THE GLOBAL HEALTH NETWORK
Enabling research by sharing knowledge

EDCTP

Sweden
Sverige

Data Sharing Toolkit

Data sharing, particularly of potentially sensitive information, is not always straightforward. This Data Sharing Toolkit has been developed to collate practical information and resources related to the topic. The Data Management Basics section includes explanations covering different aspects of working with data, from a list of free version control tools, through explanation of metadata, to real-life examples of data problems encountered by data managers handling clinical data. The Data Sharing Steps provide an overview of the core elements of the process of depositing your data into a repository. The Repository Finder has been developed in order to aid users in choosing a suitable repository to submit to. The toolkit also included an extensive collection of resources linked to working with and sharing data.

The Global Health Network and European & Developing Countries Clinical Trials Partnership (EDCTP) Editorial Team

https://doi.org/10.48060/tghn.10

EDCTP Knowledge Hub: Version 1.0 - 17th September 2021 (Knowledge Hub)

Toolkit

How do I share my data?

OVERVIEW OF THE MAIN STEPS

1. CHOOSE A SUITABLE REPOSITORY & SET UP AN ACCOUNT

- Your funder may require you to submit the data to a specific repository. Some funders have strict requirements, while others provide a list of recommended repositories.
- Journals may also require deposits to a specific repository and/or may recommend repositories.
- There may be discipline-specific and disease-specific repositories that are preferable. You can also look up repositories in your discipline using re3data and FAIRsharing.
- If your funder/journal do not provide any guidance, or if you are not familiar with repositories used in your field, we provide guidance on how to choose a suitable repository in the Repository Checklist.
- Make sure to familiarise yourself with the repository guidelines.

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2. ORGANISE YOUR DATA

- Decide on the best way to organise your data - sometimes it is best to merge several files into one dataset, but in other cases depositing separate files makes more sense.
- Structure and name your files well - for your own use and to assist others.

2

Data organisation

File organisation

File naming

3. PREPARE YOUR DATA

- Are your data clean and labelled consistently? Be explicit in your naming to ensure that others can understand your data.
- Are you using non-proprietary formats to ensure accessibility now and in the future? If you need to use discipline-specific format you may consider submitting two versions of the file – one in the discipline-specific format and one in non-proprietary format.
- If data are in a discipline-specific format you should double check that the repository will accept that format.

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4. PREPARE DOCUMENTATION FILES

- All variable labels, codes and acronyms should be either self-explanatory or explained (this can be done in a separate 'README.txt' file). You may want to include a user guide.
- Have you included your methodology, protocols and any other relevant information? Is it clear how the data were collected and processed?

4

Example: README file

Checklist of potential files to include

5. COPYRIGHT, CONSENT, PERMISSIONS

- You need to have the right to share the data: are you sure that you have obtained permission from all right-holders? This includes patient consent – make sure sensitive data are sufficiently anonymised.
- What licence are you going to share your data under?

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6. DEPOSITING THE DATA

- Once all your files are ready and you have chosen an appropriate repository it is time to deposit your data. Each repository has its own process for data deposition, so you will need to follow their guidelines.
- If your dataset was assigned an identifier (DOI) keep it for future reference – you may need to provide your funder and/or institution with it.

6

Example: submission process

How do I share my data?

Back: overview of the main steps

DIFFERENT MODELS OF ACCESS

The model of access will determine who can browse the repository, who can submit data to the repository, and who can download the data from the repository.

There are three main models of access:

- open access
- controlled access
- closed access

Open access model: there are no access barriers

repository access - anyone can browse the repository
data deposition - anyone can deposit their data in the repository
data access - anyone can access the data stored by the repository

Controlled access model: external users have to overcome certain barriers before they access

repository access - barriers have to be overcome before browsing
data deposition - barriers have to be overcome before depositing data
data access - barriers have to be overcome in order to download data

There is a range of barriers that can be implemented, for example users may have to register on the website, pay a fee, be members of a specific institution, or obtain permission from the data owner.

Access to data may also be embargoed, in which case external users cannot access the data unit the data are released for open access. Embargo could last for months or years.

Closed access model: external users cannot overcome the barriers

repository access - external users cannot browse the repository
data deposition - external users cannot deposit data
data access - external users cannot download data

Please note that access to the repository and to the data are not the same.

It is possible to have different models of access for the repository and for the data held by the repository. For instance, the repository may be open access, but the data in that repository could be under controlled access. In that case, anyone could browse the repository and view certain level of metadata, but in order to download and use the data additional steps would have to be taken to obtain the necessary permissions and access.

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

OPTIONS OFFERED BY REPOSITORIES

In addition to Different models of access (open/controlled/closed) there are several other attributes that need to be considered when choosing a suitable repository suitable for data deposition. These are outlined briefly below. We will address as many of them as we can through our Repository Tool to guide you through choosing an appropriate repository.

Scope - while many repositories accept data from all fields of research, some may be more discipline-specific.

Eligible depositors - as explained in the Different models of access some repositories may restrict who can submit their data to the repository.

Ownership - Look out for repositories that require you to transfer ownership of your data to them. Ideally, you want to retain ownership and just give the repository the right to store and maintain your data.

Data file formats - many repositories will accept all data formats, even proprietary ones. However, certain repositories may only support and accept specific formats, so do check the guidelines before submission. Remember that use of non-proprietary formats is also strongly encouraged when data sharing.

Volume and size limitations - depositories may have restrictions on how much data can be deposited as a single record. The volume of data generated by many studies will fit safely within usual size limits. However, if you have an unusually large amount of data (tens or hundreds of GBs), you may need to look to specialist repositories or contact the repository with a query - some repositories will consider larger submissions on a case-by-case basis.

Data quality - a few repositories out there that may be able to assist you with data curation, however, most repositories accept data “as is”, so you should process it beforehand to ensure that it can be reused easily.

Metadata - repositories vary in the types and sources of metadata they support. Some adhere to specific standards or follow certain guidelines (such as the OpenAIRE Guidelines).

Language - repositories may require the text to be in a specific language, may have a preference for a language (but accept more than their preferred choice), or accept all languages. If you need to deposit your data in a specific language, check that your chosen repository will accept it.

Use licences - a variety of licences may be supported by the repository. Repository may request that the depositor specify the licence themselves. Licences for data and metadata may not be the same.

Retention period - to ensure the long-term archiving of your data you should choose a repository that clearly states the length of their secured funding or at least has a sound business plan outlining how the repository will be kept alive. What will happen to your data if the repository has to be closed down?

Preservation - it is important to know whether the repository has a back up system in place. If all your data are on one specific server, they could be easily destroyed if that server fails. It is also important that the repository checks the files and back ups regularly (e.g. using checksums) to ensures file authenticity.

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

RECOMMENDED REPOSITORIES

Identifying a suitable research data repository is an important step in planning how to manage and share data. Many funding bodies and journals recommend or specify the research data repositories in which grantees and researchers should deposit data.

Repository review

To help you work through all of the options we have reviewed which repositories are recommended/specified by different funders and journals. A broad range of funding bodies and journals were identified to be included. Our aim was to identify any conflicts (what would happen if a funder specified a repository that a funder did not support?) and to facilitate the development of a list of recommended research data repositories.

Findings

Were there any conflicts?
In total, 125 individual repositories were recommended or specified by one or more funders or journals. In all cases funding bodies made recommendations rather than specifying repositories so no conflicts were identified between funder and journal repository requirements.

Recommended repositories
There was no single repository which was recommended by all 30 funders/journals.

- The majority of funders/journals make repository recommendations based on data type, for example expression and sequence data or imaging and audio data rather than research subject.
- All have the option to use another suitable/subject or field specific repository, though some list requirements that a repository must meet in order to be used (see funder requirements).

The top 10 repositories that featured most frequently in the repository recommendation lists in our review are shown below:

- Array Express - EBI
- GenBank - NCBI
- NCBI Gene Expression Omnibus (GEO)
- Figshare
- UniProt knowledgebase
- Cambridge Crystallographic Data Centre (CCDC)
- DNA Data Bank of Japan (DDBJ)
- Dryad
- European Nucleotide Archive (ENA) - EBI
- Open Science Framework

Repository Finder

In order to interpret the extensive list of recommended repositories, we applied the open access FairSharing ontology to filter by subject areas related to health, based on metadata from re3data.org, which was extracted using the re3data.org API. In addition, we used other metadata fields attached to each repository record in re3data.org to facilitate the building of Repository Finder; a decision tree web app. This essentially enables users to filter the 276 repositories in the database by answering the questions in the app, with the result being a shorter list of repositories that could fulfil the user's requirements.

Each step of this auditing and review process has led to the point where the recommended repositories list can be shortlisted from a database of thousands to a handful of repositories that closely match a particular user's specific requirements, which greatly improves and speeds up research.

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

FUNDER/JOURNAL REQUIREMENTS

Data sharing requirements vary between funding body and journal. It is rare for a funder/journal to insist on using a specific repository, instead many provide a non-exhaustive list of repositories for you to choose from, with the option to use a discipline specific or institutional repository if no suitable repository is listed. These requirements tend to be detailed in data sharing or data management policies and guidelines available online.

Many provide a list of required attributes for repositories to be considered as suitable. Again these requirements vary, as an example the guidance from the European Research Council and PLOS journals are below:

European Research Council - Open Research Data and Data Management Plans
'When looking for a depository for research data it is important to check whether the depository:

1. *stores the data in a safe way;*
2. *makes sure that the data will remain findable (via the use of a persistent identifier), as well as accessible and re-usable;*
3. *describes the data in a standard way, using accepted metadata standards;*
4. *and specifies a license governing access and re-usability of the data.'*

PLOS ONE Journal – Repository inclusion criteria
'PLOS continues to consider adding new entries to our approved list of recommended repositories. The minimum criteria for inclusion are as follows:

1. *Dataset submissions should be open to all researchers whose research fits the scientific scope of the repository. PLOS' list does not include repositories that place geographical or affiliation restrictions on submission of datasets.*
2. *Repositories must assign a stable persistent identifier (PID) for each dataset at publication, such as a digital object identifier (DOI) or an accession number.*
3. *Repositories must provide the option for data to be available under CC0 or CCBY licenses (or equivalents that are no less restrictive). Specifically, there must be no restrictions on derivative works or commercial use.*
4. *Repositories should make datasets available to any interested readers at no cost, and with no registration requirements that unnecessarily restrict access to data. PLOS will not recommend repositories that charge readers access fees or subscription fees.*
5. *Repositories must have a long-term data management plan (including funding) to ensure that datasets are maintained for the foreseeable future.*
6. *Repositories should demonstrate acceptance and usage within the relevant research community, for example, via use of the repository for data deposition for multiple published articles.*
7. *Repositories should have an entry in FAIRsharing.org to allow it to be linked to the PLOS entry.'*

You can use the questions in the **Repository Finder** decision tree tool to obtain a list of repositories which closest fit your funder/journal requirements, such as access levels. To assist you, only repositories established and recognised in the field, and those following best practices are included. Including, but not be limited to, repository's adherence to metadata standards and provision of persistent identifiers.

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

ORGANISE YOUR DATA BEFORE SUBMISSION

You need to organise your data and files in a logical way.

- Ensure that all files are clearly named, so that it is easy to distinguish between them and understand what they contain.
- If you are submitting multiple files, many of which are similar, you may want to group files together - for instance by adding an indication of a grouping in the file names.
- Provide a document that explains the content of each of the files and descriptions of the datasets. See README file.
- Make sure that any abbreviations and acronyms used are explained.
- If specific software was needed to create the files, or is needed to open them, make sure to state the software name and version. Additionally, if you use proprietary formats consider submitting the same data in non-proprietary format too.
- Double check that there are no sensitive data included in the files you are planning to submit - you need to protect the privacy of your study participants (see Anonymisation and de-identification).

For spreadsheet data:

- do not embed any charts, images, comments etc.
- each table should be submitted as a separate file i.e. do not use multiple worksheets within Excel, save each of them as a separate file
- the first row should contain the header
- do not use multiple rows for the header
- make sure the names of the variables are clear (data labelling)
- remove any empty rows and columns
- make sure there are no merged cells
- make sure there are no special (non-alphanumeric) characters or commas in the spreadsheet
- remember that formatting such as coloured text or highlighting is not machine readable - avoid using them
- avoid mixing different data types in one column
- it is best to submit standard spreadsheets as CSV files (see non-proprietary formats)

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

FILE ORGANISATION

General

It is easiest and most efficient to decide on file/folder naming and organisation right at the start of the project. Using a logical and consistent structure will save time and prevent errors in the long run. It will also make it easier for others to understand your work and to locate specific files. You can add the necessary context through documentation (e.g. README files).

You want to create a system that allows you to:

- access your files easily
- identify the files without issue
- avoid duplication
- make back ups easy

There may already be systems and approaches established at your work place - it is worth checking with your colleagues and/or collaborators. Maybe you can adapt their system rather than develop your own from scratch (unless that's what you want to do of course!).

Use folders - and create a hierarchical structure. Group files by topic, start with a few folders that encompass the broader topics and then create more specific folders within them.

Archive completed work - separate folders/files for completed projects from the ongoing projects, especially when you work on many things and have a lot of files. It is also useful to review the completed projects before archiving them: they should contain only the necessary files.

Back up your work - regardless of where the files are stored you always want to have a back up strategy in place. Make it an easy one, perform it regularly and check that your back ups worked properly and are usable.

Clinical Trials

If you are working with **clinical trial data** there may be **regulatory requirements** in place that dictate how your files and documents should be organised in order to allow auditing and inspection. Have a look at Trial Master File Reference Model.

For instance, the TMF model consists of 11 'zones', including:

- **Data Management** -- Records related to Data Management activity on the trial. Includes subject data (completed CRFs or Final EDC Data) and the design and establishment of databases.

- **Statistics** - Records related to Biostatistics and Statistical Programming activity on the trial.

Sometimes it is easier to think about the types of files you will generate and need to store and work out the groupings from there - have a look at the TMF Reference Model for what you may be required to store (the 'artifact name' column here, for 'Data Management' and 'Statistics' zones).

1
Repository Finder
Different models of access
Options offered by repositories
List of funder & journal recommended repositories
Funder requirements

2
Data organisation
File organisation
File naming

3
Example: data structure
Example: data labelling
Non-proprietary formats list
Anonymisation and de-identification

4
Example: README file
Checklist of potential files to include

5
Data ownership
Deposit licences
Use licences
Example: patient consent for data sharing

6
Example: submission process

How do I share my data?

Back: overview of the main steps

FILE NAMING

Be **consistent**, **descriptive** and **explicit** in your file naming - even if you are not planning on sharing the files with anyone. Things that seems obvious to you today, may not make much sense in six months or a year. Save yourself the trouble!

Here are a few things you should consider including in your file names:

- Project identifier - e.g. name of the project or an acronym.
- Spatial information, especially if you have separate files for different sites.
- Name and initials of the researcher, particularly for bigger teams.
- Temporal information - this could be the date the sampling took place, or the range of dates for the experiment.
- Type of data that the file contains.

In terms of format:

- Try to keep your file names short - it is possible to include plenty of information if you use abbreviations, just make sure that you include a README file that clearly explains these abbreviations and the style of naming.
- Delimit words in a specific ways and avoid using spaces - some software is not happy with spaces in names. We suggest using a mix of camelCase or dashes (-), and underscores (_) as that makes it more machine-readable.

For instance, use *openLetter_pharmaGroup_2018-07-22_V3.txt* or *file-naming-conventions_V3.txt*. Essentially, different elements of the name (such as date and version in the first example) are separated by an underscore, and the words within each of these elements are delimited either by using capital letters (camelCase) or dashes.

- Use YYYY-MM-DD format for clarity - this will also automatically sort your files in chronological order.
- Fill in the zeros - use 09 instead of 9, or even 009 if you know the records will go into triple digits

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

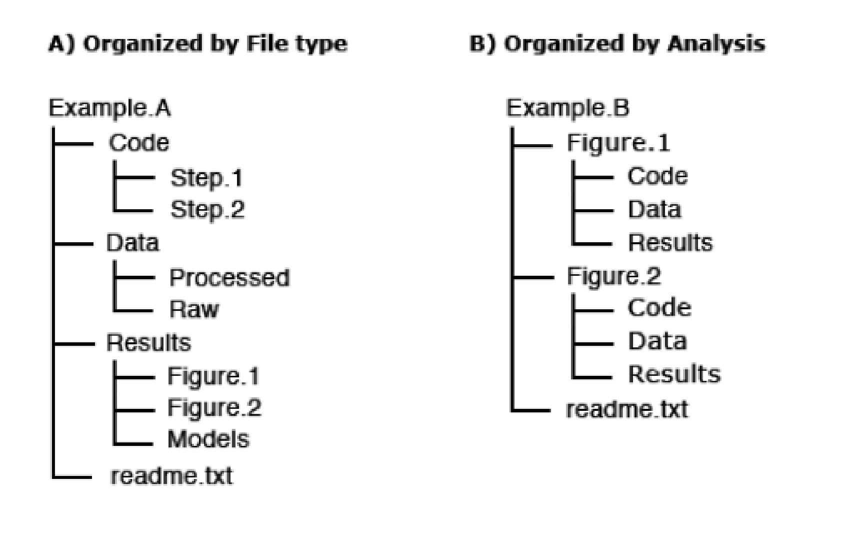
Back: overview of the main steps

EXAMPLE: DATA STRUCTURE

In order to make it easier for others to interpret your file structure it is important for it to be clear and logical.

Example 1: Best practices for creating reusable Dryad data packages

Organize files in a logical schema *Just as with file naming, it is important to generate a clear file structure that will be interpretable by others. For instance, it may be sensible to segregate the data, code, and results of the data package as in the examples below (by file type or analysis as presented in your publication). However you choose to structure the data files, it will be helpful to your fellow researchers to explain all relevant details in a README file.*



A) Organized by File type

```
Example.A
├── Code
│   ├── Step.1
│   └── Step.2
├── Data
│   ├── Processed
│   └── Raw
├── Results
│   ├── Figure.1
│   ├── Figure.2
│   └── Models
└── readme.txt
```

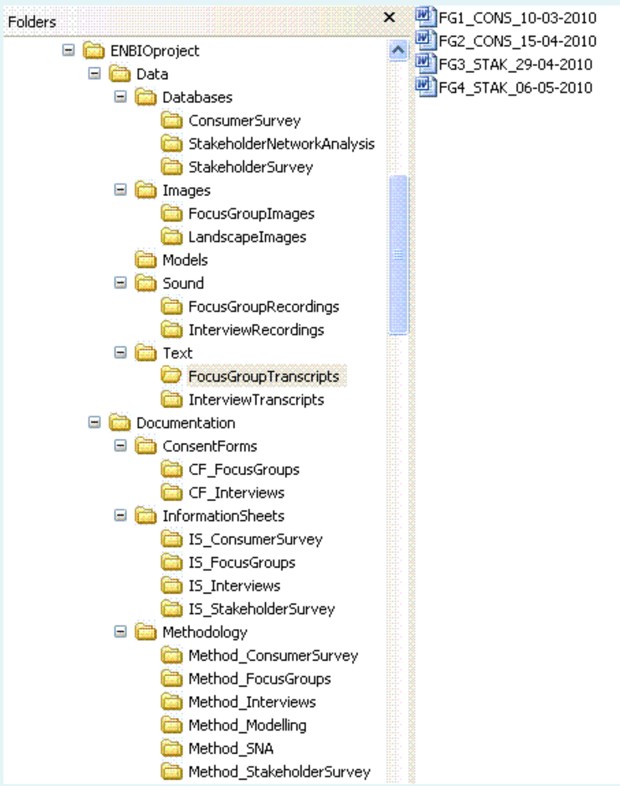
B) Organized by Analysis

```
Example.B
├── Figure.1
│   ├── Code
│   ├── Data
│   └── Results
├── Figure.2
│   ├── Code
│   ├── Data
│   └── Results
└── readme.txt
```

Example 2: UK Data Service: Organising Data

Example folder structure *In this example, data and documentation files are held in separate folders. Data files are further organised according to data type and then according to research activity. Documentation files are organised also according to type of documentation file and research activity.*

It helps to restrict the level of folders to three or four deep and not to have more than ten items in each list.



```
ENBIOproject
  Data
    Databases
      ConsumerSurvey
      StakeholderNetworkAnalysis
      StakeholderSurvey
    Images
      FocusGroupImages
      LandscapeImages
    Models
    Sound
      FocusGroupRecordings
      InterviewRecordings
    Text
      FocusGroupTranscripts
      InterviewTranscripts
  Documentation
    ConsentForms
      CF_FocusGroups
      CF_Interviews
    InformationSheets
      IS_ConsumerSurvey
      IS_FocusGroups
      IS_Interviews
      IS_StakeholderSurvey
    Methodology
      Method_ConsumerSurvey
      Method_FocusGroups
      Method_Interviews
      Method_Modelling
      Method_SNA
      Method_StakeholderSurvey
```

```
FG1_CONS_10-03-2010
FG2_CONS_15-04-2010
FG3_STAK_29-04-2010
FG4_STAK_06-05-2010
```

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back:
overview of
the main steps

DATA LABELLING - EXAMPLE

How you label your data, i.e. what names you give to the variables will affect how easy it is to work with them. It is best to keep variable names short and self-explanatory. Short names are easier to remember and type when working with data, and often easier to view in software too. Names with obvious meaning allow you (and others!) to figure out what the variables are without checking and re-checking the documentation.

However, there is no simple naming-convention that suits all data, and some variables are easier to deal with than others. For example, it is easy to see why using *ageYears* and *weight*, rather than *ay* and *w* is a good idea. But what if we had a more complex variable?

Lets say we need to distinguish between men/women, rural and urban setting, note that the variable is a percentage, and indicate that the subject was given one of 2 treatments.

In that case, for a woman from an urban population getting treatment A, we may name the variable something like:

female_treatmentA_urban_percentage.

This is pretty self-explanatory, but also rather long! On the other hand, using...

f_A_u_p

...is short, but very cryptic. Therefore, we may want to use something in between the two, for instance

fem_A_urb_perc

This way we keep the variable reasonably short and yet the name conveys quite a bit of meaning.

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

Suggested file formats

The file formats you use affect your ability to open those files at a later date. They will also affect the ability of other people to access those data. It is therefore important that you save your data in non-proprietary (open) formats whenever possible.

Ideally, the formats you use should be:

- non-proprietary
- uncompressed
- commonly used within your research community
- interoperable across variety of software and platforms

Some suggested file formats include:

- tabular data: csv
- text: plain text, RTF, XML, PDF/A, HTML, ASCII, UTF-8
- images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- geospatial: SHP, DBF, GeoTIFF, NetCDF
- databases: XML, CSV
- time series data: HDF5
- containers: TAR, GZIP, ZIP

If you need more information, the Library of Congress has published a Recommended Formats Statement that discusses this topic in great depth, and covers preferred and acceptable formats for different types of data.

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

DE-IDENTIFICATION OF SENSITIVE DATA

It is crucial that sensitive and proprietary data are handled appropriately before they are shared. You have a responsibility to your patients to ensure that their rights and privacy are protected, and that you are not putting them at risk by sharing their data.

De-identification consists of a collection of approaches and tools that are applied to data in order to remove personal (identifying) information. It is important to note that risks to individuals can remain in de-identified data, for instance inferences about individuals in the data could be made without re-identification, impacting on groups represented in the data.

Direct identifiers cannot be included in the data if you want the data to be classed as de-identified. These identifiers may relate to the individual, but also to their relatives, household members of even employers.

Examples of direct identifiers: names, geographic information below a certain level (e.g. this could mean information more specific than state or region), birth dates, telephone numbers, email addresses, biometric identifiers (e.g. finger prints), medical record numbers, account numbers, licence numbers, full-face photographs etc.

Indirect identifiers are the elements that may allow the identification of an individual through deduction.

Examples of indirect identifiers: gender, race, ethnicity, age, income, number of children, job title, place of work, medical condition etc.

This is clearly a very complex topic that cannot be sufficiently covered in a few paragraphs. As a starting point, you could explore the guides listed below. Most of the core principles are universally applicable, regardless of the more country-specific focus taken by some of these documents, so do not be put off by the fact that they were produced in a specific country.

The guides:

- **ANDS Introduction to sensitive data**
- **ANDS De-identification**
- **ANDS Publishing and sharing sensitive data**
- **ANDS Data sharing considerations for Human Research Ethics Committees**
- **NICHD DASH Data and Biospecimen Inventory De-Identification Guidance**
- **Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the HIPAA Privacy Rule**
- **De-Identification of Personal Information - NAtional Institute of Standards and Technology (U.S.)**

1
Repository Finder
Different models of access
Options offered by repositories
List of funder & journal recommended repositories
Funder requirements

2
Data organisation
File organisation
File naming

3
Example: data structure
Example: data labelling
Non-proprietary formats list
Anonymisation and de-identification

4
Example: README file
Checklist of potential files to include

5
Data ownership
Deposit licences
Use licences
Example: patient consent for data sharing

6
Example: submission process

How do I share my data?

Back: overview of the main steps

WHAT TO INCLUDE IN A *README* FILE?

Just like a README file helps your collaborators to engage in your project, a README file included with your deposited data will help other researchers understand and engage with these data. A well written README should also lower the risk of others misunderstanding your project and data, and help them make the most of the data.

You need to include enough detail so that it is clear why your data matters and clear how to get started with it. Make sure to use a non-proprietary format for this file to ensure that everyone can open it easily.

Below are some sections that you may want to consider for your project and data README files - not all of them will be applicable to all projects/data of course, but they should give you a good idea about what can be useful to include.

Project title -- One paragraph describing the project: say *what* the project is about. You could also include Motivation - and explain *why* the project has been undertaken.

Getting started -- Include instructions necessary to get a copy of the data up and running on a new machine.

Prerequisites and installing -- Is there anything that the new user will have to install on their machine before being able to use your data? List software as necessary and include installation instructions if needed.

File contents -- If you are submitting several files it is really helpful to list these and explain their contents. If there are any dependencies and/or relationships between the different files, state these too. If the datasets are very complex you may want to include a separate README document for each file and list file/dataset specific information there (explanation of variable names and calculations etc.)

Technical details -- Description of creation methods, protocols, and data sources. If the methods are published, include the references. Here you could share code style that you are using, with a code example; and explain your framework too.

Information warning --- If there are any known caveats or problems with the data, or if certain parts of the dataset are not of the highest quality it may be prudent to warn the future users about these.

Contributing -- You could include information about your code of conduct.

Versioning -- Include information about which version control system you use and - if you are including multiple versions of your data - make sure it is clear how these versions differ from each other.

Authors -- List authors and their contributions.

License -- State clearly what terms the data are licensed under: what can other researchers use your data for and what is not allowed?

Acknowledgements -- List people who contributed to your project and data. If there are any additional references that should be cited when using these data, list those too.

Metadata -- should be machine-readable and pass validation. Should include DOI (usually as the full URL).

Identifiers - make sure relevant data identifiers (e.g. DOI) and author identifiers (e.g. ORCID) are included.

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

DATA DEPOSITION CHECKLIST

- **Sensitive and proprietary data have been handled appropriately.** Double check this - it is crucial that you get this right.
- Data deposition has been **approved** by all relevant parties - you may want to include the obtained written **permissions** from your stakeholders or a data-sharing agreement.
- **README** file is included.
- **Data Dictionary** - particularly recommended for large, complex datasets. This file should give the names of all the variables, their descriptions, default units, typical ranges etc.
- **LICENSE** file - specifies under what license the data are made available. You may also want to include an attribution statement.
- **Metadata** are machine-readable and pass validation.
- Relevant **identifiers**, such as DOI, are included.
- If you need to include many references it may be worth creating a separate **CITATION** file.
- All files are **well formatted, clearly named** and saved using **non-proprietary formats**.

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

DATA OWNERSHIP

Determining the ownership of and rights relating to data can be difficult when the research project involves multiple researchers, funders and institutions. Therefore, it is important to clarify these aspects early on in the planning of a project.

Data Management Plan (DMP)
Ideally, data ownership and rights should be considered and documented in your DMP before your research starts and reviewed regularly throughout the project lifecycle.

The Digital Curation Centre guide on 'How to Develop a Data Management and Sharing Plan' provides the following guidance on what to include in your DMP regarding data ownership:

'***Data ownership*** *should be clarified and, where necessary, plans should be in place to negotiate licences at the start of the research process. If you agree/purchase licences to reuse third party data, be aware of any restrictions this places on subsequent deposit and data sharing. JISC provides lots of advice on copyright, IPR and relevant legislation such as the Data Protection Act and Freedom of Information. Institutional support is also available from experts in university libraries, records management and research offices.'*

Do you have the right to share the data?
Before you submit to a repository you need to know if you have permission to share the data. If you are not the data owner you need to make sure you have an agreement from the data owner prior to submission.

If third parties are involved data ownership and rights will usually be detailed within funding agreements or contracts. You need to be familiar with these agreements and contracts in order to know your rights when it comes to sharing the data.

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

DEPOSIT LICENCES

Why licence data?
A licence agreement is a legal arrangement between the creator/depositor of a dataset and a data repository. It is applied when you publish your data in a repository and signifies what a user is allowed to do with the data.

'Researchers (and computers) who find a dataset should immediately know what they are allowed to do with it. Stating clear re-use rights is like having a warm 'Welcome' on the doormat of your dataset. The motto is: 'open if possible, restricted if necessary'.' **CESSDA Training Data Management Expert Guide**

Which licence should you use?
As a general rule to maximise re-use you should choose a licence which allows your data to be used by the widest audience possible for the widest number of uses.

Before you start to consider which licence option might be the best fit for your dataset you should check whether the use of a certain licence is a condition of your repository of choice, your funder, publishing journal or your institution. For example, all data submitted to the Dryad repository is released to the public domain under a Creative Commons Zero (CC0) licence.

The **Creative Commons** website has a list of considerations, primarily for choosing a Creative Common licence, but they could be applied more broadly in your decision making.

Standard Licences
Whilst it is possible to generate your own bespoke licence, a standard licence would be the best fit for more research projects which do not have special requirements. Some of the standard licences available are shown below:

Creative Commons Licences provide a simple, standardised way to give your permission to share and use your creative work. There are several options available: **Creative Commons Licence chooser**.

Open Data Commons Licences are part of a set of legal tools to help you provide and use Open Data. Again there are several licencing options available: **Open Data Commons Licences**.

How to indicate which licence applies to your data
When you have chosen which licence you wish to apply to your data you need to make it clear to the user by adding both a human-readable and machine-readable statement.

Example statement from Open Data Commons: *This {DATA(BASE)-NAME} is made available under the Open Data Commons Attribution License: http://opendatacommons.org/licenses/by/{version}.*

The Creative Commons website suggests using **Author, License, Machine-readability** as a good rule of thumb to ensure that users know who to attribute, what they can do with your data and that the code you use is in a format that machines can understand.

The following guides might be a useful starting point for further information:

Open Data Institute - Publisher's Guide to Open Data Licensing
Digital Curation Centre - How to licence research data

1

Repository Finder
Different models of access
Options offered by repositories
List of funder & journal recommended repositories
Funder requirements

2

Data organisation
File organisation
File naming

3

Example: data structure
Example: data labelling
Non-proprietary formats list
Anonymisation and de-identification

4

Example: README file
Checklist of potential files to include

5

Data ownership
Deposit licences
Use licences
Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

USE LICENCES

Everybody is now buying in to data sharing but in order to maximise the potential of this data and address the inequities that exist between HICs and LMICs with regards to data ownership and reuse, it is important that all researchers have the skills to reuse shared data globally. We are planning to develop a toolkit to facilitate skill development in this area and would greatly appreciate feedback on your experiences with using shared data. To share your experiences, please send an email to info@theglobalhealthnetwork.org.

About data licences

A licence agreement is a legal arrangement between the creator/depositor of a dataset and a data repository. It defines what you (as a user of the data) can do with the data. The licence associated with the published data you wish to use should be clearly marked, if there is any ambiguity you should contact the owner of the data. You should not assume you have the right to re-use the data.

The Creative Commons website has list of considerations for licencees, primarily relating to Creative Common licences but can be applied more broadly to guide you:

Considerations

Understand the license

Read the legal code, not just the deed. The human-readable deed is a summary of, but not a replacement for, the legal code. It does not explain everything you need to know before using licensed material.

Make sure the license grants permission for what you want to do.

Scope of the licence

Pay attention to what exactly is being licensed. The licensor should have marked which elements of the work are subject to the license and which are not.

For those elements that are not subject to the license, you may need separate permission.

Some uses of licensed material do not require permission under the license. If the use you want to make of a work falls within an exception or limitation to copyright or similar rights, you may do so. Those uses are unregulated by the license.

Know your obligations.

Provide attribution. If the licence requires you, provide attribution* and mark the material when you share it publicly. The specific requirements vary slightly by licence.

Determine what, if anything, you can do with adaptations you make

Depending on what type of license is applied, you are limited in whether you can share your adaptation and if so, what license you can apply to your contributions.

Consider licensor preferences. Consider complying with non-binding requests by the licensor. The licensor may make special requests when you use the material. We recommend you do so when reasonable, but that is your option and not your obligation.

***Attribution**

An attribution requirement means that the licensor must be given due credit for the work when it is distributed, displayed, performed, or used to derive a new work. Digital Curation Centre

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

How do I share my data?

Back: overview of the main steps

EXAMPLE: PATIENT CONSENT FOR DATA SHARING

Below are two examples of text associated with data sharing from consent forms associated with data deposited in the UK Data Archive:

Example: The 10/66 INDEP mixed methods study of the economic and social impact (at household level) of residing with a care dependent older person in China, Mexico, Peru and Nigeria (UK Data Archive. 10.5255/UKDA-SN-852071)

'**Who will have access to any personal information that I provide?**
... In accordance with the policies of the funders of the project, the data that has been collected will, after the end of the project, be made available to other researchers for legitimate scientific research purposes. This is to ensure that the best and fullest possible use is made of the information that you have provided. This process will be managed by the <insert>, who will be responsible for approving or rejecting applications to use the data. Any data released to other groups will be fully anonymised, that is, any information that could possibly identify you or others will be removed.'

Example: Doing TB Differently (UK Data Archive. 10.5255/UKDA-SN-852112)
'***Confidentiality***
Your contributions to the project will be treated with confidence. They will not be used other than for the purposes described above and third parties will not be allowed access to them (except as may be required by the law). If you request it, you will be supplied with a copy of your contribution to check it before we publish. Your data will be held in accordance with the Data Protection Act. Data will be lodged with the UK Data Archive Service so it is available for secondary data analysis. Should you request it, you will not be personally identifiable in any of the output posted on this site. If you strongly object to your contributions being used in this way, we shall reserve the right to withhold this data from the public realm.'

The UK Data Service and Inter-university Consortium for Political and Social Research (ICPSR) websites offer guidance on the type of language to use and avoid for data sharing.

1

Repository Finder

Different models of access

Options offered by repositories

List of funder & journal recommended repositories

Funder requirements

2

Data organisation

File organisation

File naming

3

Example: data structure

Example: data labelling

Non-proprietary formats list

Anonymisation and de-identification

4

Example: README file

Checklist of potential files to include

5

Data ownership

Deposit licences

Use licences

Example: patient consent for data sharing

6

Example: submission process

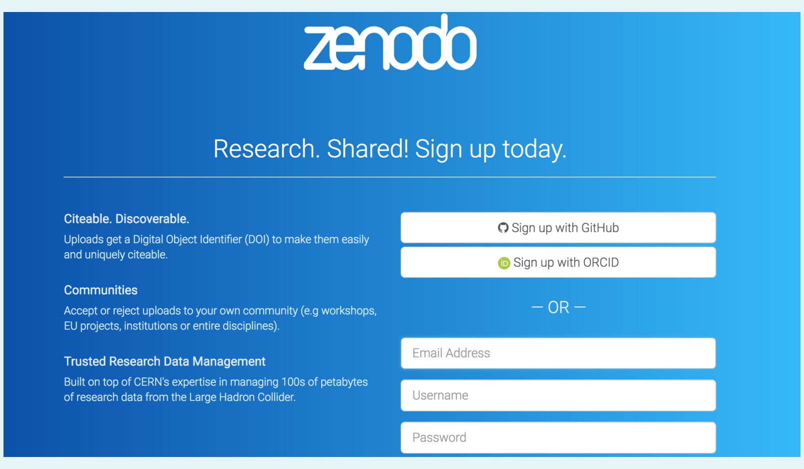
How do I share my data?

Back: overview of the main steps

SUBMISSION PROCESS - EXAMPLE: ZENODO

Below you will find screenshots of the deposition process for Zenodo. Of course the details will be specific to Zenodo, but it should give you a general idea of what to expect from a repository.

First, you will need to sign up to the repository. Zenodo offers three options - signing up with GitHub, ORCID or using your email address.



zenodo

Research. Shared! Sign up today.

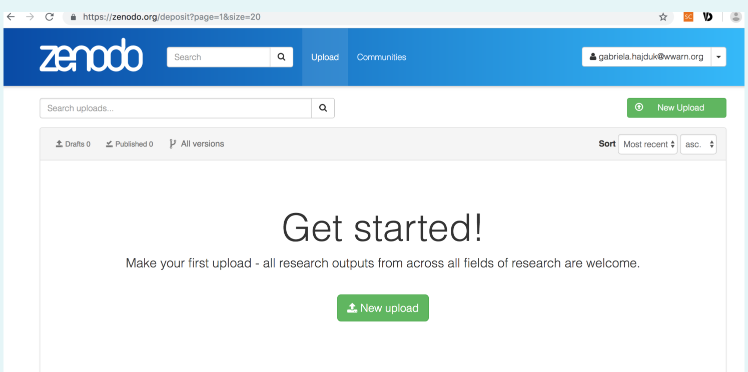
Citeable. Discoverable.
Uploads get a Digital Object Identifier (DOI) to make them easily and uniquely citeable.

Communities
Accept or reject uploads to your own community (e.g workshops, EU projects, institutions or entire disciplines).

Trusted Research Data Management
Built on top of CERN's expertise in managing 100s of petabytes of research data from the Large Hadron Collider.

Sign up with GitHub
Sign up with ORCID
— OR —
Email Address
Username
Password

Once you signed up and logged in you will be able to click the Upload button and the top of the screen and get started with your deposition.



https://zenodo.org/deposit?page=1&size=20

zenodo · Search · Upload · Communities · gabriela.hajduk@wwarn.org

Search uploads... · New Upload

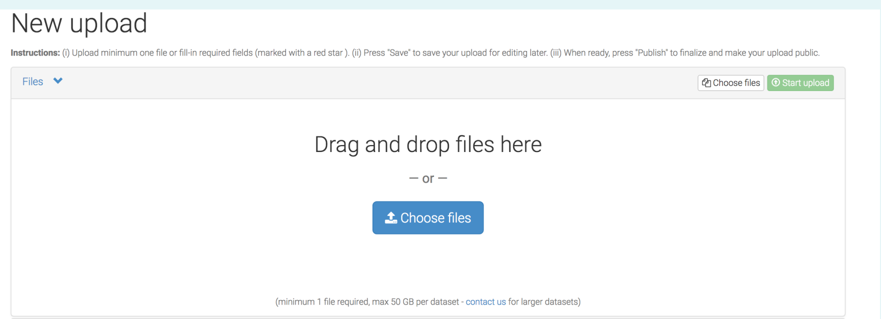
Drafts 0 · Published 0 · All versions · Sort Most recent · asc.

Get started!

Make your first upload - all research outputs from across all fields of research are welcome.

New upload

You should prepare all your files for deposition beforehand - it will be much easier that way. Upload your files either by dragging-and-dropping them onto the browser window or clicking the blue button and choosing them that way.



New upload

Instructions: (i) Upload minimum one file or fill-in required fields (marked with a red star). (ii) Press "Save" to save your upload for editing later. (iii) When ready, press "Publish" to finalize and make your upload public.

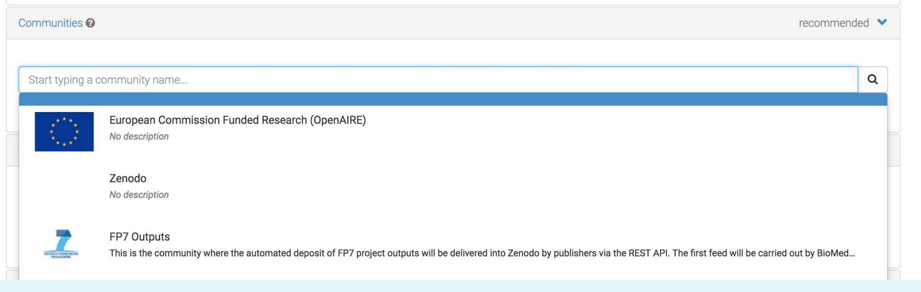
Files · Choose files · Start upload

Drag and drop files here
— or —
Choose files

(minimum 1 file required, max 50 GB per dataset - contact us for larger datasets)

Zenodo offers 'communities' you can assign your deposition too. Lets say that your study has been funded by European Commission (OpenAIRE) - you may want to assign the submission to that community.

If you are not sure whether there is a suitable community you should choose, have a browse first. You can search for different communities here.



Communities · recommended

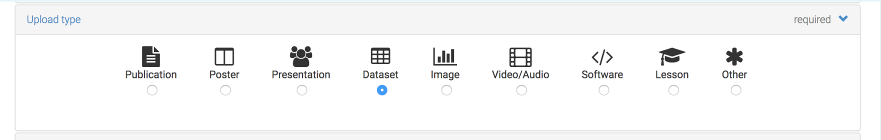
Start typing a community name...

European Commission Funded Research (OpenAIRE)
No description

Zenodo
No description

FP7 Outputs
This is the community where the automated deposit of FP7 project outputs will be delivered into Zenodo by publishers via the REST API. The first feed will be carried out by BioMed...

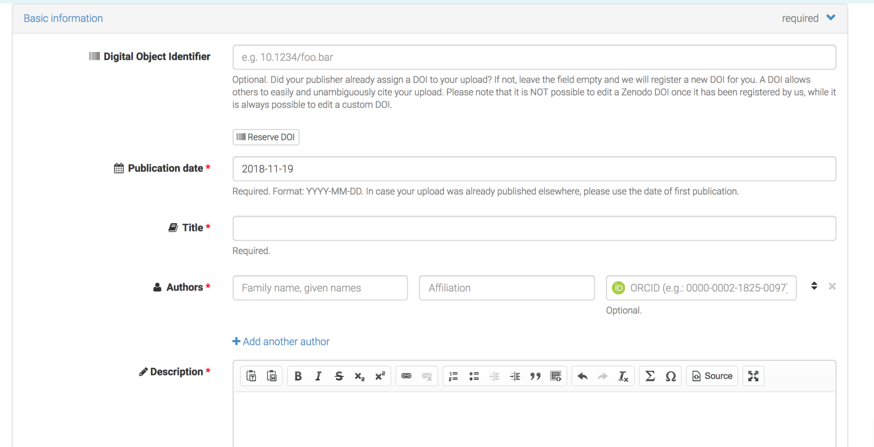
You can then indicate what you are uploading - is it a poster from a conference that you want to share? Or a dataset for archiving?



Upload type · required

○ Publication · ○ Poster · ○ Presentation · ● Dataset · ○ Image · ○ Video/Audio · ○ Software · ○ Lesson · ○ Other

Fill out the information about your submission - if you have prepared a README file you should be able to copy-and-paste a lot of the content into appropriate sections here.



Basic information · required

Digital Object Identifier: e.g. 10.1234/foo.bar
Optional. Did your publisher already assign a DOI to your upload? If not, leave the field empty and we will register a new DOI for you. A DOI allows others to easily and unambiguously cite your upload. Please note that it is NOT possible to edit a Zenodo DOI once it has been registered by us, while it is always possible to edit a custom DOI.
Reserve DOI

Publication date *: 2018-11-19
Required. Format: YYYY-MM-DD. In case your upload was already published elsewhere, please use the date of first publication.

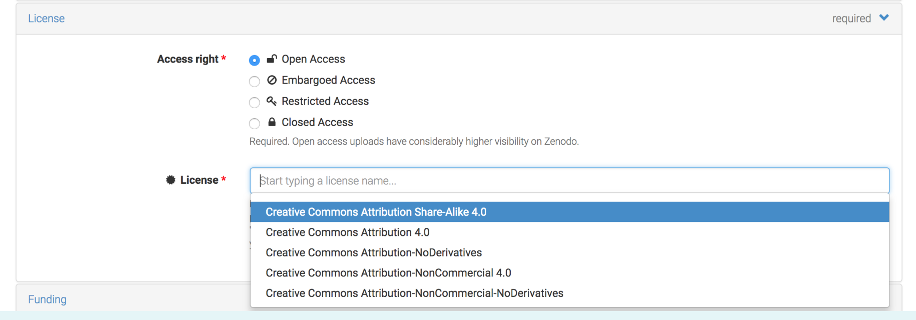
Title *:
Required.

Authors *: Family name, given names · Affiliation · ORCID (e.g.: 0000-0002-1825-0097)
Optional.
+ Add another author

Description *: B I S x₂ x² … Σ Ω Source

Importantly, you need to choose the type of access and license to be used.

If you choose to embargo your data, you will need to provide the end date for the embargo (when your data will become open access). If you choose Restricted Access (a.k.a. controlled access) you will be given a chance to specify the conditions under which you grant users access to the files in your upload.



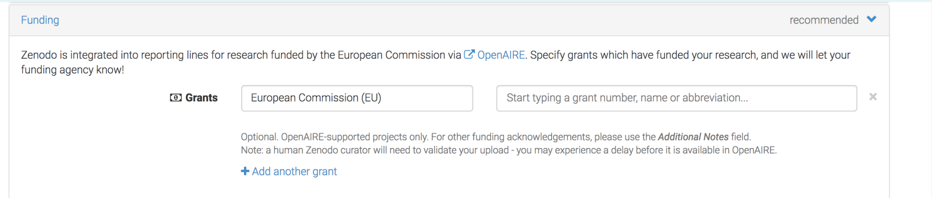
License · required

Access right *:
● Open Access
○ Embargoed Access
○ Restricted Access
○ Closed Access
Required. Open access uploads have considerably higher visibility on Zenodo.

License *: Start typing a license name...
Creative Commons Attribution Share-Alike 4.0
Creative Commons Attribution 4.0
Creative Commons Attribution-NoDerivatives
Creative Commons Attribution-NonCommercial 4.0
Creative Commons Attribution-NonCommercial-NoDerivatives

Funding

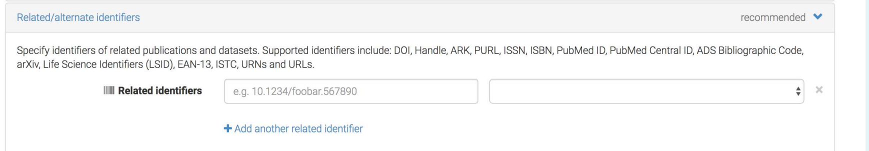
Let Zenodo know who your funders are and if there are any other identifiers (such as PubMed ID) that are related to your deposition.



Funding · recommended

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: European Commission (EU) · Start typing a grant number, name or abbreviation...
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

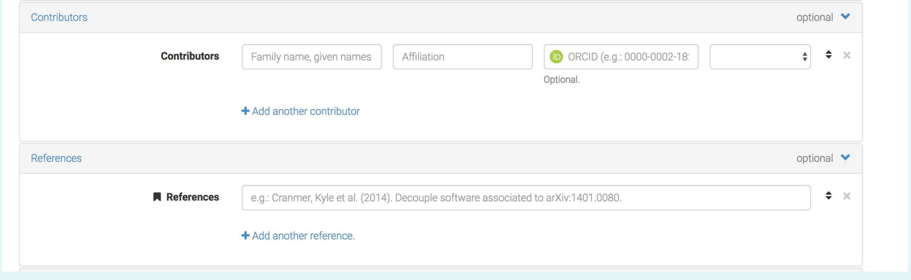


Related/alternate identifiers · recommended

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: e.g. 10.1234/foobar.567890
+ Add another related identifier

You will also have a chance to add contributors as well as relevant references. Again, if you have written a README file, you should be able to refer to it as you are filling this form out.



Contributors · optional

Contributors: Family name, given names · Affiliation · ORCID (e.g.: 0000-0002-18
Optional.
+ Add another contributor

References · optional

References: e.g.: Cranmer, Kyle et al. (2014). Decouple software associated to arXiv:1401.0080.
+ Add another reference.

Zenodo then provides a few more optional sections so you can indicate any conferences, theses, or book chapters that your deposition is related to.

Finally, once you are happy with all the information you can publish your submission to Zenodo!

As you can see, the process is not complicated, provided you do your homework beforehand and prepare your data, files and documentation well. Preparation will allow you to copy across the necessary information quickly and efficiently.

1

Repository Finder
Different models of access
Options offered by repositories
List of funder & journal recommended repositories
Funder requirements

2

Data organisation
File organisation
File naming

3

Example: data structure
Example: data labelling
Non-proprietary formats list
Anonymisation and de-identification

4

Example: README file
Checklist of potential files to include

5

Data ownership
Deposit licences
Use licences
Example: patient consent for data sharing

6

Example: submission process