Repositories for research data

Definition

Repositories for research data are used to publish or archive digital data from our research and to make it accessible to interested parties. The motivation for this is that research results from publications[1] become traceable and reproducible with the help of the published research data. It is also possible to publish a research dataset without a corresponding publication so that as many other researchers as possible can use it for their research questions.

The publication of the research data of a project is already required or recommended by many third-party funders and is part of good scientific practice[2]. When choosing a data repository, attention should be paid to various quality features that can contribute to the understanding and visibility of the data set.

Quality features for repositories

Persistent identifiers (PID) for data records : To be able to reference a data record as clear as possible, it needs an identifier that refers to the corresponding resource on the Internet. For this purpose, the repository (rarely also from its institution or publishers) usually assigns a DOI (Digital Object Identifier) for the data record. Other PID systems, which are also widely used in a specific field, are URN (Uniform Resource Name) or hdl (Handle System).

Persistent identifiers for authors: When specifying the authors, it is also important to be able to provide them with a unique identifier. This remains even if the researchers change their names or change their place of work and is therefore important to permanently link the data set with the relevant people. Common identifiers for scientific authors include *ORCID* and *ResearcherID*.

Metadata: Metadata is used to describe the data record in as much detail as possible. This includes information such as authors, survey and publication date, organization/institute, file format, keywords or departments. Depending on the specialization, more specific information may also be common, such as geolocation.

Download and export options: The repository should offer the possibility of being able to download the data record or even parts of it. Furthermore, there should also be various export options for citing, e.g. bib for latex or JSON-LD for websites to easily reference the data set in different systems. Displaying the citation of the data record in different styles as plain text can also be helpful.

Description or documentation: It should be possible to describe the creation and context of the data set in a text field. However, this can also be implemented with a so-called readme file in the data record itself.

Access options: Various access options give you the option of restricting access to the data record. The data record can either be freely accessible, restricted only for a certain user group, on request or only freely available after a specific embargo period.

Repositories for research data

Licenses: The license of a data record defines how it can be used. By selecting a suitable license, restrictions regarding attribution, monetization, modification or republication can be set. Standard licenses are offered by CC (Creative Commons), ODC (Open Data Commons) and GPL (GNU General Public License), for example.

Overview / preview of the data record: Since many users immediately want information or an overview in the form of visualizations for a data record, it makes sense if the repository already offers folder structures, image or PDF viewers without having to download the data record first.

Versioning: It can always happen that a data record must be changed or updated. In this case, it makes sense to be able to create the data set as a new version instead of having to enter it as a new data set.

Registration and processing: Registration and logging in should be made as easy as possible, for example via the ORCID of the author. After publication, the author should still be able to edit the data record, in particular the metadata, at any time without having to contact the support desk.

Discovery by search engines: The records of a repository are not always made public by indexing for Internet search engines. By looking for a sample data record by title, description or keywords in common search engines, you can find out whether it is indexed there and can, therefore, be found.

The listed quality features are not always apparent at first glance. However, it often helps to look at a few sample data sets in the repository to get a first impression.

The **FAIR principles** provide principles for making research data sustainable and reusable. Research data repositories help to fulfill some of these principles. For example, uploaded data records are automatically provided with a persistent identifier and important metadata must be filled in before publication. Nevertheless, the documentation and compliance with subject-specific standards must be taken over by the researcher himself.

Offers and services

In the following, we would like to present some recommended offers and services. Re3data can serve as the first point of contact for the search for a suitable subject-related repository.

Re3data.org lists over 2000 repositories from various fields. These are presented there together with information on compliance with important quality characteristics. Several German institutes and sponsors are already supporting this platform so that researchers can use it as a support for finding suitable repositories.

Repositories for research data

Interdisciplinary repositories for research data

If it was not possible to find a suitable subject-related repository via the re3data directory, then general research data repositories are available as an alternative.

All the following repositories basically offer their service free of charge, i.e. Data records can be uploaded free of charge and made public. However, some of the offers have restrictions on the number of files or the size of a data record.

Zenodo is a cross-disciplinary repository for all types of scientific data sets, such as also for publications, source code or presentations. Among other things, it has an integration of GitHub and social media statistics. The platform is funded by the European Commission, originated from the *OpenAIRE* project and is hosted at *CERN*.
Link: zenodo.org

With *B2Share*, the *EUDAT* project offers a platform to make data accessible and interdisciplinary. With *B2NOTE*, *B2SAFE*, *B2HANDLE*, *B2DROP*, *B2FIND*, *B2STAGE*, *B2ACCESS*, and *easyDMP* further services are offered to simplify the handling of research data. The platform originated from the Horizon 2020 funding program of the European Union.
Link: b2share.eudat.eu

The *OSF Home* repository of the *Open Science Framework* is open source and is intended to help scientists to make their research processes as open and transparent as possible. With *OSF preprints*, *meetings*, *institutions*, *registries* and *collections*, other services are offered to support this approach. The offering was developed by the *Center for Open Science* and is hosted in the United States.
Link: osf.io

Dryad is an interdisciplinary solution for the publication of all types of research data. The open source platform is supported by various American institutes and sponsors. Dryad has been hosted by the *California Digital Library* since 2019. New services are to be developed in cooperation with *Zenodo*.
Link: datadryad.org

Figshare is a repository for research data, which is especially designed for visual graphic data sets such as posters, diagrams or videos. The platform is offered by the *Nature Publishing Group*, which is part of the *Springer Nature* publishing group and is based in Great Britain.
Link: figshare.com

The *Thuringia Digital Library* is a local solution for the publication of research data. In addition to the open access publication of publications, it also offers researchers from Thuringian universities the publication of data records.
Link: db-thueringen.de