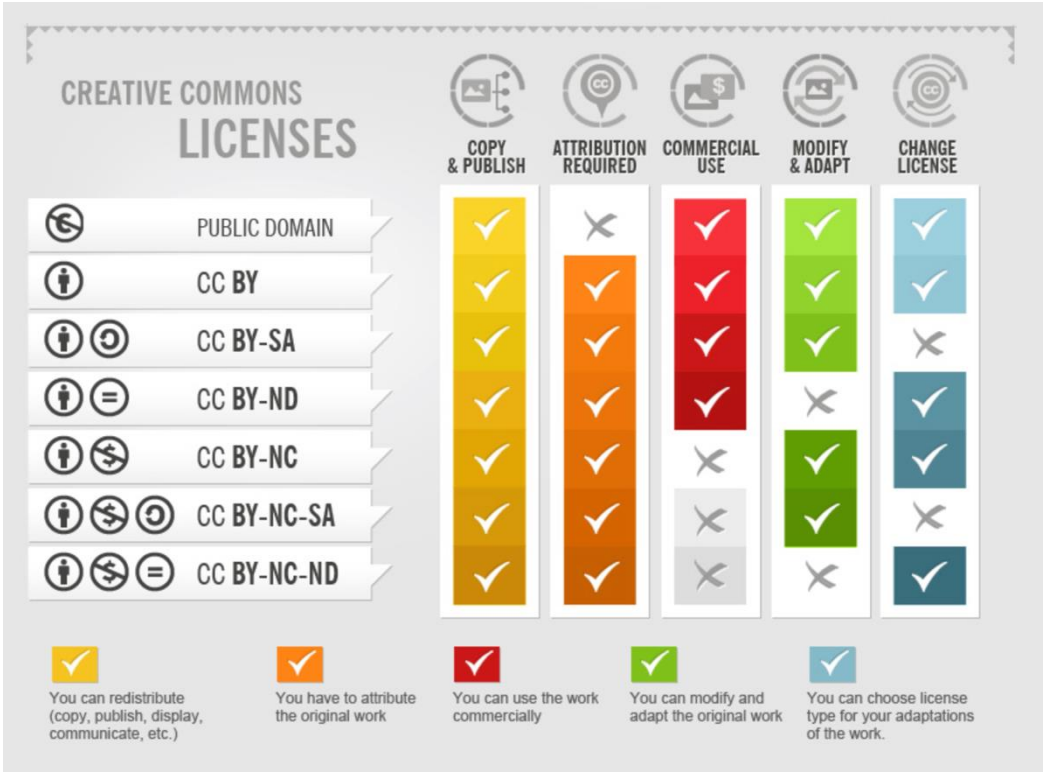
This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use. If you remix, adapt, or build upon the material, you must license the modified material under identical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

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

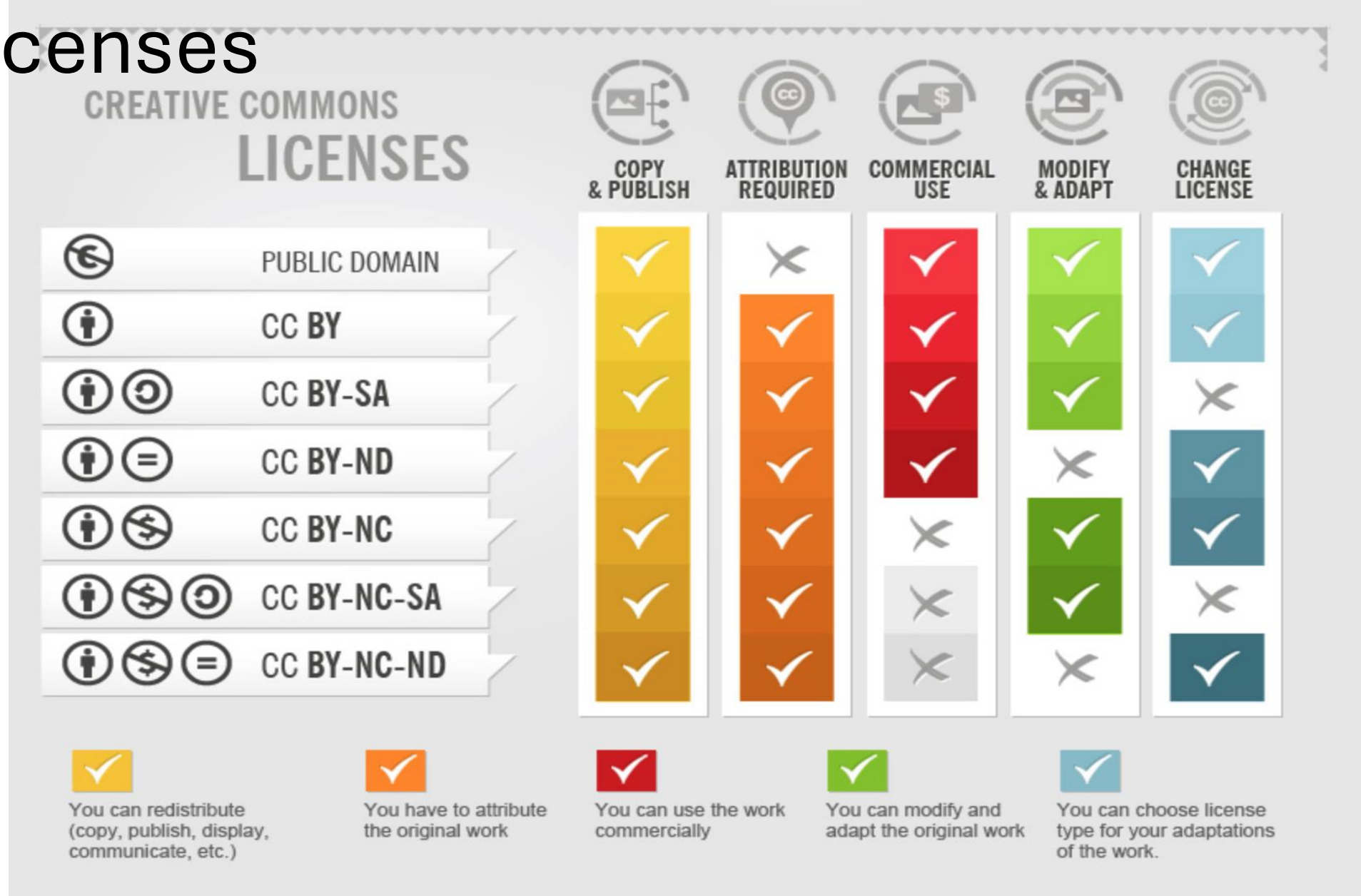
This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, and only so long as attribution is given to the creator. The license allows for commercial use.

- **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.



Licenses: Creative Commons Licenses



Data Repository quality: Data Citation

Support Data citation;

Data Citation – what is it?

The provision of a reference to dataset(s) describing data properties enabling credits and attribution, discover, interlinking, and access.

To support attribution, reward and encourage the reuse of data for new research questions.

Why it is important to cite data

- To support proper *attribution* and *credit*;
- To support *collaboration* and *reuse* of data;
- To enable *reproducibility* of findings;
- To foster *faster* and more *efficient* research progress;
- To provide the means to *share data* with future researchers.

Source:
https://datacite.org/cite-your-data.html

Components of a data citation

very similar to citing publications; many "correct" formats to use. It is suggested including the following information:

- creator(s) or contributor(s);
- title of dataset;
- date of publication;
- (persistent) identifier (e.g. DOI) or URL;
- Publisher;
- version, if relevant;
- Last date accessed, if relevant;

Data citation examples

Cite APA

Assante, M. (2015). *FigShare harvested dataset licenses at 31-12-2014* [Data set]. figshare. https://doi.org/10.6084/M9.FIGSHARE.1311684

Cite Harvard

Assante, M. (2015) "FigShare harvested dataset licenses at 31-12-2014." figshare. doi: 10.6084/M9.FIGSHARE.1311684.

Cite Chicago

Assante, Massimiliano. "FigShare Harvested Dataset Licenses at 31-12-2014." figshare, 2015. https://doi.org/10.6084/M9.FIGSHARE.1311684.

DataCite Search

Locate, identify, and cite research data with the leading global provider of DOIs for research data

https://search.datacite.org/

26,240,291 Works

2,302,649 People

2,436 Repositories

Find what you're looking for by searching millions of records with extensive, reliable metadata.

Share your data and reuse the data of others to create the highest impact in the research community.

Cite your research sources with confidence, and receive proper credit when your work is reused.

Connect your research – publications, datasets, software, authors, institutions, and funding data all in one place.

Create DOI on Datacite Video

https://www.youtube.com/watch?v=yr6M6x5z3XU

Data Repository Certification

Data Repository Certification

(if certified the quality assessment has already been done for you)

Importance of Certification

- ***National and international funders*** are increasingly likely to mandate open data and data management policies that ***call for the long-term storage and accessibility of data.***

- If we want to be able ***to share data, we need to store them in a trustworthy data repository***. Data created and used by scientists should be managed, curated, and archived in such a way to preserve the initial investment in collecting them.

- Certification can be an important contribution to ***ensuring the reliability and durability of data repositories*** and hence the potential for sharing data over a long period of time. ***By becoming certified, repositories can demonstrate to both their users and their funders*** that an independent authority has evaluated them and endorsed their trustworthiness.

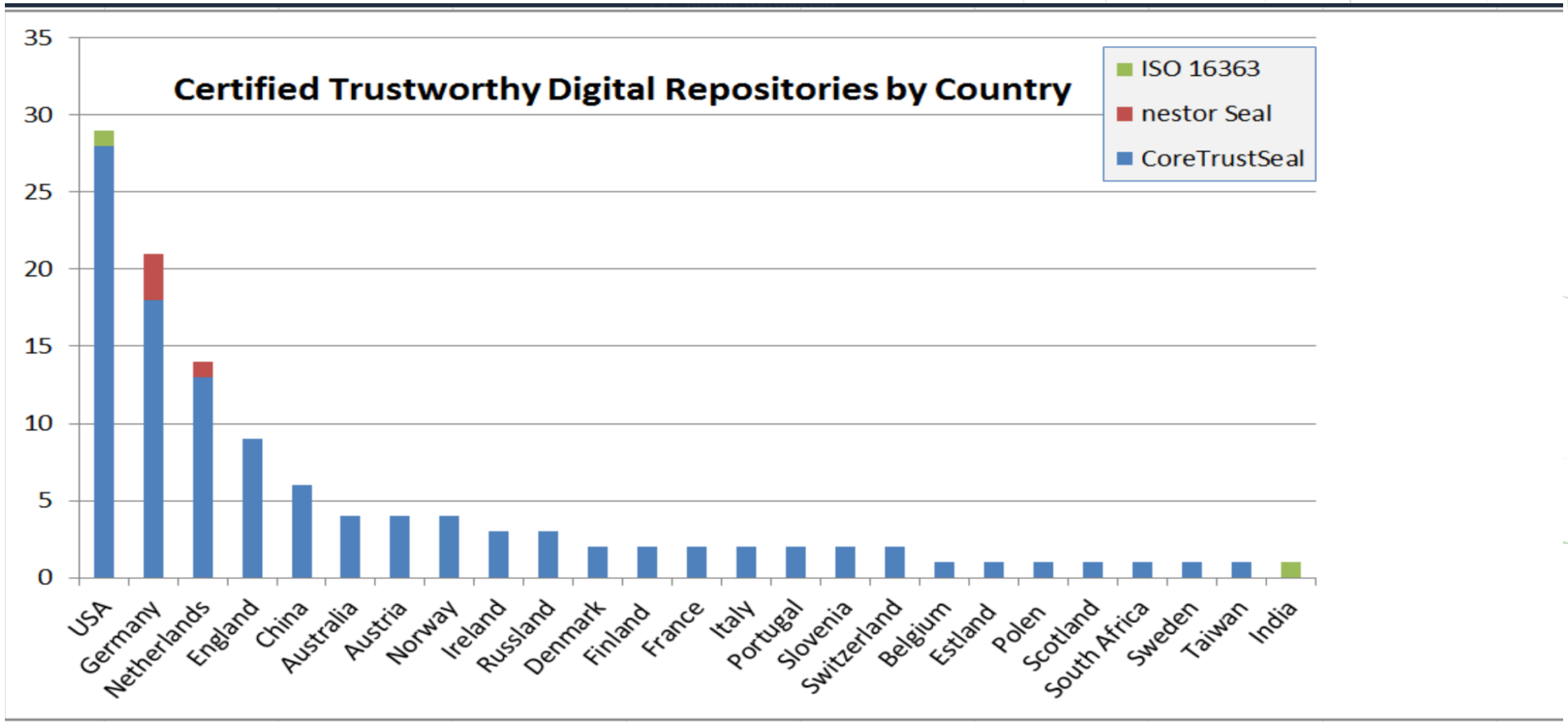
CoreTrustSeal Standards and Certification Board. (2019). CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022 (v02.00-2020-2022). Zenodo. https://doi.org/10.5281/zenodo.3638211

European Framework for audit and certification

3 certification instruments, with increasing degrees of complexity and depth, are available:

- CoreTrustSeal (CTS)
- Nestor-Seal
- ISO 16363

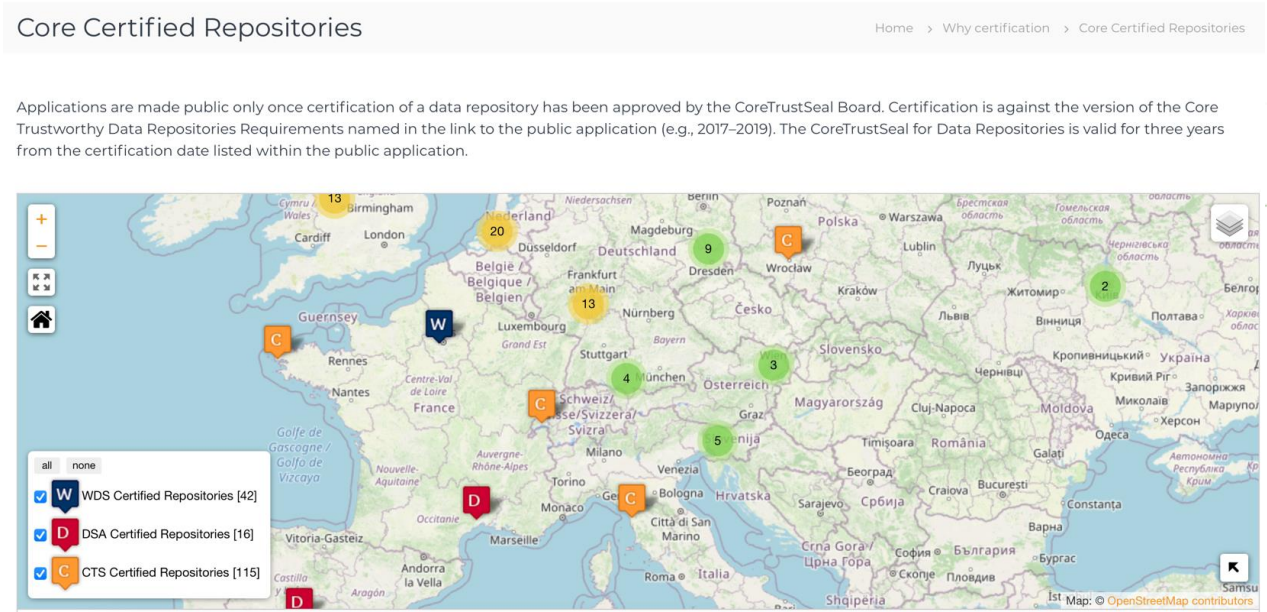
Certified Repositories by Country at 2022



Core Trust Seal (CTS) is the most common used

What is CoreTrustSeal (CTS)

- A certification organisation launched by WDS (World Data System of the International Science Council) and DSA (Data Seal of Approval).
- All repositories having CTS cert. listed at https://www.coretrustseal.org/why-certification/certified-repositories/



Core Certification and its Benefits

Core certification involves a minimally intensive process whereby data repositories supply evidence that they are sustainable and trustworthy:

- A repository first conducts an internal self-assessment, which is then reviewed by community peers.
- The repository complies with the 14 requirements from CTS

CoreTrustSeal Standards and Certification Board. (2019). CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022 (v02.00-2020-2022). Zenodo. https://doi.org/10.5281/zenodo.3638211

CoreTrustSeal (CTS) Requirements (1/2)

Main requirements that a repository with a CTS cert. adheres to (R0 and R1 are about context and mission of the repo):

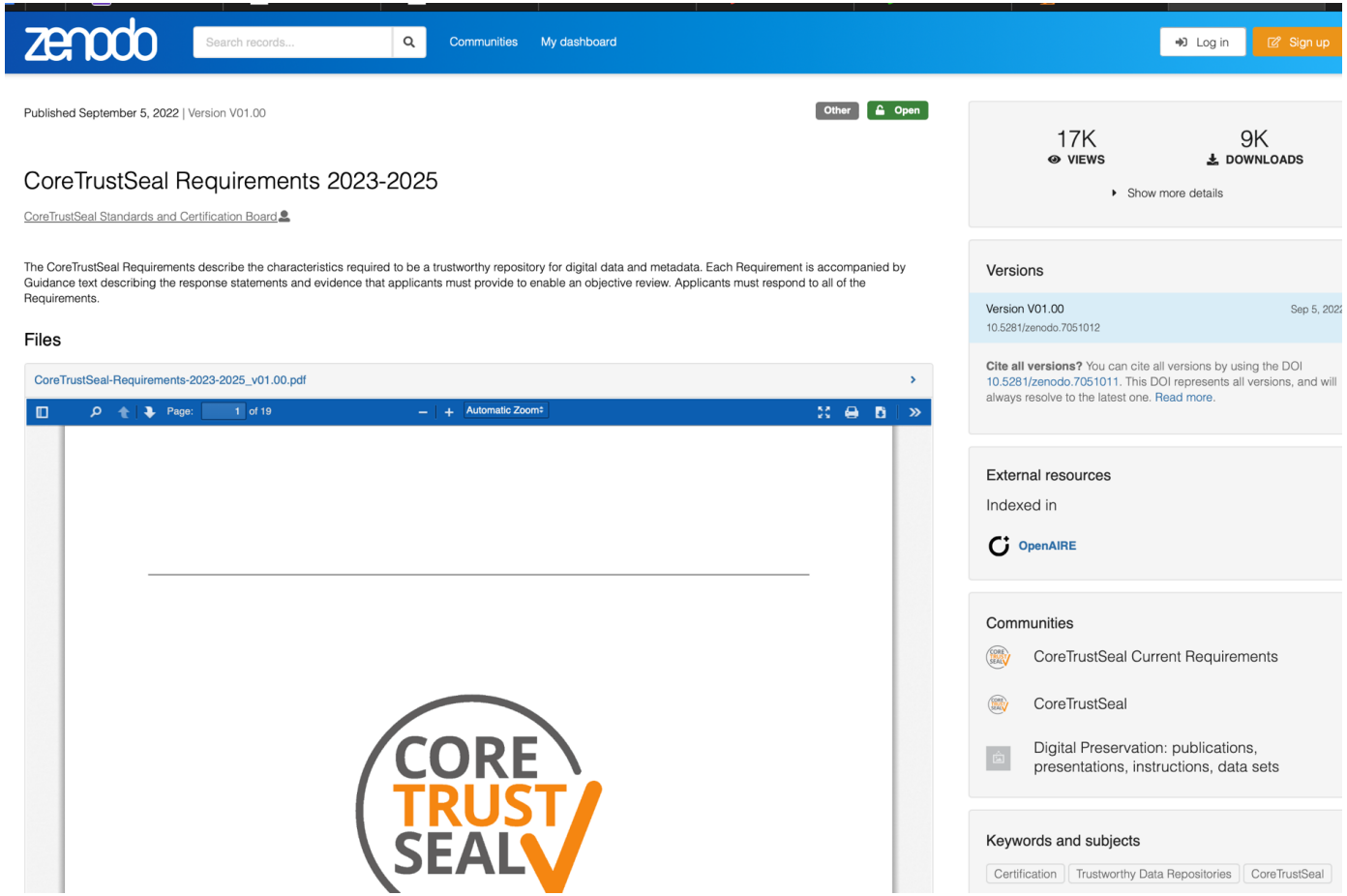
- **R2:** The repository maintains all applicable **licenses covering data access** and use and monitors compliance.
- **R3**: The repository has a continuity plan to ensure **ongoing access** to and preservation of its holdings.
- **R7**:The repository guarantees the **integrity and authenticity of the data.**
- **R8**: The repository accepts **data and metadata** based on defined criteria to ***ensure the relevance and understandability for users***.

CoreTrustSeal (CTS) Requirements (2/2)

- **R10:** The repository assumes responsibility for the **long-term preservation** and manages this function in a planned and **documented** way.
- **R11:** The repository has appropriate expertise to address **technical data and metadata quality** and ensures that **sufficient information** is available for end users to make quality-related evaluations.
- **R13**: The repository enables users to **discover the data** and **refer to them in a persistent way** through proper citation.
- **R14**: The repository enables reuse of the data over time, ensuring that **appropriate metadata** are available to support the understanding and use of the data.

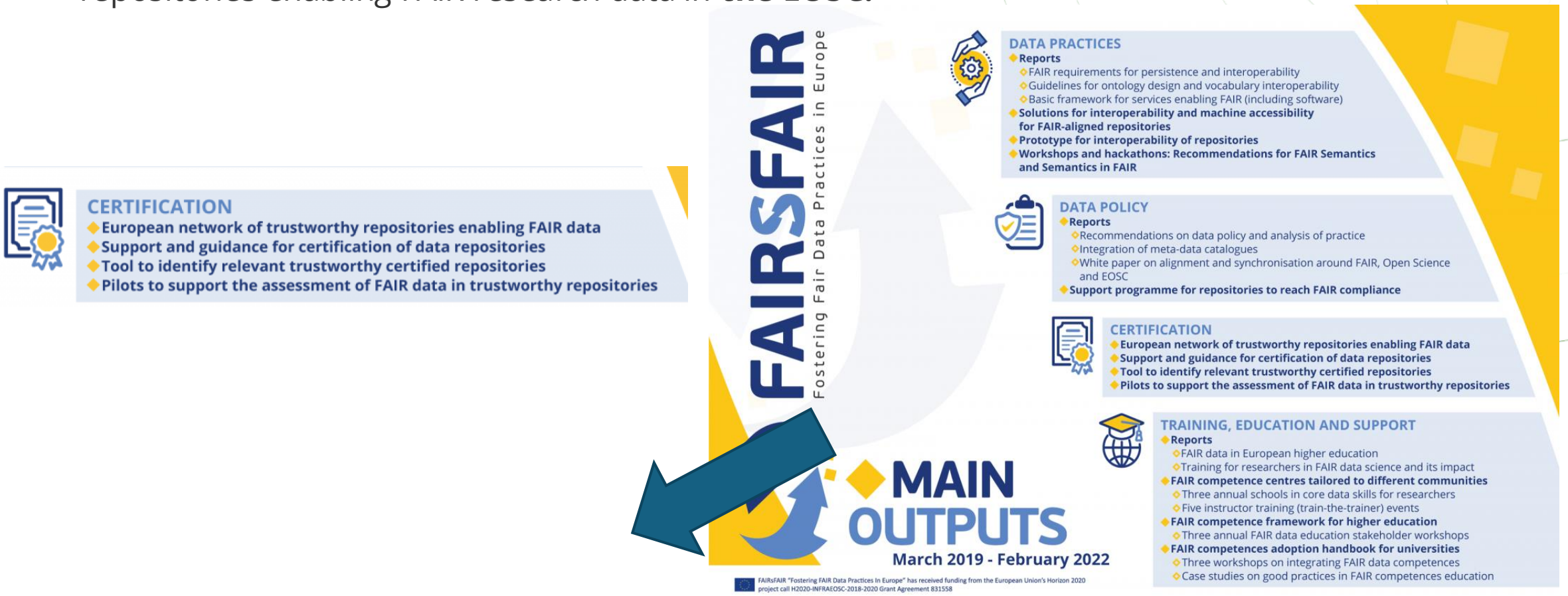
CoreTrustSeal Standards and Certification Board. (2019). CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022 (v02.00-2020-2022). Zenodo. https://doi.org/10.5281/zenodo.3638211

More on CoreTrustSeal (CTS) Requirements



Repositories FAIR Certification is ongoing

WP4 will support the co-development and implementation of certification schemes for trusted data repositories enabling FAIR research data in **the EOSC.**



Thanks!

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani

THANKS!

IR0000032 – ITINERIS, Italian Integrated Environmental Research Infrastructures System
(D.D. n. 130/2022 - CUP B53C22002150006) Funded by EU - Next Generation EU PNRR-Mission 4 “Education and Research” - Component 2: “From research to business” - Investment 3.1: “Fund for the realisation of an integrated system of research and innovation infrastructures”

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani