

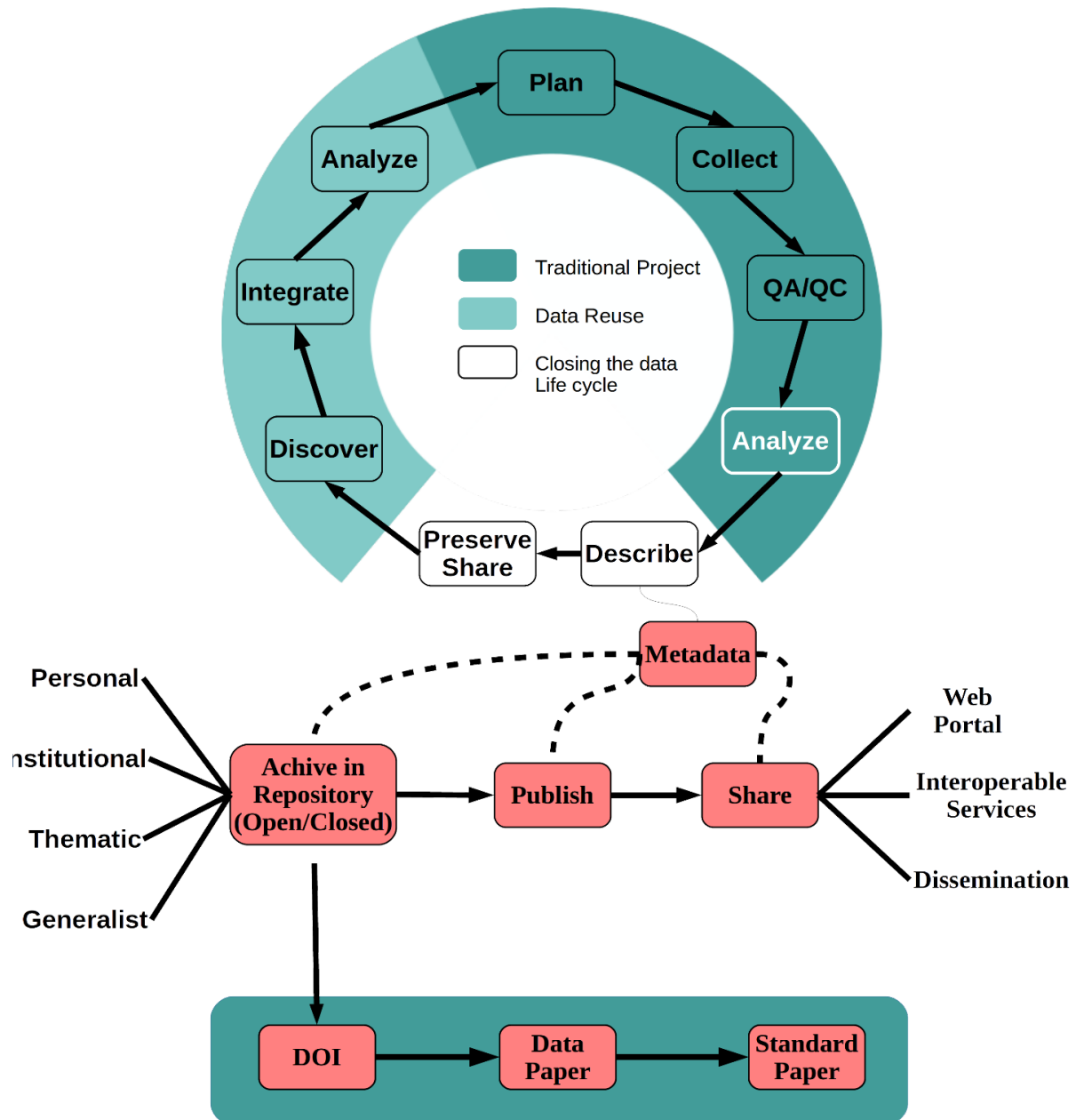
## FAIR Checklist for Data & Software sharing (Psychology)

Public data sharing has become a popular practice among research psychologists at the University of Sheffield to increase the transparency and openness of their research discoveries. However, perceived barriers included lack of training and guidance, technical difficulties in sharing data, concerns on misuses by others, privacy of personal data as well as commercial sensitivity. Here, we describe FAIR – a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable. We also introduce you to a checklist to provide guidance on where and how data can be shared and reused effectively. The checklist involves all stages of the research data life cycle.

### FAIR checklist elements in Research Data life cycle

1. **Findable**: DOI; Metadata; Repository
2. **Accessible**: Privacy of personal and sensitive data; Authentication and Authorisation; Standard protocol; Obligations and Rights (Data Policy)
3. **Interoperable**: Data format and standards (Describe)
4. **Reusable** - Storage and Archive; Licensing (Share); Publish

# Research Data Life Cycle



Research Data Life Cycle image source from: <https://doi.org/10.5281/zenodo.1149040>

## 1. Findable

The first step in (re)using data is to find them. Machine-readable metadata and data should be easy to find when you share your data. Automatic discovery of datasets and services is

essential. Findable data sharing requires globally unique, persistent identifiers, rich metadata to describe your data that includes the persistent identifier, and indexing by repositories or registries in specialised topical domains ([go-fair.org](https://go-fair.org)).

## 1.1 General

### 1.1.1 Does your data have an unique and persistent identifier?

A globally unique and persistent identifier (e.g., digital object identifier, [DOI](#); Open Researcher and Contributor ID, [ORCID](#)) needs to be assigned to your identifiable data item with data elements and metadata.

Tips: it could be provided by good repositories and services like open science framework ([OSF](#)), [UK data Archive](#) and University of Sheffield's data repository [ORDA](#).

### 1.1.2 Do you supply your data with rich metadata?

Rich metadata is to explain your data which met with [metadata standards](#) (e.g., Data Documentation Initiative, [DDI](#); Data Catalogue Vocabulary, [DCAT](#); Resource Description Framework in Attributes, [RDFa primer](#)).

Tips: Need to know what to include in metadata? See [An example of things your metadata might need to cover](#).

[An example of things your metadata might need to cover:](#)

- Specifies your persistent identifier
- Describes your data structurally and contextually
  - 1) A structural metadata describes the data itself (in a README file or documentation):
    - basic description of the research
    - file size, filenames/naming conventions and file structure/relationship between files
    - Methodology including how data was created and processed
    - illustration if applicable
    - abbreviations if applicable
  - 2) A contextual metadata describes (in repositories):
    - where documents are created
    - Who created them

- When documents are created
- An example of [survey metadata](#) for structural description (Lavrakas, 2008) should address at least :
  - A definition of survey variables. Each variable must contain a valid value, a code, an identifying name, and so on.
  - Data model. This model describes relationships between the variables regarding their categories and hierarchies.
  - Route instructions. This defines order, such as how questions are asked.
  - Relationships. Whenever relationships impose restrictions on the values of variables, the restrictions must check the consistency among the collected data.

### 1.1.3 Is your (meta)data registered or indexed in a searchable resource using repositories?

(Meta)data needs to be deposited online in a researchable resource with catalogue. You could use a subject-specific repository (e.g., [OSF](#), [UK data service](#)) or institutional repository (e.g., [ORDA](#)).

It can be useful to keep metadata even the data is not available (any longer). Especially for sensitive data or deleted data, it provides archival information as well as gives others some indication of what kind of data they can expect if they request the data. (Also see 2.1.4 A special consideration on personal data and sensitive data)

If open data is not available, provide instructions on how to request the data (e.g., contacting the author and material transfer agreement [MTA](#) between institutions).

Tips:

- To find more subject-specific repositories, visit: [re3data](#) or [opendoar](#)
- If you use a subject-specific repository to deposit your (meta) data, maybe you could link/integrate it with your institutional repository.

## 1.2 Secondary data

(1) Keep a record of data, where and how it was found.

(2) Check on 1.1 which(s) would apply.

### 1.3 Software and code

(1) If the original software and code has been used, provide information on software used to process data like giving full details including version used.

(2) If the adaptive or modified software and code has been used, provide information on what details has been adjusted and make it findable (1.1 and 1.2).

(3) Software is often made public through GitHub. However, this repository does not equip the software with a persistent identifier and does not ensure longevity of storage. To achieve that [Zenodo](#) stores each version of your software and gives each version a separate DOI. This allows for long-term accessibility while acknowledging the need for updating your software. This process can be automated for you - contact [Code Clinic](#) for support.

(4) Register your software and code in a community registry if applicable.

Tips:

- Find [Awesome Research Software Registries](#)
- Book a [Code Clinic](#) appointment or contact [code-clinic-helpers-group@sheffield.ac.uk](mailto:code-clinic-helpers-group@sheffield.ac.uk).

(5) Check on 1.1 which(s) would apply.

## 2. Accessible

Once the user finds the required data/code, they need to know how they can be accessed, possibly including authentication and authorisation. Therefore, the (meta)data can be retrieved by identifier with a standardised protocol which is open, free and universally implementable and allows for an authentication and authorisation procedure. However, FAIR does not mean completely open. The access to your data/code needs to be under condition or restrictions where appropriate. Your metadata should be available, even if the data is no longer available ([go-fair.org](#)).

### 2.1 General

#### 2.1.1 Make sure that your persistent identifier works

Following the persistent identifier (PI) should take you to the metadata (highly sensitive data only) and/or data.

Tips: We suggest you try on your own or by your colleagues by searching the PI to test whether it takes you to your data object.

#### 2.1.2 Do you use a standardised protocol?

The protocol by which data can be retrieved follows recognised standards, for examples, Hypertext Transfer Protocol ([HTTP](#)) or File Transfer Protocol ([FTP](#)).

Tips: Don't worry if these terms sounds unfamiliar to you. Most places will provide these two common types as examples.

#### 2.1.3 Does the access procedure include authentication and authorisation?

Especially also for potential commercial use of FAIR data, companies need to have a clear appreciation and legal position on their ability to (re)use your data. Non-licensed data, although 'open' in the mind of most academics, should be avoided by most major companies, due to legal risks (for instance for patient privacy or intellectual property reasons). We therefore consider appropriate licensing of Data Objects (or even individual data elements within them) as key to FAIR data publishing.

Tips: see more in 4. Reusable about Licensing.

#### 2.1.4 A special consideration on personal data and sensitive data

If your data is particularly sensitive you can still share it for future research through restricted access. Most repositories allow this.

(1) Provide an email, telephone number, name of a contact person if applicable.

Alternatively, you could list a role in the department (e.g., Departmental Director of Research) who can discuss the access to the data later.

(2) Ensure participant consent is in place from the start of the project and consider:

a. using a comprehensive and negotiable consent to include plans for long-term storage, analysis, sharing of research data and reusing the data.

b. plan for changes in content given by parents when children participants grow up.

Tips:

- deleting data after a certain period of time even its access has been restricted.
- including a risk assessment of the longitudinal sensitivity.
- a possibility of re-consent in the future if applicable.

(3) Prepare a time-limited data sharing contract requiring pre-registered data analysis plan from potential requesting academics for sensitive clinical data which has strict conditions for data sharing (e.g. data from NHS patients).

(4) Be aware of General Data Protection Regulation ([GDPR](#)) and special category data.

Tips: To learn more about what is a [special category data](#).

(5) Use anonymisation or pseudonymisation, but be aware that re-identification may still be possible.

Tips:

- particularly considering platform usernames and IP addresses might be identifiable from online studies.
- a possible re-identification within participants' institution, such quoting an expression frequently used by an individual in speech and/or writing.

(6) Discuss and make arrangements for data sharing with external partners if applicable.

(7) Metadata should be accessible, even if the data is not available. Either because of restricted access or data being no longer available, metadata leaves the trace that the data is stored in a restricted form and gives information of the file properties.

Tips: more questions about personal and sensitive data? Contact Research Services or the departmental open science lead/ research and ethics committees

## 2.2 Secondary data

Be sure about your obligations and rights when access to secondary data:

(1) Are there restrictions on sharing the data and data derived from them?

(2) Do arrangements have to be made with the original researcher(s) if there is no clear guide available?

(3) If you are using secondary clinical data from NHS, you need NHS/IAPT licence to access material for specific uses (rather than general use), and have ethical approval for this specific secondary data use by NHS Research Ethics Committee.

(4) Check on 2.1 which(s) apply.

## 2.3 Software and code

Be sure about your obligations and rights:

(1) Check on the copyright and restrictions.

(2) If you need to adapt and develop on proprietary software and code, do you need to make an agreement personally or via institution?

(3) Check on 2.1 which(s) apply.

## 3. Interoperable

The data usually needs to be integrated with other data. Therefore, (meta)data should conform to recognised formats and standards. In addition, the data needs to interoperate with applications or workflows for analysis, storage, and processing between humans and machines ([go-fair.org](https://go-fair.org)).

### 3.1 General

#### 3.1.1 Do you use an open format?

Data should be provided in commonly understood and preferably open format. If your raw data or processed data is formatted in a specialised or proprietary way, consider transferring your data to an open or more widely accessible format. Also, It is helpful to provide information on details how format has been transformed (e.g., your scripts or listing what software/packages you have been using).

Tips:

- See [Table of Accepted and Recommended File Formats](#).
- Visit <https://5stardata.info/en/> for more detailed guidance.
- A useful resource: [FAIR Data Point \(DP\) specification](#). DP is a software that, from one side, allows data owners to expose datasets in a FAIR manner and, for another side, allows data users to discover properties about offered datasets (metadata) and, if license conditions allow, the actual data can be accessed.



### Table of Accepted and Recommended Data Formats

Data type	Recommended	Accepted
Textual	Txt, <u>rtf</u> , XML	HTML, MS Word, NVivo
Tabular	CSV, SPSS portable (.por), structured XML	SPSS, MS Excel, MS Access
Image	Uncompressed TIFF	JPEG, GIF, RAW, PNG
Audio	FLAC	MP3, WAV
Video	MP4, OGG	AVCHD

#### 3.1.2 Do you specify data structure and formats in your metadata?

You need to provide metadata that specifies the data structure and formats and follows relevant standards (e.g., Data Documentation Initiative, [DDI](#); Data Catalogue Vocabulary, [DCAT](#); Resource Description Framework in Attributes, [RDFa primer](#)).

Tips:

- see [A logical file structure](#) and [Table of Accepted and Recommended File Formats](#).
- A useful resource: [FAIR Data Point \(DP\) specification](#).

#### 3.1.3 Do you use consistent and informative terms?

(1) Controlled, consistent vocabularies, keywords, thesauri or ontologies should be used where possible.

(2) Come up with [consistent filenames](#) rules to make your file names informative, and make sure that everyone involved in the research follows them.

#### How to make your file names informative?

- Use consistent punctuation, spelling, version numbers and abbreviations
- Use YYYYMMDD date format, if you need dates in file names to keep track of versions
- Make filenames concise but meaningful
- Record naming conventions in README file

- Avoid repetition within filenames and folders
- If you have individual files for each participant, in some cases you can add a script that generates anonymous, consistent file names

### 3.2 Secondary data

(1) Check on 3.1 which(s) apply

### 3.3 Software and code

(1) Qualified references and links are provided to other related data/software and code if applicable.

(2) Check on 3.1 which(s) apply.

## 4. Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings. (Meta)data should conform to community norms and be clearly licensed to specify permission for potential users. This should provide a close link to 1. Findable ([go-fair.org](https://go-fair.org)).

### 4.1 General

#### 4.1.1 Has your data been stored properly?

When conducting your research

If you collect your data in a physical form, it is recommended that you digitise it. All digital data should be stored in a secure place:

- Research data storage on the X:drive is recommended
  - IT Services can provide up to 10TB per research group (more storage can be requested if needed)
  - Secure and backed up several times a day
  - Can be shared with co-researchers / supervisors
  - <https://students.sheffield.ac.uk/it-services/research/storage/standard>
  - <https://students.sheffield.ac.uk/it-services/research/storage/using>
- University Google drive can also be used
  - Useful for sharing data with external research partners
  - Not suitable for highly sensitive data
- If not possible to use University drives

- Backup and test backups
- Use 3-2-1 rule (3 copies, 2 different media, 1 off-site)

If you cannot digitise it or physical data is important for validation/reproduction of your research, make sure you have access arrangements for physical data.

When you are publishing your research:

A good [repository](#) provides the dataset with a persistent identifier such as DOI e.g. Open Science Framework ([OSF](#)), Online Research Data ([ORDA](#)), and ensures longevity of the storage. OSF is a general repository, ORDA is an University of Sheffield's institutional repository. You might also want to use a subject-specific repository:

- PsychArchives - psychology-specific, manuscripts, data and code
- PsyArXiv - psychology preprints
- Social Science Open Access Repository (SSOAR) Share software and code created to process data, or details of proprietary software used.
- You can find more repositories on [openoar](#) or [re3data](#).

Dropbox, ResearchGate, Academia.edu are not repositories. GitHub is not a permanent repository, which needs to add DOI and ensure longevity of storage (min. 10 years).

If you are sharing sensitive data, you can use only restricted data access possible repositories like [UK Data Service](#) and Inter-university Consortium for Political and Social Research [ICPSR](#). Controlled access of data allows for validation and reuse of your data, whilst maximising protection of participants' identity. Controlled access can be requested in most data repositories, then any researcher interested in your data contacts you directly, describing who they are and what purpose they need the data for so you can decide whether it is appropriate to share it or not.

If data can only be made available on request, you should give details of how access can be requested in a Data Availability Statement.

#### 4.1.2 Comply with your data management plan

Make sure that your data sharing complies with your reviewed and updated data management plan. A data management plan is usually required by funders. There are some templates available on [DMPonline](#).

Tips:

What is a data management plan:

<https://www.sheffield.ac.uk/library/rdm/whatisrdm>.

Where to do a data management plan: [DMPonline](#).

Who to contact: [rdm@sheffield.ac.uk](mailto:rdm@sheffield.ac.uk).

#### 4.1.3 Has each data file been well-described in metadata?

Your shared data needs to be accurately and thoroughly described in metadata with relevant attributes. These could contain information about:

- why was the data generated/collected
- what does the data set include
- how, when and by whom was the data generated
- how, when and by whom was the data processed
- how can the data be reused properly
- variable names are explained or self-explanatory (i.e. defined in the research field's controlled vocabulary).
- version of the archived and/or reused data is clearly specified and documented.

Tips: Also see 1.1.2

#### 4.1.4 Do you publish with a clear and accessible usage licence?

When publishing research data in a repository the choice of licence defines who can re-use the data, and for what purposes. These licences are designed to be open, which means the material released under these licences can be freely re-used without explicit permission being required, as long as the conditions of the licence are adhered to.

The most commonly used for sharing academic data and in academic publishing is the Creative Commons ([CC](#)) set of licences. For example, the CC-BY licence being the main one used for open access papers and books.

## How to choose your licence?

- Have you checked funder requirements before choosing a licence?
- Creative Commons ([CC](#)) licences provide a range of options. Are you clear about which option is suitable for your (meta)data?
  - BY - Attribution: The name of the creator must be provided and where technically possible a link to the original material and the CC licence
  - ND - No Derivatives: The work may be modified, but the modified version may not be distributed.
  - SA - Share Alike: The work may be modified, but the modified version may be distributed only under the same licence as the original
  - NC - Non-commercial: The material may be used only for non-commercial purposes
- NHS services/data ([IAPT](#)) sometimes require different copyright and licences ([open government licence v3.0](#)).

If CC does not meet your data's need, go to <https://choosealicense.com/> or contact [copyright@sheffield.ac.uk](mailto:copyright@sheffield.ac.uk)

### 4.1.5 Do you publish openly through repositories?

Repositories should be organised thematically and are easier for secondary access. Make sure that you use a logical file structure based on your topics.

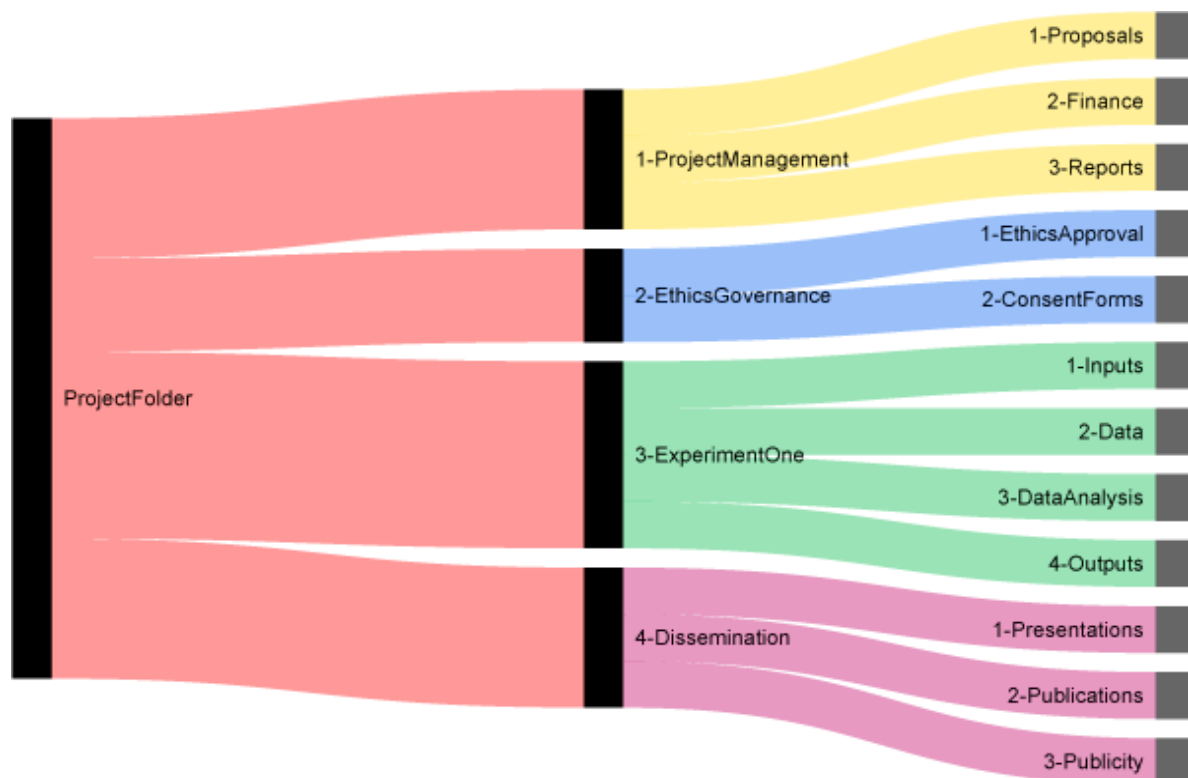
Tips: see [An example of logical file structure](#)

When submitting data to a repository, make the details that you give are clear and informative.

If you cannot share all of your data, do you share metadata and some of your data (e.g. fully anonymised, analysed or sample data) openly through a repository (Also see 2.1.4 A special consideration on personal data and sensitive data in 2. Accessible).

Tips: also see 1.1.3

[An example of logical file structure](#)



#### 4.1.5 When publishing with a data availability statement

The best way to inform readers, if and how underlying data can be accessed, is to include a data availability statement. Not only does this enable readers to access the data, it enables them to cite your data more easily and accurately if they reuse it. A data availability statement is usually required by funders and journals.

If data is available through a repository, you should include the dataset DOI, which provides a direct link to the dataset. If data can only be made available on request, you should give details of how access can be requested in a Data Availability Statement. It is also recommended to mention whether and how you will communicate with potential users about any updates and changes after publication.

Examples of Data Availability Statements ( see <https://www.sheffield.ac.uk/library/rdm/publish>)

- Data supporting this publication can be freely downloaded from the University of Sheffield Research Data Repository at <https://doi.org/<your doi here>>, under the terms of the Creative Commons Attribution (CC BY) licence.

- Data supporting this publication include personal information, and may be obtained by contacting <group email address>@sheffield.ac.uk. A signed Data Sharing Agreement may be required to comply with patient consent.
- Data supporting this publication are confidential, and can only be supplied by our industrial partner, <name>.

Further examples of data access statements can be found on the [University of Manchester library website](#).

#### 4.1.6 Check other conditions and terms

When sharing your data, you need to double check with conditions and terms agreed with or set by participants (consent also see 2.1.4), external/ commercial partners and third-party data providers.

You might also need to check if [Export Control Legislation](#) applies, when sharing data outside the UK.

#### 4.2 Secondary data

(1) If you have used data in your research that is publicly and permanently available, do you share a link (to the repository access with a DOI) rather than the actual data (raw downloaded file)?

(2) Clarify the restrictions and terms on original data and your processed data

(3) Check on 4.1 which(s) apply

#### 4.3 Software and code

(1) Archive along with the data/sample data if possible.

(2) Create link in ORDA to version used to produce final dataset if multiple versions of created software / code available on GitHub.

(3) Software and code meets relevant domain standards ([The Dublin Core Metadata Initiative](#), [CodeMeta](#), [Schema.org](#)).

(4) Choose a licence for your software (e.g., [GNU GPL v3.0](#)).

(5) Make your code citable ([CodeMeta](#), [Citation File Format](#)).

(6) Give full details including version used if using proprietary software used to process data.

(7) If your code is not open due to the nature of the language or restriction of copyright, make sure that you have included or indicated the sources of all required supplementary software and code.

(8) Check on 4.1 which(s) apply.

#### More useful resources

University of Sheffield

Data management: <https://www.sheffield.ac.uk/library/rdm>

Research service: <https://staff.sheffield.ac.uk/research-services>

Code clinic: <https://rse.shef.ac.uk/support/code-clinic/>

External

RDMkit: <https://rdmkit.elixir-europe.org/researcher>

Data Management Skillbuilding Hub: <https://dataoneorg.github.io/Education/>

NIHR:

<https://www.nihr.ac.uk/documents/nihr-position-on-the-sharing-of-research-data/12253>