FAIR data and software - a checklist for ScHARR researchers

Researchers should consult the ScHARR IG Policy, Section 5 - Information Sharing [ScHARR Information Governance Policy | ScHARR | The University of Sheffield] for additional important guidance before beginning research data management planning.

What are FAIR Data and Software?

The FAIR data principles aim to make data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable, while making them 'as open as possible, as closed as necessary'. In the context of research software, the principles have been developed into FAIR4RS (FAIR for Research Software).

Findable
The first step in (re)using data/software is to find them. Data and software - and the metadata that describe them - should be easy for both humans and machines to find.

Accessible
Once the user finds details of data/software that may be helpful to them, it should be clear if / how they can be accessed.

Interoperable
Where data/software can be made available, it should be possible to use them with commonly used applications and in conjunction with other data.

Reusable
Metadata and data/software should be well-described so that they can be replicated and/or it is clear if and how they can be used in different settings.

FAIR in the context of health research

Health data can be personal, sensitive, and difficult to anonymise, and participant-level data is considered too sensitive to share without additional controls outlined in a data sharing agreement (DSA). Even where these concerns do not apply, there may be important ethical and data privacy reasons why you may wish to control or restrict other researchers' access to research data.

- Contexts within which it is unlikely to be possible to share data and software include the following:
    - Data may be drawn from an **existing context/dataset** (e.g. NHS patient data) where the data owner does not allow any further sharing of the data, either in its original form or when processed, linked, redacted or deidentified.
    - **Primary (including participant-level) data** may be impossible to deidentify to a satisfactory degree, or doing so may substantially impact upon its usefulness.

- While sharing raw data openly without controls may be impossible in the context of many ScHARR research projects, an understanding of FAIR principles and practices can:
    - Help all researchers involved in the study, during the study and after, to store and access material in a **consistent, effective and straightforward** way.
    - Where possible and appropriate, enable at least some of the data and software that support publications to be shared in order to **validate and authenticate** findings.
    - Where it is not possible or appropriate to share data and software, enable other materials (e.g. methodological materials - see below) to be shared in order to **benefit other researchers and validate the study**.

'As open as possible, as closed as necessary': FAIR vs Open data

Given concerns around the sharing of health data, it is important to remember that **FAIR data is not the same as open data**; rather, it is a set of practices that enable good data management and help to open up aspects of your research to others in ways (and to degrees) that are appropriate to the data concerned. If it's not appropriate to share, you might consider taking one or more of the following steps, all of which can incorporate one or more aspects of the FAIR principles:

- **Sharing analysed, sample or dummy data**
  You could share analysed or sample data in a repository if your dataset contains identifying information. Alternatively, dummy datasets for quantitative data or NVivo codebooks for qualitative data can allow you to be transparent while protecting your participants.

- **Sharing methodological resources from your study, but not data**
  Materials you could share, and which would benefit other researchers, might include analysis code, blank questionnaires, database specifications, protocols, and case report forms.

- **Using an embargo and/or making data available upon request only**
  When using ORDA and other repositories, you can apply an embargo, either to the entire content of the record, or just to uploaded data. This can be applied for a set time period, or to an end date. If appropriate, you can provide contact details to request access to data stored under embargo in a repository or other suitable storage (e.g. University X: drive). If using this option, consider what ethics and governance requirements you will need to implement for those requesting access.

- **Restricting access to sensitive data**
  It might be appropriate to restrict access to data in a repository. This may be especially useful where you have qualitative transcripts or issues of commercial sensitivity.

- **Metadata-only record**
  For highly sensitive data, the most appropriate option may be a metadata -only record which gives details of your dataset without making any other aspects of it available.

If you have any questions or concerns about sharing data, you can seek advice from the Library's Research Data Management team (rdm@sheffield.ac.uk).

With these provisos, the following information may be useful in increasing the 'FAIRness' of your research data and software:

Making your data and software FAIR

Before your project starts, you should create a Data Management Plan, to include all of the below information to help you to organise and store your data securely. DMPOnline provides templates, guidance and Library feedback to support this process.

Data Management Plan checklist

Data Storage	
Digital data should be stored securely using either:	
- University research data storage	
- University Google Shared Drive (for data **not involving participants** only)	
Have you got a logical naming structure in place for your project filestore? *This is good data management practice in enabling the project team and potentially others to locate and understand a file's relationship to the filestore as a whole, including consistent file versioning. You can find guidance here.*	
Have you created a README file which includes:	
- File structure *Details of the hierarchical structure by which project files are organised.*	
- Methodology *An overview of the study's methodology, including details of how the data have been collected, cleaned, processed and analysed.*	
- Any other information that can help you and others understand the data *Readme files enable all researchers on the project (at the time of the study or a later date) to gain an overview of the data. They are also useful to other researchers in instances where it is possible to share the dataset. You can find out more information about readme files here and here.*	
Use research data storage to ensure files are backed-up appropriately, following IT Services' guidance for research data storage.	
Data Collection Plans	
What will be the format of the data collected? *E.g. physical or digital, and in what file format?*	
Will you digitalise any physical data? If so, how and when?	
Identify any non-uniform variables across datasets you plan to merge.	
Include a plan for post-study deletion plans in your DMP.	

Include a clear missing data strategy.	
Who will have access to the data?	
Data archiving and/or sharing plans	
Where do you intend to store your data after the study, and to whom (if applicable) will it be made available? (see also below)	
Ensure you have created sufficient documentation, including (where appropriate) a data dictionary.	
Where applicable, discuss options for data sharing with external partners.	

Before you start collecting or analysing your data…

Make sure participants or providers of any data you collect (or are using) have consented to long-term storage and sharing of data, if you plan to do this.

Collecting your own data	
Have you completed a Data Protection Impact Assessment (DPIA) for your project? *(Applicable for research projects that may impact the privacy of individuals and / or involve the use of personal data: GDPR and data protection | Governance and Management | The University of Sheffield)*	
Have the participants consented to long-term data storage?	
Where applicable, have the participants consented to sharing their responses with other researchers (with an appropriate degree of anonymisation if necessary)?	
Have you put measures in place to ensure that privacy and confidentiality will be maintained?	
Will the data include sensitive and personal data?	
Have you got ethical approval for collecting, storing and, where applicable, sharing the data?	
Reusing other health data	
Have you checked to see if the data you plan to use are easily downloadable/accessible?	
Have you considered whether you need funding or training to access the dataset you plan on using?	
Are your intended uses of the data in line with the licence/s under which you are accessing them?	
If you are using international data, consider data privacy laws and protection of both the UK and the country of origin, including accessing the data appropriately and securely.	
Do you have permission to share or link to the data (in its processed or redacted form if	

necessary)?	

At the end of your research…

Ensure data and/or metadata is stored in a repository (e.g. a discipline-specific repository or the University repository, ORDA) and given a DOI. *A searchable directory of available research data repositories can be found at re3Data.*	
Where it is possible to make data available for sharing, consider what data and scale are most appropriate and potentially useful. *Any participant level data should be minimised, i.e. all direct and indirect identifiers, and free text fields removed.* *See also the ONS best practice for applying disclosure control to health statistics: Health statistics - Office for National Statistics (ons.gov.uk) and the Handbook on Statistical Disclosure Control for Outputs*	
Select an appropriate licence or conditions for reuse of your data and software. Make any delay to the release of data as short as possible and within funder requirements. *Use a Creative Commons licence for data and a software-specific licence (e.g. MIT or Apache) for code, selecting the most appropriate levels of restriction. Participant level data may only be shared using the most restrictive CC BY-NC-ND licence.*	
Include a data availability statement in publications emerging from the study, giving a link to stored data/metadata using the DOI provided by the repository.	
Where possible, share the software and code created to process data, or details of proprietary software used.	
Import code from Github to your chosen repository. *This means that a copy of the code at the time of the study's completion is archived and given a DOI. If you are using Github then take care not to accidentally upload sensitive data at the same time, especially if you are storing your code and data in the same folder. Github is hosted externally and it is not possible to guarantee that any sensitive data uploaded accidentally has not been accessed.*	
Store your data for a minimum of 10 years in line with the University's data retention schedule, or in line with funder/data provider requirements.	
Where you have used data from existing datasets that are publicly available on a long term basis, provide a link to the data rather than sharing it again.	

Helpful Resources

NIHR Open Research - see especially 2.1.1 (Spreadsheet data), which can help you to increase the accessibility and reusability of tabulated data.

Guidelines for implementing FAIR open data policy in health research

Uploading items to ORDA (Useful tips)