

Understanding Data Repositories

A Beginners' Guide for Researchers & Data Stewards (v3.0)

Table of contents

- [Glossary](#)
- [Repository vs. Archive](#)
- [Data repositories: what and why?](#)
- [What's in it for you as a researcher?](#)
- [Types of Data Repositories](#)
- [FAIR Data Repositories](#)
- [Characteristics of Trustworthy Data Repositories](#)
- [Certified Data Repositories](#)
- [Finding Trustworthy Data Repositories](#)
- [Data Appraisal, Data Selection and Data Depositing](#)
- [Final remarks](#)

Glossary and Other Concepts

Glossary (1)

- **Data publishing**

“Research data publishing is the release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for reuse and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way. Data publishing occurs via dedicated data repositories and/or (data) journals which ensure that the published research objects are well documented, curated, archived for the long term, interoperable, citable, quality assured and discoverable—all aspects of data publishing that are important for future re-use of data by third party end-users.”

Source: Austin, C.C., Bloom, T., Dallmeier-Tiessen, S. e.a., Key components of data publishing: using current best practices to develop a reference model for data publishing, *International Journal on Digital Libraries*, 18, 77–92 (2017). <https://doi.org/10.1007/s00799-016-0178-2>.

➤ *This means the researcher **releases their data for others to access/use**. They publish their data in dedicated data repositories and/or (data)journals in order to achieve this goal.*

- **Data archiving**

Data archiving is the practice of making and keeping information accessible so the information can be reusable, now and in the future. Not all information will be permanently kept, this happens on the basis of [appraisal and selection](#). In order to be permanently accessible, preservation activities need to occur.

Source: [Nationaal archief – Wat betekent archiveren?](#)

➤ *This is the practice of making information accessible and consultable for **the long term** (>10y). This practice requires planning, procedures, continuous management (cf. preservation activities), structural resources.*

Glossary (2)

• Preservation activities

“The management and protection of digital information to ensure authenticity, integrity, reliability, and long-term accessibility. This includes the utilization of management activities, strategies, best practices and standards, and policies and procedures to **guarantee ongoing access** to digitized and born-digital information, **despite the challenges of technological change, media deterioration, hardware/software obsolescence, human error, and intentional harm**. Digital preservation efforts seek to provide accurate and authentic rendering of content, while ensuring its future functionality and usability over time.”

Source: [Dictionary of archives terminology \(SAA\) – Digital preservation](#)

➤ *These are all technical activities needed to make sure data/information is **consultable and usable** for the **long term** (>10y).*

Examples of preservation activities:

- If a .pdf is no longer used in 20 years, all information that is stored in a .pdf format may need to be converted to a format that is readable. Ideally, the original format needs to be preserved.
- To make sure data will remain FAIR, there needs to be a technology watch. This means that the data will be regularly checked to see if they are still findable, accessible, interoperable and reusable.

Glossary (3)

- **Research data repository**

An online research data repository is a database infrastructure for the management, the storage and dissemination of research data. The repository ensures its database is searchable and the data is discoverable through metadata. There are different types of research data repositories: there are specialized disciplinary repositories, more general purpose repositories, national repositories, institutional repositories and journal specific repositories.

➤ *Large database infrastructure to publish data. The researcher can manage, store and distribute datasets there.*

- **Digital Archive**

The organisation, policies, processes and procedures, financial management, staff, data management, data security and available hard- and software in its entirety; which make the curation and consultation possible of digital archival documents that need to be preserved.

Source: [Nationaal archief – Wat is een e-Depot?](#)

➤ *The infrastructure that is needed to ensure information can be curated and consulted for the long term (>10y).*

Repository vs. Archive

Digital archive

Purpose = preservation and curation of information that is process-related, for the long term:

- Guard authenticity
- Check integrity
- Document context
- Appraisal and selection
- Technology watch
- Make information available
- Dispose of information

E.g., [The National Digital Archives of Poland](#)

E.g., [E-Depot Nationaal archief \(NL\)](#)


Data repositories

Purpose= publication of information/data. This should be available immediately.

- Like articles that are published in journals

E.g., [DANS \(Data Archiving and Networked Services\)](#)

E.g., [SODHA \(Social Sciences and Digital Humanities Archive\)](#)

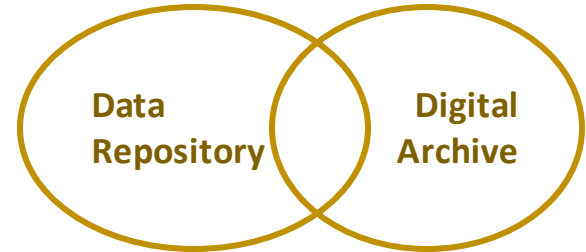


Even if you see the word “archive”, there’s a high chance they’re talking about repositories



Repository vs. Archive

- There is usually an overlap between the two:
 - Some repositories use more extensive [preservation activities](#)
 - However, data in repositories are usually not meant to last for the long term (> 10 years)
- How about repositories that call themselves “*archive*”?
 - Typically, such repositories curate data beyond what traditional data repositories do
 - Still, never assume the degree of data curation, and always double check what the repository offers, based on your needs!
- Examples of repositories that perform some level of data curation:
 - [Finnish Social Science Data Archive – FSD](#)
 - [Pangaea \(Data Publisher for Earth & Environmental Science\)](#)



TIP: Data Curation

“Data curation is the *active and on-going management of data* through its *lifecycle of interest and usefulness* to scholarship, science, and education; curation activities enable data **discovery** and **retrieval**, maintain **quality**, add **value**, and provide for **reuse** over time.”

Cragin, Melissa; Heidorn, P. Bryan; Palmer, Carole L.; Smith, Linda C. (2007). "[An Educational Program on Data Curation](#)". ALA Science & Technology Section Conference. Retrieved 7 October 2013.

Hence, data curation is concerned with data

- Generation
- Maintenance
- Management
- Appraisal and selection (see later)
- Quality ([preservation activities](#))
- Storage
- Sharing

“Digital or otherwise, data rarely stand alone. They are inextricable from research methods, theories, instruments, software, and context. Sustaining access to research data requires curation of individual objects and relationships among them.”

Borgman, C (2015). [Big data, little data, no data: Scholarship in the networked world](#). Cambridge, Massachusetts: MIT Press. pp. 272. ISBN 978-0-262-02856-1.

In other words, data curation requires data to be consciously and intentionally evaluated for current and (potential) future value, put and maintained in context with other relevant information that will help them be understood.

TIP: Data Curation on a Scale

Data curation is all about “**what** data are worthy of preserving, **why**, for whom, by **whom**, and for **how long**?”

Borgman, C (2015). [Big data, little data, no data: Scholarship in the networked world](#). Cambridge, Massachusetts: MIT Press. pp. 13. ISBN 978-0-262-02856-1.

- Think of data curation on a scale:



- Typically, archives perform a higher degree of data curation than repositories
- NOTE: Data curation is never at 100%, it can be endlessly perfected!

Data Repositories

What are they and Why we need them?

Data repositories: What?

- An online research data repository is a database infrastructure for the *management*, the *storage* and *dissemination* of research data
- Repositories are used to *publish* data and/or metadata after a research project is finalised. They provide prolonged access to data and/or metadata
- Repositories ensure that their database is *searchable*, and the data is *discoverable* through metadata
- There are different types of research data repositories:
 - E.g., discipline-specific repositories, general-purpose repositories, national repositories, institutional repositories, journal specific repositories, etc.

TIP: Do not deposit all data in a repository.

- There are classified data that, for various reasons, should not be available to the general public
- Examples:
 - Personal data regarding human research subjects
 - Data that may lead to the identification of participants
 - Data that may lead to identity theft or may make the participant vulnerable (e.g. religion, political opinion, sexual orientation, genetics, etc.)
 - Data for Dual-use: data that can be used for both civilian and military applications
 - Data that are related to the possible valorisation of research and/or economic activities based on the research

Data repositories: Why?

Depositing your data in a data repository:

- ✓ Facilitates data sharing & reuse (rise of citation rate)
- ✓ Makes verification of research results in publications easier
- ✓ Avoids unnecessary data collection/creation
- ✓ Makes possible to integrate several datasets
- ✓ Helps valorize research output
- ✓ Makes sure your data doesn't get lost
- ✓ Is increasingly required by publishers/funders



Source: [Foster – Loss of data cartoon](#)
License: [Attribution 4.0 International \(CC BY 4.0\)](#)



In summary: **Good scientific practice!**

What's in it for you as a researcher?

You build scientific **reputation**:

- Increased compliance with data management requirements brings you more *credibility* as a scientist
- Especially if your research is publicly funded (because you give back to the taxpayers by make your data “public good”)

You get more **citations**:

- Sharing your data openly can lead to a greater *visibility* and impact:
 - “papers with publicly available datasets receive a higher number of citations than similar studies without available data” ([Marwick, Boettiger & Mullen, 2018](#))
 - “a citation advantage, of up to 25.36% ($\pm 1.07\%$), with articles that have a category 3 Data Availability Statement—those including a link to a repository via a URL or other permanent identifier” ([Colavizza et al., 2020](#))

Helps with **funding**:

- Funders increasingly recognise datasets as outputs, which you can mention in funding applications ([Deutsche Forschungsgemeinschaft, 2022](#))

In short, you **contribute** to science on a whole other level AND **get rewarded** for it!

Types of Data Repositories

Discipline-specific repository

- Linked and relevant to a specific scientific discipline
- Advantages:
 - + Enhanced data visibility & reusability
 - + Customized features to handle data of that particular field of study (e.g., domain-specific metadata standards, specific documentation).
 - *It is advisable to inform yourself about these features in advance!*
 - + Designated for the scientific community to watch over the quality of data: Fellow members of the community possess similar knowledge on your field. This way they can provide a more thorough check.
- Recommended to choose discipline-specific repository instead of a general repository – whenever possible
 - E.g., [ADS](#) (Archeology Data Service), [Marine Data Archive](#) (oceanography geosciences)

General-purpose repository

- Accepts data from all disciplines
- Advantages:
 - + Often well-known solutions with large user communities
 - + Usually has strong institutional backing
 - + Indexed by major search engines (e.g. Google)
- Disadvantages:
 - Possible lack of centralized quality control
 - Increased risk of inadequate documentation
 - Possible loss of usable data



National repository

- Linked to a specific country
- Relies on government infrastructure
- Government initiative
- Sometimes set up by government agencies and regulators as part of a long-term strategy to protect and optimize the value of a nation's natural resources.

Source: Blinston, K. and Blohm, K., "Creating value from National Data Repositories", *CGG White Paper*, New York and Paris, 2017.

Examples:

- [ReShare](#) for the UK
- [Pôle National de Données de Biodiversité](#) (National Biodiversity Data centre, France) (both national & discipline-specific repository)

Institutional repository

- Financed and maintained by a particular institution, capturing the intellectual output of the host institution
- Content can be purely scholarly, but often contains other outputs, generated not only by researchers but also by other staff
 - This means the repository may include administrative, teaching and research materials.
- Sometimes open and interoperable, helping the institution disseminate its output
- Collection, storage, and disseminating of information are a part of the scholarly communication

Source: S. Haridasan and R. Khusboo, "Institutional repository initiatives in Indian Universities: An evaluative study", *Journal of Library and Information Science*, 37.2(2012), 71-87. Available online: https://www.researchgate.net/publication/303725097_institutional_repository_initiatives_in_Indian_Universities_An_evaluative_study
- Often multidisciplinary
- E.g., [Apollo](#) (University of Cambridge), [RDR](#) (University of Leuven)
 - Special case: [4TU.ResearchData](#) is an international data repository for science, engineering and design built and led by four technological universities in the Netherlands, with services available to anyone around the world

Project-specific repository

- A data repository with a focus on a specific research project
- E.g., [Aberdeen Birth Cohorts](#)

The University of Aberdeen followed all children born in Aberdeen in 1921, 1936, and 1950-1956 as they grow and age, gathering data concerning those children and depositing in this repository.

- E.g., [The Scientific Drilling Database](#)

Operated by GFZ German Research Centre for Geosciences, this repository provides drilling data that is created in the scope of the Scientific Continental Drilling Program (ICDP) and is openly accessible and reusable.

Journal-specific repository

- The majority of Journals will refer to a list of “recommended, trustworthy repositories”
- Some journals have established their own repository (= journal-specific repository)
- E.g., [GigaDB](#), linked to the journal [GigaScience](#)
- Mind your contract with the journal before you deposit your data!

FAIR Data Repositories

The role of “FAIR” in choosing a repository

What is “FAIR”?

Research data in repositories ideally comply to the **FAIR principles**:

1. **Findable:** Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets;
2. **Accessible:** Stored for long term such that they can be easily accessed and/or downloaded with well-defined license and access conditions (Open Access when possible), whether at the level of metadata, or at the level of the actual data content;
3. **Interoperable:** Ready to be combined with other datasets by humans as well as computer systems;
4. **Re-usable:** Ready to be used for future research and to be processed further using computational methods.

Source: Wilkinson, M., Dumontier, M., Aalbersberg, IJ., Appleton, G., Axton, M., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3: 160018

Findability (1)

“Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets”

How can the data in a repository be findable?

- **The repository assigns a globally unique and persistent identifier (PID) to the data. E.g., DOI**

“A persistent identifier is a long-lasting reference to a digital resource. An *identifier* is a label which gives a unique name to an entity: a person, place, or thing. Unlike URLs, which may break, a *persistent* identifier reliably points to a digital entity.”

Source: [Orcid – What are persistent identifiers \(PIDs\)?](#)

Findability (2)

e.g., DOI (Digital object identifiers): persistent identifiers for books, datasets, journal articles ...

For example: <https://doi.org/10.1109/5.771073>

An entity that provides DOI's is DataCite for example.

e.g., ORCID ID: provides a persistent digital identifier to a person/researcher. They can connect their ID with their professional information — affiliations, grants, publications, peer review, and more. For more information, visit [ORCID](#).



ORCID



Findability (3)

- **Why should a repository assign persistent identifiers?**

- Provides continued access to data, even if the location changes (Unlike URL's)
- Stable cross-linkage of digital resources
- Ensures your research outputs and activities are correctly attributed to you
- Reduces form-filling (enter data just once, link to other locations)
- Improves recognition and discoverability for you and your research outputs

Source: [Digital Preservation Coalition – Persistent identifiers](#)

- **To maximize findability, it is essential that your research (meta)data are indexed in different catalogues and services.**

- The data repository should, for this reason, be linked to these platforms. Cf. [Datacite](#), [OpenAIRE](#), [Google Dataset Search](#)...
- NOTE: You deposit your data only once to a repository of your choice, not to multiple repositories. Just make sure that the repository that you choose is findable.

Accessibility (1)

“Stored for long term such that they can be easily accessed and/or downloaded with well-defined license and access conditions (Open Access when possible), whether at the level of metadata, or at the level of the actual data content”

What level of accessibility should a data repository provide?

- The repository should be **open access by default**: data freely shared with anyone
- **Is there a possibility to opt out of Open Access?**
 - Access restrictions!
 - **Restricted access**: Data can be downloaded and reused if certain conditions are met. These data are confidential or classified (e.g., personal, dual use, intellectual property, third party, etc.)
 - **Closed access**: data that under no circumstances can be shared with other researchers
 - **Embargo**: Data can be freely shared after an embargo-period has passed. During that period data are inaccessible.

Accessibility (2)

Access conditions for data archived as restricted-access must be specified in a **data use agreement** (also important for *reusability*).

This might typically include:

- Evaluation of the reuse request by an ethical committee, if applicable
- Clear statement of the purpose of the study
- Non-disclosure agreement
- Clear acknowledgement of the data provider and the data source in publications resulting from the use of the data
- Guarantee that the data will be safely stored and eventually deleted
- Etc.

Contact your institution's legal department to learn about how data use agreements are made in your institution.

Interoperability (1)

“Ready to be combined with other datasets by humans as well as computer systems”

To be interoperable, a data repository must:

- **Make use of persistent identifiers**

- Ensures your data can be found through different information platforms (permanent and persistent “link”).

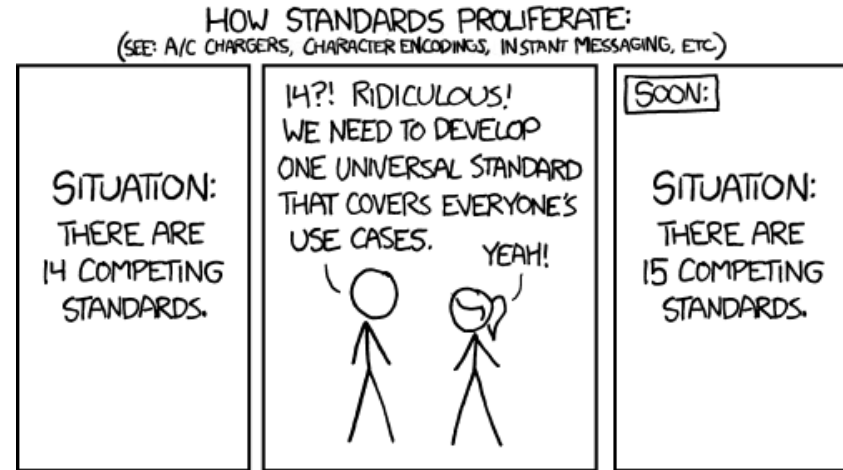
Makes your data also more [findable](#).

Interoperability (2)

- **Comply with metadata standards**

Thanks to these standards, metadata can be exchanged between information systems. This way your data can also be more [findable](#).

- The **structure** to which the metadata for your research data needs to adhere. These are **common components** of metadata (date, names, places ...) that give a description and provide the context to your data. There are a lot of metadata standards, even domain-specific ones. The more specific the standard, the more accurate the researcher can describe their data.



Source: [XKCD - Standards](#)

License: [Creative Commons Attribution-NonCommercial 2.5 License](#).



More information on *metadata* and *metadata standards*?

Check out the **EUTOPIA Metadata Training Guide**: [Understanding Metadata: A Beginners' Guide for Researchers & Data Stewards](#)

Interoperability (3)

Other ways for a repository to be interoperable (technical):

- It makes, **preferably**, use of **standardised formats**:
 - **Open and documented**: technical specifications are available
 - **Stable**: formats can only be altered after following a specific procedure
 - **Software-independent**: formats will be supported by software of multiple producers and open-source initiatives
 - **Independent of a producer**
 - e.g., “.txt” or “.pdf” instead of “.doc”; “.csv” instead of “.xls” or “.sav”, etc.
- It makes use of a clear exit strategy:
 - A repository should provide a clear way in which they will transfer the data that is stored in their repository in case they discontinue their operation, or a calamity occurs.

Reusability

“Ready to be used for future research and to be processed further using computational methods”

How repositories should make sure their data is reusable:

- **Terms of reuse** should be determined, at least:
 - Attribution
 - Copyright requirement
 - Control on commercial exploitation
- If severe restrictions on reuse: [data use agreement](#)
- ...

Reusability - Licenses

Ideally, data repositories allow researchers to **attach a reuse license** to the open data. This way, the conditions of reuse are instantly clear to potential re-users.

- E.g., Creative Commons

Note that licenses contribute to both [accessibility](#) and *reusability*.

Reusability - Licenses

- Always consider **interoperability issues between licenses!** CC0 and CC-BY licenses are the licenses that are most conducive to open access to and reuse of data
- Inform yourself about the **implications of restrictions** in the license you decide to use
 - Some of them might have undesirable side effects and/or ambiguities associated with them
- It's useful to follow the [recommendations of the European Commission](#):
 - **Creative Commons Attribution 4.0 International Public License (CC BY 4.0) for all content:** This means that the content (documents, data) can be re-used provided that the source is acknowledged; the re-user may not suggest that the Commission is endorsing the use made of this content.
 - **Commons Universal Public Domain Dedication deed (CC0 1.0) for raw data, metadata or other documents of comparable nature:** This means in practice that such content can be considered as in the public domain and can be used without any further requirement or obligation.

Reusability: Creative Commons (CC)

Step 1: Choose License Features

Publishing under a Creative Commons license is easy. First, choose the conditions that you want to apply to your work.



Attribution

All CC licenses require that others who use your work in any way must give you credit the way you request, but not in a way that suggests you endorse them or their use. If they want to use your work without giving you credit or for endorsement purposes, they must get your permission first.



ShareAlike

You let others copy, distribute, display, perform, and modify your work, as long as they distribute any modified work on the same terms. If they want to distribute modified works under other terms, they must get your permission first.



NoDerivs

You let others copy, distribute, display, and perform only original copies of your work. If they want to modify your work, they must get your permission first.



NonCommercial

You let others copy, distribute, display, perform, and (unless you have chosen NoDerivs) modify and use your work for any purpose other than commercially unless they get your permission first.

Step 2: Get a License

Based on your choices, you will get a license that clearly indicates how other people may use your creative work.



Attribution
CC BY



Attribution — ShareAlike
CC BY-SA



Attribution — NoDerivs
CC BY-ND



Attribution — NonCommercial
CC BY-NC



Attribution — NonCommercial — ShareAlike
CC BY-NC-SA



Attribution — NonCommercial — NoDerivs
CC BY-NC-ND

CC creative commons LICENSES	Copy & Publish	Attribution Required	Commercial Use	Modify & Adapt	Change License
Public Domain	✓	✗	✓	✓	✓
BY Attribution	✓	✓	✓	✓	✓
BY-SA Attribution ShareAlike	✓	✓	✓	✓	✗
BY-ND Attribution NoDerivs	✓	✓	✓	✗	✗
BY-NC Attribution NonCommercial	✓	✓	✗	✓	✓
BY-NC-SA Attribution NonComm ShareAlike	✓	✓	✗	✓	✗
BY-NC-ND Attribution NonComm NoDerivs	✓	✓	✗	✗	✓

Source: [Creative Commons Licenses - Open Textbooks and Resources for Faculty-Research Guides at George Washington University \(gwu.edu\)](#)



More on FAIR? Here are some resources.

- Learn more:

- [GO FAIR](#) initiative
- [FAIRsFAIR](#) project



- Test your knowledge via the [FAIR Aware](#) tool  FAIR Aware

Characteristics of Trustworthy Data Repositories

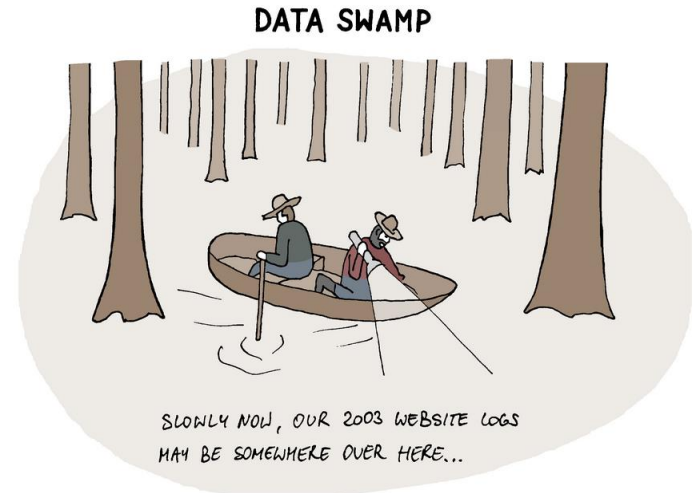
How to know which repository to trust?

A trustworthy data repository should be...

Not a data swamp!

Ideally, it should,

- Comply with **FAIR** data principles
- Comply with a **metadata** standard
- Be linked to data **creator** (e.g., via [ORCID ID](#))
- Be discoverable via (meta)data **catalogues**
- Include **access/reuse information** (e.g., license)
- **Curate data** to a certain extent (think of “*data archive*”)
- Ideally, also **certified**... (see next section)



 Dataedo /cartoon

Plot@Dataedo

Source: [Data Swamp – Dataedo Data Cartoon](#)

License: [Creative Commons Attribution-NoDerivs 3.0 License](#)



Repository Selection Criteria

Trustworthy repositories should meet the following minimum criteria:

1. Provision of Persistent and Unique Identifiers (PIDs)

- Allow data discovery and identification
- Enable searching, citing, and retrieval of data
- Provide support for data versioning

2. Metadata

- Enable finding of data
- Enable referencing to related relevant information, such as other data and publications
- Provide information that is publicly available and maintained, even for non-published, protected, retracted, or deleted data
- Use metadata standards that are broadly accepted (by the scientific community)
- Ensure that metadata are machine-retrievable

3. Data access and usage licenses

- Enable access to data under well-specified conditions
- Ensure data authenticity and integrity
- Enable retrieval of data
- Provide information about licensing and permissions (in ideally machine-readable form)
- Ensure confidentiality and respect rights of data subjects and creators

4. Preservation

- Ensure persistence of metadata and data
- Be transparent about mission, scope, preservation policies, and plans (including governance, financial sustainability, retention period, and continuity plan)

Source: [Science Europe – Practical Guide to the International Alignment of Research Data Management](#)

Looking for a repository to deposit your data?

- Check with your institution what the best options are
 - When in doubt, go *institutional* (if your institution has a repository) or *discipline-specific* – but definitely, go reputable
- Make sure you can
 - Get a persistent and unique identifier with that repository
 - Choose the degree of openness of the data deposited
 - Set a reuse license
 - Etc.
- Do not share all data (e.g., personal data? Consider only your institutional repository, if available.)
- Ideally, go *certified!* (see the next section)

Certified Data Repositories

How important is certification?

Reputation = Certification ?



Reputation of a repository can be assessed in three levels:

1. **Listed** in the *re3data* or *BioSharing* registries, or broadly recognised in the research domain
2. **Endorsed** by a relevant funder, journal, or learned/ professional society
3. **Certified** to an appropriate international standard

Hence, **certification** is one of the highest reputations a repository can get.

NOTE! Only a small number of repositories are certified and there are many good data repositories that have not been certified (yet) for various reasons.

- E.g., Zenodo has not been certified by CoreTrustSeal because it's not domain-specific

Remember: While certification is an important criterion for trustworthiness, it is not be the only one.

Who Certifies Repositories?

According to the EU framework, certification happens via 3 standards in 3 levels:

1. Level 1: Basic (e.g., [CoreTrustSeal](#))
 - Self-assessment by the repository, supported by appropriate public evidence submitted for peer review
 - No on-site visit required
2. Level 2: Extended (e.g., [nestor Seal](#))
 - Self-assessment template to be completed and submitted to nestor, based in Germany
 - Handful of repositories certified this way – not very active
3. Level 3: Formal (e.g., [ISO 16363](#))
 - Most advanced certification
 - On-site visit required



Most widely known: CoreTrustSeal



- **CoreTrustSeal** offers to any interested data repository a core-level certification based on the *DSA–WDS Core Trustworthy Data Repositories Requirements catalogue and procedures*
- Universal catalogue of requirements reflecting the core characteristics of trustworthy data repositories

- DSA: Data Seal of Approval
- WDS: World Data Systems



joined forces to create
CoreTrustSeal

At the time of writing, *CoreTrustSeal* certifies only discipline-specific repositories, NOT general-purpose repositories such as Zenodo.

Certification: must-have or nice-to-have?

In principle, prefer certified repositories to non-certified repositories,

HOWEVER, remember:

- Only a certain number of repositories is certified!
- The “good practices” in your discipline/field matters a lot:
 - Is a non-certified repository recommended more than a certified one? Then exercise your discretion.

Finding Trustworthy Repositories?

Where to look?

How to find a trustworthy repository?: Search engines


- Narrow your search with a number of parameters
- Most widely recognized inventory of research data repositories:
















- Others often based on Re3data:
 - E.g., [Repository Finder](#), [Data Deposit Recommendation Service](#), etc.

Example: *re3data*

[re3data](#) has a nice and easy way of informing you about what you can expect from a repository:

- Next to the name of each repository, you have six little icons either 'lit up' or darkened depending on whether a feature is available in that repository
 - Here are the icons 
 - Shortcut to finding out some basic information

REMEMBER → Always check the functions of the repository, *before* you deposit your data

	The research data repository provides additional information on its service.
	The research data repository provides open access to its data.
	The research data repository provides restricted access to its data.
	The research data repository provides closed access to its data.
	The terms of use and licenses of the data are provided by the research data repository.
	The research data repository provides a policy.
	The research data repository uses DOI to make its provided data persistent, unique and citable.
	The research data repository uses URN to make its provided data persistent, unique and citable.
	The research data repository uses ARK to make its provided data persistent, unique and citable.
	The research data repository uses handle to make its provided data persistent, unique and citable.
	The research data repository uses Purl to make its provided data persistent, unique and citable.
	The research data repository uses a persistent identifier system to make its provided data persistent, unique and citable.
	The research data repository is either certified or supports a repository standard.

Source: [FAQ | re3data.org](https://faq.re3data.org)

Example: *re3data* (continued)

ICSSR Data Service: Social Science Data Repository

ICSSR Data Service: Indian Social Science Data Repository

Subject(s) Humanities and Social Sciences Social Sciences Statistics and Econometrics Social and Behavioural Sciences Economics Empirical Social Research

Content type(s) Scientific and statistical data formats Raw data other

Country India

The "ICSSR Data Service" is culmination of signing of Memorandum of Understanding (MoU) between Indian Council of Social Science Research (ICSSR) and Ministry of Statistics and Programme Implementation (MoSPI). The MoU provides for setting-up of "ICSSR Data Service: Social Science Data Repository" and host NSS and ASI datasets generated by MoSPI. Under the initiative, social science research institutes, NGOs, individuals and others dealing with social science research are also being approached to deposit / provide their research datasets for hosting into the repository of ICSSR Data Service. The ICSSR Data Service includes social science and statistical datasets of various national-level surveys on debt & investment, domestic tourism, enterprise survey, employment and unemployment, housing condition, household consumer expenditure, health care, etc., into its repository. ICSSR Data Service aims to facilitate data sharing, preservation, accessibility and reuse of social science research data collected from entire social science community in India & abroad. The Information and Library Network (INFLIBNET) Centre, Gandhinagar has been assigned the task of setting-up the data repository.



- Open access
- No PID
- No certification

Slovenian Social Science Data Archives

ADP

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

Content type(s) Structured text Scientific and statistical data formats

Country Slovenia


The Slovenian Social Science Data Archives (Slovenski Arhiv Družboslovnih podatkov - ADP) were established in 1997 as an organizational unit within the Institute of Social Sciences at the Faculty of Social Sciences, University of Ljubljana. Its tasks are to acquire significant data sources within a wide range of social science disciplines of interest to Slovenian social scientists, review and prepare them for digital preservation, and to disseminate them for further scientific, educational and other purposes.



- Restricted access
- PID provided
- Certified

List of recommended repositories

- [Recommended by the European Commission](#)
- [Recommended repositories by Elsevier](#)
- [Recommended repositories by Nature](#)
- [Recommended repositories by PLOS](#)
- [Recommended repositories by ELIXIR](#)



*Not necessarily
certified
repositories!*

Data Appraisal, Data Selection and Data Depositing

What to keep and How?

Depositing Data: Starts with DMP

Ideally, most (if not all) decisions around data depositing and publishing (which data, in which repository, and under which conditions) should be

- Made in the beginning of the research project (not always possible)
- Shared with the project collaborators (if any)
- Mentioned in your initial (if possible) and final DMP

... so that you can continuously appraise your data accordingly.

Appraisal and selection of research data

- **Your GOAL:** Document and preserve everything that is needed to *verify* and/or *replicate* your study, as well as *reuse* your data by another researcher.
- Is it necessary to keep ALL your research data?
 - No, keeping large amounts of data will be **costly**
 - Hence, choose wisely
- How to appraise your data?
 - Take into account the evaluation criteria in the following pages

Evaluation criteria (Dorey et al. 2022)

- Preservation intent:

Does the owner of the dataset want to keep the materials? Is there an obligation to stakeholders? Researchers have different needs, which aspects are important to keep? This can already be indicated in the DMP.

- Value:

Does data show enough proof of research activities? Does it have possible social, scientific or historic value? How valuable is data to communicate current knowledge/ perspectives. Who will use the data, why and why in this format?

- Uniqueness:

Is data unique, can it be reproduced easily? Sometimes better to keep data, even if it is reproduced easily: e.g., when it has an impact on the environment, like if you work with very polluting chemicals.

Source: Dorey, J., et al, *Appraisal guidance for the preservation of research data*, 2022. <https://doi.org/10.5281/zenodo.5942236>.

License: [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)



Evaluation criteria (Dorey et al. 2022) – continued

- Rights, Restrictions, and Potential for redistribution:

Can data be made accessible or are there restrictions? Does the data contain sensitive and/or personal information? Is there authorization to redistribute information? If not, is this still the case for long term preservation?

- Preservability of content and context (full documentation):

Is there enough documentation to retrieve the context of the data creation? Is the data kept in a format that allows preservation and access?

Source: Dorey, J., et al, *Appraisal guidance for the preservation of research data*, 2022. <https://doi.org/10.5281/zenodo.5942236>.

License: [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/)



Evaluation Criteria (DCC 2014)

The following criteria from DCC (2014) also helps to put data appraisal into perspective:

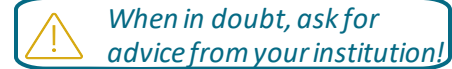
1. Check for indications that it *must* be kept considering **legal or policy compliance risks**
2. Consider potential **reuse purposes** - what aims *could* the data meet?
3. Identify which data *should* be kept as it may have **long-term value**
4. **Weigh up the costs** - which data management costs have already been incurred and therefore contribute to its value, and how much more is planned and affordable? Where will the funds to pay these costs come from? Considering these questions will give you the cost element of your data appraisal and should help identify any need for external advice, e.g., on how to deal with any shortfall in the budget.
5. **Complete your data appraisal** - this will list what data *must*, *should* or *could* be kept to fulfil potential reuse purposes. The appraisal should also summarise any actions needed to prepare the data for deposit, or the justification for not keeping it.

DCC (2014). 'Five steps to decide what data to keep: a checklist for appraising research data v.1'. Edinburgh: Digital Curation Centre. Available online: <https://www.dcc.ac.uk/guidance/how-guides/>



Depositing Data: Step-by-step

- STEP 1: Appraise your data (data evaluation and selection)
 - If the data includes personal data, the GDPR applies, meaning that the data cannot be deposited in an open access repository
 - If personal data are completely anonymized, the GDPR doesn't apply, and data can be deposited in open access
 - *If the data is classified in any way, DO NOT deposit them in a repository – even if you restrict the access*
- STEP 2: Attach a license (incl. Terms of Use and/or data use agreement)
- STEP 3: [Selection a repository](#)
- STEP 4: Upload your data



TIP: Check if your deposited data are FAIR

F-UJI is a fun and interesting tool to test the *FAIRness* of a dataset

- Developed within the context of the FAIRsFAIR project
- Based on 16 out of 17 core FAIR object assessment metrics developed within FAIRsFAIR
- Used to assess how FAIR a dataset uploaded in a repository is
- **Still under development** (*at the time of writing in September 2022*)
- Results not formal and definitive but informal and informative



Example: Assessment in F-UJI

- [Dataset](#) in Zenodo (version 99 uploaded on August 25, 2022)
- DOI entered in [F-UJI](#) & assessed
- F-UJI gives a score to the dataset in three levels:
 - Overall
 - Each aspect of FAIR
 - Criteria for each aspect of FAIR

(see next page for the results)

August 25, 2022

Dataset Open Access

BIP4COVID19: Impact metrics and indicators for coronavirus related publications

Thanasis Vergoulis, Ilias Kanellos, Serafeim Chatzopoulos, Danae Pla Karidi, Theodore Dalamagas

This dataset contains impact metrics and indicators for a set of publications that are related to the COVID-19 infectious disease and the coronavirus that causes it. It is based on:

1. The COVID-19 dataset released by the team of Semantic Scholar¹ and
2. The curated data provided by the LitCovid hub².

These data have been cleaned and integrated with data from COVID-19-TweetIDs and from other sources (e.g., PMC). The result was dataset of 592,857 unique articles along with relevant metadata (e.g., the underlying citation network). We utilized this dataset to produce, for each article, the values of the following impact measures:

- **Influence:** Citation-based measure reflecting the total impact of an article. This is based on the PageRank³ network analysis method. In the context of citation networks, it estimates the importance of each article based on its centrality in the whole network. This measure was calculated using the PaperRanking (<https://github.com/divis/PaperRanking>) library⁴.
- **Influence_{alt}:** Citation-based measure reflecting the total impact of an article. This is the Citation Count of each article, calculated based on the citation network between the articles contained in the BIP4COVID19 dataset.
- **Popularity:** Citation-based measure reflecting the current impact of an article. This is based on the AttrRank⁵ citation network analysis method. Methods like PageRank are biased against recently published articles (new articles need time to receive their first citations). AttrRank alleviates this problem incorporating an attention-based mechanism, akin to a time-restricted version of preferential attachment, to explicitly capture a researcher's preference to read papers which received a lot of attention recently. This is why it is more suitable to capture the current "hype" of an article.
- **Popularity alternative:** An alternative citation-based measure reflecting the current impact of an article (this was the basic popularity measure provided by BIP4COVID19 until version 26). This is based on the RAM⁶ citation network analysis method. Methods like PageRank are biased against recently published articles (new articles need time to receive their first citations). RAM alleviates this problem using an approach known as "time-awareness". This is why it is more suitable to capture the current "hype" of an article. This measure was calculated using the PaperRanking (<https://github.com/divis/PaperRanking>) library⁴.
- **Social Media Attention:** The number of tweets related to this article. Relevant data were collected from the COVID-19-TweetIDs dataset. In this version, tweets between 23/6/22-29/6/22 have been considered from the previous dataset.

We provide five CSV files, all containing the same information, however each having its entries ordered by a different impact measure. All CSV files are tab separated and have the same columns (PubMed_id, PMC_id, DOI, influence_score, popularity_alt_score, popularity_score, influence_alt score, tweets count).

The work is based on the following publications:

255,702 36,735
 views downloads
[See more details...](#)

Indexed in



Publication date:
August 25, 2022

DOI:
DOI: [10.5281/zenodo.7022581](https://doi.org/10.5281/zenodo.7022581)

Keyword(s):
 COVID-19 coronavirus scientometrics bibliometrics

Related identifiers:
 Cites
<https://pages.semanticscholar.org/coronavirus-research> (Dataset)
<https://github.com/divis/PaperRanking> (Software)

Supplement to
www.biorxiv.org/content/10.1101/2020.04.11.037093v2 (Preprint)

Communities:
 Coronavirus Disease Research Community - COVID-19
 Zenodo

License (for files):
[Creative Commons Attribution 4.0 International](#)

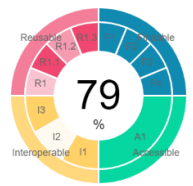


Example F-UJI: Assessment Results

Evaluated Resource:

BIP4COVID19: Impact metrics and indicators for coronavirus related publications		Save Download (JSON) New
FAIR level: ⓘ	advanced	
Resource PID/URL:	10.5281/zenodo.7022581	
DataCite support:	enabled	
Metric Version:	metrics_v0.5	
Metric Specification:	https://doi.org/10.5281/zenodo.4081213	
Software version:	2.0.0	
Download assessment results:	(JSON)	
Save and share assessment results:		

Summary:

		Score earned:	Fair level:
	Findable:	7 of 7	advanced
	Accessible:	3 of 3	advanced
	Interoperable:	3 of 4	moderate
	Reusable:	6 of 10	moderate

Findable

FsF-F1-01D - Data is assigned a globally unique identifier.	✔
FsF-F1-02D - Data is assigned a persistent identifier.	✔
FsF-F2-01M - Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability.	✔
FsF-F3-01M - Metadata includes the identifier of the data it describes.	✔
FsF-F4-01M - Metadata is offered in such a way that it can be retrieved programmatically.	✔

Accessible

FsF-A1-01M - Metadata contains access level and access conditions of the data.	✔
FsF-A1-03D - Data is accessible through a standardized communication protocol.	✔
FsF-A1-02M - Metadata is accessible through a standardized communication protocol.	✔

Interoperable

FsF-I1-01M - Metadata is represented using a formal knowledge representation language.	✔
FsF-I2-01M - Metadata uses semantic resources	⚠
FsF-I3-01M - Metadata includes links between the data and its related entities.	✔

Reusable

FsF-R1-01MD - Metadata specifies the content of the data.	✔
FsF-R1-1-01M - Metadata includes license information under which data can be reused.	✔
FsF-R1-2-01M - Metadata includes provenance information about data creation or generation.	✔
FsF-R1-3-01M - Metadata follows a standard recommended by the target research community of the data.	✔
FsF-R1-3-02D - Data is available in a file format recommended by the target research community.	✔

Final Remarks

Conclusions and Recommendations

When choosing a data repository:

Choose a repository that:

- Gives your submitted dataset a **persistent and unique identifier**
- Provides a **landing page for each dataset**, with metadata that helps others find it, tell what it is, and cite it
- Helps you to **track** how the data has been used
- Responds to **community needs and/or is certified** as a ‘trusted data repository’
- Offers clear terms and conditions that meet **legal requirements** – e.g., for data protection and allow reuse without unnecessary licensing conditions

Do yourself a favour: Plan Ahead.

- The decision to publish or share data should ideally be made in the beginning of the research project, including depositing data in a repository
 - You can still re-evaluate your decisions later on
- Begin with the end in mind: Write your intentions in your Data Management Plan
- Any idea how you will appraise your data (what to keep)? Write that in your DMP too

Have fun contributing to (open) science!

Contributors

Özgün Ünver (Research & Data Management, Vrije Universiteit Brussel)

Julie Jordens (Centre for Academic and Secular Humanist Archives, Vrije Universiteit Brussel)

This guide was prepared with the valuable contributions from (in alphabetical order) Pieter De Bruyn, Elisa Maes, Jone Paesmans, and Lennart Stoy from Vrije Universiteit Brussel.

For any queries, contact dmp@vub.be.

License

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



Available on Zenodo at <https://doi.org/10.5281/zenodo.7258306>