Data Repository: FAQ

Find the answers to your questions regarding this repository service, data publishing and preservation here.

What is the Data Repository service?

The Data Repository service is a data publication platform for research data sets. Any data that has been used during research projects and is referred to in published articles can be stored on the platform in order to long-term preserve it and make it available to others.

Who can use the Data Repository?

The Data Repository is open to all researchers and scientists who are affiliated to research institutions, universities as well as to individual researchers. To deposit research data registration is required, as well as for access to and downloading files of specific closed data sets.

Why would I publish my data set?

By storing and publishing your data set in a repository you can guarantee the data will be available and preserved for a long time and make it findable and reusable for other researchers. For example, as soon as you have published an article, it is worthwhile to make the corresponding research data that was used and produced during the research project available to other researchers or anyone else. These people are then able to replicate the steps and reproduce the conclusions made in the published research article.

Depending on the conditions in your funding contract, publishing in a trusted repository might successfully conclude your research project as you are complying to the requirements stated by the financier and/or publisher concerning the handling of your research data.

Can I publish any data I have or find elsewhere online?

No, you cannot (re)publish any data that does not belong to you, since you need to have the rights and ownership of the data and in case the data is not yours, explicit permission of the original owner is required. Also, the service is meant for publication of *research* data, any other data will be rejected.

So what does this all cost?

Larger data sets need data publication contracts that can be requested and negotiated by contacting us. There is no limit on the number of deposits and/or collections you create.

What is needed for publication?

In order to publish a data set you need one or more files that can be published and adhere to the basic metadata requirements, like providing the title and description of your publication. It is important that you are allowed to publish the data and, if applicable, own the intellectual property corresponding to the data and metadata you are publishing.

Furthermore, for large publications the following information is needed:

- How many data sets are being published?
- Which file formats are used and what kind of data is contained?
- How many files are included in the data sets and what is their file size distribution?
- How many downloads are expected in the short and long term?
- Which publications (articles) are related to the data set publications?
- For which duration the data set needs to be available?

In order to get the data to our premises, it is useful to know where your data is currently residing.

What is metadata and why should I add it to my publication?

Metadata is the accompanying information that describes and enriches the data set and its files. It makes sure the data set can be understood, discovered, correctly cited and authorised for use to other parties. Metadata is added during the creation of the publication and should not be altered after the publication is finalised.

In general, several types of metadata are distinguished: technical, descriptive and additional metadata. Technical metadata is automatically gathered and assigned by the repository, like file and system information and persistent identifiers. Descriptive and possibly additional metadata need to be provided by the data producer and/or owner of the data set. Basic (mandatory) metadata includes among others the title, author, description, date and license of the data set. If there is more descriptive metadata available which is useful to add to the publication, this can be added during the ingest of the data.

Many metadata are assigned to the publication using a schema, usually called the metadata schema. Basic metadata is defined in the default metadata schema, but communities can also create their own schemas that can be used for publication of data sets.

What is a persistent identifier?

A persistent identifier (PID) is a unique Digital Object identifier and consists of a prefix number followed by a unique string of characters, for instance a UUID. Prefixes are assigned by internationally recognised DONA (Digital Object Numbering Authority) agencies and hosted at a PID service provider. The Data Repository creates PIDs that are based on the Handle system. The repository's principle PID provider for allocation and resolution of PIDs is hosted at SURF.

You can read more about PIDs on the website of Persistent Identifiers for eResearch.

How can I organise my data?

There are several ways to organise your data. Generally, files that belong to a single data set can be published in a single deposit. If your data set is very large and can be divided naturally or easily into different deposits, this is certainly possible. These deposits can thereafter be grouped into a collection with its own metadata. From this collection another user can easily find each deposit in that collection.

If you have many collections and deposits to make, it can be tedious to do this through the website. Please contact the advisors of SURF in order to obtain assistance in ordering your data set deposits.

My data set is very *very* large, what can I do to publish it?

If your data set exceeds several terabytes, it can be difficult or even impossible to upload the data through a browser to the Data Repository. This largely depends on the size of your data set, the speed of your network connection and other technical limitations, such as browser timeouts. Also, the repository currently does not allow file sizes larger than 4 GB and a total size of 10 GB and 10 files for a new deposit when using the online deposit workflow through your browser.

To avoid these problems, please contact the advisors of SURF to set up efficient data transfers to a separate storage location. Once transfers have completed, the advisors will help creating the deposits and collections.

How can I start publishing my data set?

First contact us to explain and discuss your publication needs. Once you have approved your request and you have been given access, you can start depositing right away by starting with the first step of the online deposit workflow and start uploading your data.

If you need to publish large data sets or have special requirements for your publication, also contact us.

Other questions

If you have any other questions not mentioned here, please contact us. We will get back to you as soon as possible.