THE FAIR Data and Software Project Checklist for the School of Bioscience*

*(*This FAIR checklist is prepared for the School of Bioscience; however, the resources are specialised for the Molecular Microbiology: Biochemistry to Disease (MMBD); Development, Regeneration and Neurophysiology (DRN); Molecular and Cellular Biology (MCB) clusters)*

The main purpose of the project is to introduce FAIR data and software principles to the School of Biosciences. The FAIR (Findable, Accessible, Interoperable, Reusable) data does not only aim to raise awareness of data sharing, but also aims to gain good data management practices. Therefore, it is a useful guidance explaining the principles of the FAIR data as well as good data management practices. It is prepared aiming for a wider audience so that everyone from PhD students to established researchers can benefit. There are checklists for FAIR data and software that you can quickly apply for your data. You can find subject specific resources to deposit your data for MMBD, DRN, MCB clusters. The detailed information for each part and the links for additional reading are given in the sections.

Author: Deniz Simsek

1. Introduction

1.1. What/ Why is FAIR Data?

The '*FAIR Guiding Principles for scientific data management and stewardship'* were published in 2016, providing guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of research data. The principles of the FAIR data provide several benefits for data producers, publishers and science funders, leading to faster knowledge discovery and innovation.

Benefits of The FAIR Data:

- Promote good data management, which is very important for knowledge discovery, innovation, and the re-use of data
- Provide proper collection, annotation, and archiving of data
- Long-term care and preservation of data
- Unlock productivity through data sharing
- Increase reproducibility
- Leading to greater collaborations and more efficient workflows
- Researchers gain credit for their data as the FAIR concept has been increasing among various institutes
- Enable 'Machine actionable' data to reduce the costs and risks of data discovery

Sources and relative links:

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

A Guide to the FAIR Principles in Biopharma:
https://frontlinegenomics.com/a-guide-to-the-fair-principles-in-biopharma

How Springer Nature is supporting FAIR data: Khodiyar, Varsha; Laine, Heidi; O'Brien, David; Rodriguez-Esteban, Raul; Turkyilmaz-van der Velden, Yasemin; Baynes, Grace; et al. (2021): Research Data: The Future of FAIR White paper. figshare. Journal contribution.
https://doi.org/10.6084/m9.figshare.14393552.v1

1.2. The Four Basics of FAIR: A Summary of The FAIR Process

'Findable': Your data should be discoverable with metadata, identifiable, and locatable by means of a standard identification mechanism. Make your data **findable** by ensuring it:

- Has a persistent identifier
- Has rich metadata
- Is searchable and discoverable online

'Accessible': Your data should be always available and obtainable; even if the data is restricted or not available any more, the metadata should be open. Make your data **accessible** by ensuring it:

- Is retrievable online by using standardised protocols

- Has restrictions if necessary

'Interoperable': Your data should have the standard format to be reanalysing and reusing by other researchers. It should have the properties to be integrated with other data, applications, and workflows. This allows data exchange and reuse between researchers, institutions, organisations, or countries. Make your data **interoperability** by using:

- Common formats and standards, keywords
- Controlled vocabularies, ontologies

'Reusable': You should sufficiently describe and share your data with minimal restrictive licences that can allow broader reusing of it. Make your data **reusable** by ensuring it:

- Is well-documented to ensure that your data can be interpreted correctly by other researchers: create a README file.
- Has clear licence information that informs others to the limitation of reusing your data

Sources and relative links:

Guides for researchers: How to make your data FAIR:
https://www.openaire.eu/how-to-make-your-data-fair

FAIR Cookbook: https://faircookbook.elixir-europe.org/content/home.html

FAIR principles: https://www.go-fair.org/fair-principles/

Brief guide for Biomedical Data Producers, Stewards, and Funders to make data FAIR, Top 10 FAIR Data & Software Things:
https://librarycarpentry.org/Top-10-FAIR//2018/12/01/biomedical-data-producers/

''A love letter to your future self": What scientists need to know about FAIR data:
https://www.natureindex.com/news-blog/what-scientists-need-to-know-about-fair-data

1.3. FAIR Checklist Data and Software

The FAIR Data Checklist	
Findable	
❑ You have a rich metadata	*3.1. Organising and Documentation*
❑ The metadata is available online in a searchable repository	*4.4. Where and How to Share?*
❑ Your data has a persistent identifier	*4.4. Where and How to Share?*
❑ The metadata includes clearly and specifically the persistent identifier	*3.1. Organising and Documentation*
Accessible	
❑ The persistent identifier directly takes you to data and/or metadata	*4.4. Where and How to Share?*
❑ Your data is retrievable online by using standardised protocols	*4.4. Where and How to Share?*
❑ If necessary, the access procedure has authorisation and authentication steps	
❑ Metadata is always available even if the data is restricted or not shared	*4.2. What to Share?*
Interoperable	
❑ Your data is clearly documented to be commonly understood	*3.1. Organising and Documentation*
❑ Your data and metadata conform to common formats and standards	*2.2. Plan and Design Experiments; 3.1. Organising and Documentation; 4.3. Preparing Data to Share*
❑ You used ontologies and controlled vocabularies to define your metadata where possible	*3.1. Organising and Documentation*
Reusable	
❑ Your data is well documented by a README file text that allows others to understand it	*3.1. Organising and Documentation*
❑ Your data and metadata follow relevant standards	*3.1. Organising and Documentation*
❑ Your data has a clear and accessible licence that informs others for data reuse data	*4.3. Preparing Data to Share*

The FAIR Software Checklist	
Findable	
❏ You have a metadata (codemeta) describes your software	*6.1. What is FAIR for Software? Translating FAIR Principles to A Software*
❏ You describe your software with relevant ontologies	*6.1. What is FAIR for Software? Translating FAIR Principles to A Software*
❏ Your software is registered in a software registry	*6.1. What is FAIR for Software? Translating FAIR Principles to A Software*
❏ You get a unique persistent identifier for your software	*6.1. What is FAIR for Software? Translating FAIR Principles to A Software*
Accessible	
❏ The persistent identifier directly takes you to software and metadata	*4.4. Where and How to Share?*
❏ Your software can be downloaded	*6.1. What is FAIR for Software? Translating FAIR Principles to A Software*
❏ If necessary, the access procedure has authorisation and authentication steps	
Interoperable	
❏ You explained the functionality of your software by using ontologies	*6.1. What is FAIR for Software? Translating FAIR Principles to A Software*
❏ You use standard formats for inputs and outputs	*6.1. What is FAIR for Software? Translating FAIR Principles to A Software*
Reusable	
❏ You properly documented your software. You have a The Docs page that describes how to install, run and use your software	*6.1. What is FAIR for Software? Translating FAIR Principles to A Software*
❏ You state how to cite your software	*6.1. What is FAIR for Software? Translating FAIR Principles to A Software*
❏ Your software has a clear and accessible licence that informs others for reuse	*6.1. What is FAIR for Software? Translating FAIR Principles to A Software*

2. Data Creation

2.1. Research Data

Data is the outcome ‘you use and produce during your research life cycle’ (The Turing Way). It covers datasets, software, code, workflow, models, figures, tables, images and videos, interviews, articles. The data can be both digital and non-digital.

Research data often has a life cycle which starts from establishing a research plan, following data creation, analysing, sharing, archiving and reusing data. The overview schematic representation of the research data life cycle is given below.

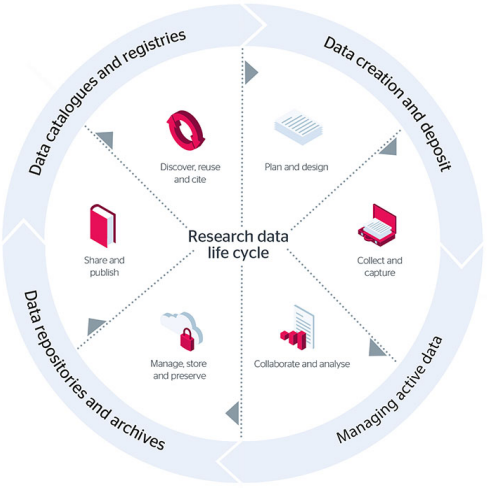

Figure1. The overview diagram of the research data life cycle. (The image is taken from Kaye et. al. 2017)

Sources:

Kaye, J., Bruce, R. and Fripp, D., 2017. *Establishing a shared research data service for UK universities*. Insights, 30(1), pp.59–70. DOI: http://doi.org/10.1629/uksg.346

The guide for reproducible research, The Turing Way: https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-data.html

2.2. Plan and Design Experiments

At the start of your project, you should make some decisions to manage your data during the project. Therefore, start developing suitable procedures before data collection. That will help to ensure consistency and quality during data collection.

1. Create your data management plan (DMP):
    - Use DMPonline platform to create DMP: https://dmponline.dcc.ac.uk/
    - Contact with the University of Sheffield library for further information: rdm@sheffield.ac.uk

2. Consider ethics policy:
    - Think about the ethics policy if your research involves human and animal subjects, and make sure that consent includes your plans for long-term storage and sharing of data.
3. Choose appropriate file formats:
    - Consider which file formats you will use to store your data
    - Use one file format for data collection and analysis.
    - Convert your data to another format for archiving when the project is finished.
    - Use standard and interchangeable data formats for long term use and storage.
    - Common best file formats:
        - Textual data: XML, TXT, HTML, PDF/A (Archival PDF)
        - Tabular data (including spreadsheets): CSV
        - Databases: XML, CSV
        - Images: TIFF, PNG, JPEG
        - Audio: FLAC, WAV, MP3

Sources and relative links:

Data management Plan (DMP):

The Digital Curation Centre: Example DMPs and guidance:
https://www.dcc.ac.uk/resources/data-management-plans/guidance-examples

Data Management Planning, The University of Sheffield Library website:
https://www.sheffield.ac.uk/library/rdm/dmp

Guide for reproducible research: https://the-turing-way.netlify.app/reproducible-research/rdm

Data Management Planning: https://dataoneorg.github.io/Education/lessons/03_planning/index.html

Foster Open Science, Open Resources for Research Data Management:
https://www.fosteropenscience.eu/taxonomy/term/143

Research Data Management Toolkit: https://www.jisc.ac.uk/guides/rdm-toolkit

An online guide containing good data management practices applicable to research projects from beginning to end. The ELIXIR Research Data Management Kit (RDMkit):
https://rdmkit.elixir-europe.org/index

Guides for Researchers: How to create a Data Management Plan for H2020 projects:
https://www.openaire.eu/how-to-create-a-data-management-plan

Guides for Researchers: How to identify and assess Research Data Management (RDM) costs:
https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs

Ethics Policy:

Ethics Policy, The University of Sheffield website:
https://www.sheffield.ac.uk/research-services/ethics-integrity/policy

The Royal Society, Research Ethics: https://royalsociety.org/journals/ethics-policies/research-ethics/

Good practice: Research and Data Ethics: https://www.fosteropenscience.eu/node/2822

File Formats:

Choosing Formats, Research Data Management:
https://www.data.cam.ac.uk/data-management-guide/creating-your-data/choosing-formats

Recommended Formats UK Data Service:
https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/

The Library of Congress, Recommended Formats Statement:
https://www.loc.gov/preservation/resources/rfs/TOC.html

Guides for Researchers: Data formats for preservation:
https://www.openaire.eu/data-formats-preservation-guide

2.3. Data Collection and Analyse

1. Collect your data according to your experimental method.
2. If you collect or create physical data, digitise them if possible.
    - Non-digital data examples: handwritten laboratory notebooks, journals, surveys, cell and tissue samples.
    - Why digitise data?: Increased efficiency, ease of access, long term preservation, data recovery.
    - Cell and tissue specimens may be adhered to glass slides. There are methods allowing them to 3D scan and convert into 3D digital objects which can then be virtually sectioned.
3. Analyse and interpret your data.
4. Store your both raw and analysed data securely.
    - Preferably in University research data storage or University Google drive: https://students.sheffield.ac.uk/it-services/research/storage

Sources and relative links:

How to deal with non-digital data? https://www.openaire.eu/non-digital-data-guide

3. Organise and Store Data

3.1. Organising and Documentation

1. Organise your data properly that enables you and others to find and use them easily.
    - Create your files and folders with consistent and meaningful names.
    - Think about vocabulary, punctuation, order, numbers and dates when you are naming the files and folders.
    - Check existing procedures in your team/ department and adapt to them.
    - When you make changes on the files, use a revision numbering system. For example, you can demonstrate major changes to a file as v01 (first version), v02 (second version). Minor changes can be shown with increasing decimal numbers as v01_01 indicates that minor changes were made to the first version.
    - Separate folders for ongoing and completed work.

- Ensure that your files are backed up.

2. Write a documentation including all the necessary information to properly understand, interpret and reuse the given data later by yourself and others.

- This documentation is often in the form of a plain text file called a **README file,** and has been stored with the data files together.
- **A README plain text file** should contain the following information:
    - Basic description of the research
    - Inventory of files and the relationship between them
    - For each filename: a short description of what data it includes, describing relationships between the tables, figures or sections.
    - For tabular data: definitions of column headings and row labels, data codes, and measurement units
    - Methodologies, protocols, sampling techniques, equipment used with settings and calibrations, and any data processing steps
    - Software, code, and algorithms
    - Classification systems and abbreviations
    - Details of third-party data, whom to contact with questions

3. Create a **metadata**, which gives clear details about your data, when you are depositing your data in a repository.

- What is metadata? “documentation about the data that describes the content, quality, condition, and other characteristics of a dataset. More importantly, metadata allows data to be discovered, accessed, and reused” - (DataONE Education module)
- Metadata has two major components (Soranno, 2019):
    - Dataset origin and context: This part includes details on the study itself containing the keywords associated with the dataset, the authors, organisations, funders, timeframe, organisms, location of the study, and the study design and methods. This part allows the others to discover your data.
    - Data dictionary: It is a collection of specific variables in the dataset. This part provides detailed information for future researchers to understand, interpret, and use the data in your dataset. This information includes date and time formats, unique identifiers, definitions of variables, units of measurement, missing data codes, and other factors.
- The difference between ‘README file’ and ‘metadata’:
    - **A README file text** includes information within a dataset that enables data to be understood and reused **(I: interoperability; R: reusable).**
    - **Metadata** includes details about a dataset in a repository record that enables data to be findable and accessible by others (**F: Findable; A: Accessible; R: reusable).**

- How to create a metadata?
    - Some repositories allow you to create metadata automatically when you deposit your data.
    - However, if you need to create a metadata yourself, use a metadata standard in your research area or general metadata schemes (*metadata standard list is given in section 5.2*).
    - Metadata needs to be standardised in order to be useful.
    - To standardise your metadata, define a semantic data model that will help make your data searchable and findable, so more shareable in the same discipline:
        - Choose community- and domain-specific **ontologies**, along with **controlled vocabularies** and **keywords** to describe the dataset entities.
        - **Ontologies** are the structural frameworks for organising information by extracting relevant data. They are used in artificial intelligence, software engineering, and library science. They are helpful when organising data.
        - **Controlled vocabularies** an organised arrangement of words to retrieve relevant data. A controlled vocabulary enables a consistent way to describe data. Using them will enhance the findability of data.
    - Make the metadata machine-readable so that it can transform the information into a file format (.XML) that is easily readable and interpretable by computers.
    - Store the metadata file always with your dataset in the same repository.

Sources and relative links:

Organising:

Useful source about organising your data from The University of Cambridge Library website:
https://www.data.cam.ac.uk/data-management-guide/organising-your-data#Naming

UK Data Service: Versioning
https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/versioning/

Documentation and metadata:

README file:

Great Learning, README File- Everything you Need to Know:
https://www.mygreatlearning.com/blog/readme-file/

Useful websites helping to create documentation:
https://www.makeareadme.com/; https://readme.com/documentation

GitHub, about READMEs:
https://docs.github.com/en/repositories/managing-your-repositorys-settings-and-features/customizing-your-repository/about-readmes

Metadata:

Metadata for Data Management: A Tutorial: Controlled Vocabularies: https://guides.lib.unc.edu/metadata/controlled-vocab

National Microbiome Data Collaborative: Introduction to metadata and ontologies: https://microbiomedata.org/introduction-to-metadata-and-ontologies/

A Semantic data model: https://www.gooddata.com/blog/what-a-semantic-data-model/

What is an ontology and why do we need it?: https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

Controlled vocabulary: https://www.jisc.ac.uk/guides/metadata/controlled-vocabulary

Ontology Development 101: A Guide to Creating Your First Ontology: https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

Soranno, P.A. (2019), *Six Simple Steps to Share Your Data When Publishing Research Articles*. Limnology and Oceanography Bulletin, 28: 41-44. https://doi.org/10.1002/lob.10303

Good practice: BOLSTM classifying relations via long short term memory networks along biomedical ontologies: https://www.fosteropenscience.eu/node/2287

Guide to writing 'readme' style metadata, Cornell University, Research Data Management Service Group: https://data.research.cornell.edu/content/readme

3.2. Store and Archive Data

Another important step after creating and organising your data is to carefully store them in a safe place with backups.

1. Storage: Decide what you need to keep and store securely, ideally in University research data storage or University Google drive:

https://students.sheffield.ac.uk/it-services/research/storage

2. Backups: Good backup practices are essential to prevent data loss. Therefore, make sure you are aware of the following tips to backup your data properly.

- Make sure the computer networks backup files automatically.
- Find out how often backups happen and how long they're kept.
- Make 2 or 3 backups of all important data which are not stored on a networked file server.
- Keep one backup in a different place from others, and use multiple different types of storage media.

3. Selection: You have to decide what to keep because storage costs money and requires effort. Storing large amounts of data makes it difficult to find and access the important files you really need.

4. Archiving: Identify data that is no longer actively needed and move it to long-term storage systems.

Sources and relative links:

Useful source about looking after your data from The University of Cambridge website: https://www.data.cam.ac.uk/data-management-guide/looking-after-your-data

Good source about deciding what to keep and what to delete: How to Appraise and Select Research Data for Curation: https://www.dcc.ac.uk/guidance/how-guides/appraise-select-data

Five steps to decide what data to keep:
https://www.dcc.ac.uk/guidance/how-guides/five-steps-decide-what-data-keep

More information about data archiving:
https://www.druva.com/glossary/what-is-data-archiving-definition-and-related-faqs/

Information about data archiving from Radboud University:
https://www.ru.nl/rdm/archiving-data/what-data-should-archived/

Guides for Researchers: Raw data, backup and versioning:
https://www.openaire.eu/RAW-DATA-BACKUP-AND-VERSIONING

4. Share Data

4.1. Why is Sharing Data Important?

Research data is a valuable resource as it requires a large amount of time, money and effort. Therefore, sharing data, enabling it to be understood and reused by other researchers, is very important for both scientific development and individual researchers.

At the end of your research, where possible, share data that validates your research, especially if it has potential for reuse.

Benefits of sharing research data (UK Data archive):

- encourages scientific investigations
- promotes innovation
- leads to new collaborations between data users and data creators
- maximises transparency
- provides important resources for education and training
- encourages the improvement research methods
- reduces the cost of duplicating data collection
- increases the impact and visibility of research
- promotes the research that created the data and its outcomes

Sources and relative links:

UK Data Archive, Managing and Sharing Data:
https://dam.ukdataservice.ac.uk/media/622417/managingsharing.pdf

Good practice: Understanding Data Sharing:
https://dataoneorg.github.io/Education/lessons/02_datasharing/index

Good practice: *FOSTER, Managing and Sharing Research Data*:
https://www.fosteropenscience.eu/node/2328

A Guide to the FAIR Principles in Biopharma:
https://frontlinegenomics.com/a-guide-to-the-fair-principles-in-biopharma/

Five selfish reasons to work reproducibly:
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0850-7#Sec11

4.2. What to Share?

You should decide what data to share and to what extent. Remember that not all data has to be made open. Data can be restricted and still be FAIR. *The FAIR aims to make your data as open as possible and as closed as necessary.*

As a minimum, the raw data is required to recreate your findings. However, there are some situations where you cannot share your raw/all data openly because of the **personally and commercially sensitive data.** In these cases, you can only share metadata so that others can understand what you did in your research.

- Examples of sensitive data:
    - *Personal data:* names or identification numbers, physical, physiological, genetic, mental, economic, cultural, or social characteristics.
    - *Confidential data***:** trade secrets, investigations, data protected by intellectual property rights.
    - *Biological data:* endangered plant or animal species, where their survival is dependent on the protection of their location data
- Some ways to prepare sensitive data for storage and sharing:
    - **Anonymization:** It irreversibly prevents any way of identifying the data subject. A tool to anonymise data: *Amnesia*, https://amnesia.openaire.eu/
    - 'Anonymisation of personal research data is the only effective solution to comply with both data protection legislation and the requirements of the Open Research Data Pilot.' (OpenAIRE Horizon 2020 Fact Sheets)
    - **Pseudonymization:** Additional information is required to re-identify the data subject. It allows data to be traced back to its origins.
    - **Encryption:** It will make your data completely meaningless to those who may try to access.
    - Another option is to archive data under a closed licence in a trustworthy repository such as Institutional data archives.

Observe all terms of participant consent regarding data sharing. Check if you are permitted to share third-party data, or data derived from them. Don't assume you can share data just because they are available online.

Sources and relative links:

Useful source about sharing your data from The University of Cambridge:
https://www.data.cam.ac.uk/data-management-guide/sharing-your-data

Good practice: Data Protection and Ethics: https://www.fosteropenscience.eu/node/2330

Guides for Researchers - How to deal with sensitive data:
https://www.openaire.eu/sensitive-data-guide

An implementation story making data FAIR but not open:
https://zenodo.org/record/6302908#.Yoo9cKjTW5c

4.3. Preparing Data to Share

1. Transfer your data to an open or more widely accessible format if they are in a specialised / proprietary format.
    - Using standard and open data formats ensures longer-term usability of data. For more information visit: https://ukdataservice.ac.uk/learning-hub/research-data-management/#format-your-data
2. Select an appropriate **licence or conditions** for reuse of your data and software.
    - A licence tells people how they can use your data, and options often include the Creative Commons licences.
    - Information about CC licences: https://creativecommons.org/about/cclicenses/
    - How to choose a right licence: https://creativecommons.org/choose/#
    - You should be aware that the licence cannot be revoked and you must own/control the work.
3. Place a **data availability statement** in theses and other publications.
    - It should be written in the manuscript text itself ideally after the title and affiliations.
    - The statement should be short and include the information where the data and metadata are available with DOI, or contact details for access request. (e.g. a DOI for the dataset in a repository)
4. Check whether individual repositories charge costs for depositing data, and if these will be covered by your funding.
5. Discuss options for data sharing with external research partners.

Sources and relative links:

Before applying a licence on your work: Considerations for licensors and licensees: https://wiki.creativecommons.org/wiki/Considerations_for_licensors_and_licensees

How to Licence Research Data: https://www.dcc.ac.uk/guidance/how-guides/license-research-data

A tool help you choose an appropriate licence: https://ufal.github.io/public-license-selector

Sharing Research Data, The University of Sheffield Library: https://www.sheffield.ac.uk/library/rdm/publish

4.4. Where and How to Share?

1. Share data openly through a **repository** that provides your dataset with a DOI e.g. ORDA, or a subject-specific repository. (*repository and database lists are given in section 5.1*).
    - How to choose a repository? You should look for a repository that does the following:
        - Stores the data safely
        - Assigns a persistent identifier that makes the data findable
        - Describes the data appropriately (metadata)

        - Adds licence information
    - What is a persistent identifier?
        - Persistent identifier is a long lasting reference to a digital source.
        - DOIs, URLs, PURLs are some examples of persistent identifier models.
        - Identifiers allow you to cite other people's work and be cited.
        - DOI (Digital Objective Identifier) is a permanent identifier that always points directly to your data, even if the actual location of your data changes or goes offline.
        - The main purpose of the DOI is to encourage sharing and citation.
        - The DOI system has many applications such as Crossref and DataCite.
            - Crossref to manage citations in scholarly publications (https://www.crossref.org/)
            - DataCite: to help you locate, identify and cite data (https://datacite.org/index.html)

2. Search for an appropriate subject-specific repository for your research data. Use the tools below to determine this and sign up to share.

- Re3data: https://www.re3data.org/
- FAIRsharing: https://fairsharing.org/

3. If using a subject-specific repository is not possible, share your data in a general repository. Visit the links given below for general repository comparison.

- *The General Repository Comparison Chart and FAIRsharing Collection:*
    - https://zenodo.org/record/3946720#.YsSbIXbTW5d
    - https://fairsharing.org/GeneralRepositoryComparison

4. If you have used data in your research that is publicly and permanently available, share a link rather than sharing the actual data.

5. Store and share data for a minimum of 10 years after the end of the project, or in line with funder requirements.

Sources and relative links:

Good practice: *A demonstration of searching for research data repositories using the Re3data directory:* https://www.fosteropenscience.eu/content/re3data-demo

Useful article showing how to share data: *Six Simple Steps to Share Your Data When Publishing Research Articles*: https://aslopubs.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/lob.10303

Identifiers, Jisc: https://www.jisc.ac.uk/guides/rdm-toolkit/identifiers

Ten quick tips for sharing open genomic data: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006472

Guides for Researchers, How to find a trustworthy repository for your data?: https://www.openaire.eu/find-trustworthy-data-repository

A Guide to the FAIR Principles in Biopharma: https://frontlinegenomics.com/a-guide-to-the-fair-principles-in-biopharma/

The University of Sydney, Library resources: Data publication: Step 6 - Get a persistent identifier https://libguides.library.usyd.edu.au/datapublication/step6

5. Resources

5.1. Repositories and Databases

5.1.1. General-purpose Repositories and Databases

General-purpose data repositories accept a wide range of data types in a wide variety of formats. They don't attempt to integrate the deposited data, so usually resulting in a hugely diverse repository.

- **Data Archiving and Networking Services (DANS)**: http://www.dans.knaw.nl/; https://dans.knaw.nl/en/data-services/easy/; is a repository for research data that allows long term and secure archiving. Assigns a DOI, and certified with Core Trust Seal.
- **DataDryad**: https://datadryad.org/stash; provides integration between repositories and journal publishers, which is making it easier for journal publishers to facilitate better data deposit.
- **DataHub**: http://datahub.io; a repository provides a fast way for individuals, teams and organisations to publish, deploy and share their data.
- **Dataverse7**: https://dataverse.org/; assigns DOI, deposits include metadata, data files, and any complementary files. Metadata is always public, even if the data are restricted or removed for privacy issues. Provides machine-accessible interfaces to search the data, access the metadata and download the data files.
- **FigShare**: http://figshare.com; is a repository that enables users to deposit their data, papers, code, media and other research outputs in a citable, shareable and discoverable way.
- **NIH FigShare**: https://nih.figshare.com/; a repository to make datasets resulting from NIH funded research more accessible, citable, shareable, and discoverable.
- **Mendeley Data**: https://data.mendeley.com/; is a secure cloud-based repository where you can store your data, ensuring it is easy to share, access and cite.
- **Open Science Framework (OSF)**: https://osf.io/; provides a system for organising scientific projects, including data, code, and protocols. It simply makes your project publicly available.
- **ORDA**: https://orda.shef.ac.uk/; The University of Sheffield's data repository. It is provided by figshare and enables university research data to be preserved, discovered, and accessed.
- **Zenodo**: http://zenodo.org/; allows for deposition of data, code, analysis, and manuscripts and has semantic versioning, assigns persistent identifiers.

5.1.2. Subject-specific Repositories and Databases:

5.1.2.1. Genomics:

- **Gene Expression Omnibus (GEO)**: https://www.ncbi.nlm.nih.gov/geo/; data house for quantitative gene expression, gene regulation, and epigenomic data, including data from RNA-seq, ChIP-seq, Hi-C, bisulfite sequencing, and microarrays.
    - File formats: CRAM, BAM, SFF, HDF5, FASTQ, bedGraph, bigBed, WIG, bigWig, general feature format (GFF), gene transfer format (GTF), GEOarchive

- **Sequence Read Archive (SRA)**: https://www.ncbi.nlm.nih.gov/sra/; you can deposit high-throughput sequencing reads that do not fit into GEO.
    - File formats: CRAM, BAM, SFF, HDF5, FASTQ
- **Genbank**: https://www.ncbi.nlm.nih.gov/genbank/: deposit DNA and RNA sequence data, which contains sequence of genomic DNA, mRNA, noncoding RNA, plasmids, and synthetic constructs.
    - File format: FASTA
- **The European Genome-phenome Archive (EGA)**: https://ega-archive.org/; deposits sensitive genetic and phenotypic information from human participants, allows controlled access to the data upon request.
    - File formats: CRAM, BAM, FASTQ, VCF, SFF, HDF5
- **International Nucleotide Sequence Database Collaboration (INSDC):** https://www.insdc.org/; includes GenBank, SRA, DNA Data Bank of Japan (DDBJ), and European Nucleotide Archive (ENA) repositories. The INSDC members can use data submitted to any of these repositories and automatically make it available in the others.

5.1.2.2. Proteomics:

- **Proteomics Identifications Database (PRIDE)**: https://www.ebi.ac.uk/pride/; deposits mass spectrometry proteomic data.
- **Mass Spectrometry Interactive Virtual Environment (MassIVE)**: https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp; deposits mass spectrometry proteomic data and also stores reanalysed results of available datasets.
- **The ProteomeXchange consortium**: https://www.proteomexchange.org/; includes main proteomics data repositories such as PRIDE and MassIVE, enabling a centralised sharing system for mass spectrometry proteomics data.
- **PeptideAtlas SRMexperiment library (PASSEL)**: http://www.peptideatlas.org/passel/; deposits targeted selected reaction monitoring (SRM) proteomic data.
- **Panorama Public**: https://panoramaweb.org/; deposits targeted proteomic data analysed using Skyline.
- File Formats for proteomics data:
    - Raw data come in a proprietary vendor file format, such as .raw (Thermo Scientific, Waltham, MA, USA), .wiff (SCIEX, Framingham, MA, USA), or .d (Agilent, Santa Clara, CA, USA).
    - Share peak files in the standard mzML file format.
    - Provide identification data and quantification data in the standard mzTab format.

5.1.2.3. Structural Biology:

- **Protein Data Bank (PDB)**: https://www.rcsb.org/; a repository for atomic coordinates of nucleic acids, proteins, and larger assemblies.
    - File formats: PDB, mmCIF.
- **Electron Microscopy Data Bank (EMDB)**: https://www.ebi.ac.uk/emdb/; deposits 3D reconstructions from processed EM data.

    - File formats: MRC, CCP4.
- **EMPIAR**: https://www.ebi.ac.uk/empiar/; stores EM raw and processed image data.
    - File formats: TIFF, HDF5, MRC, MRCS, DM4, EER, IMAGIC, SPIDER, SCIPION, EMDB-SFF, AMIRA, STL, VTK, VTP, OBJ, AVI, JPEG, PNG, EMX, BLENDER, TXT.
- **Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC)**: https://www.proteindiffraction.org/; deposits X-ray diffraction raw data.
    - File formats: Raw diffraction data formats.
- **Biological Magnetic Resonance Bank (BMRB)**: https://bmrb.io/; a repository for NMR raw data. Additionally, you can deposit raw NMR data, in the form of restraints, in the NMR Restraints Grid (https://restraintsgrid.bmrb.io/).
    - File formats: CCPN, mmCIF, PDB, NMR-STAR, X-PLOR.
- **Worldwide Protein Data Bank:** http://www.wwpdb.org/; an organisation that maintains archives of macromolecular structure. It has the following members: EMBD, BMRB, Protein Data Bank Japan (PDBj), Protein Data Bank in Europe (PDBe), Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB-PDB).
  (For more information: https://en.wikipedia.org/wiki/Worldwide_Protein_Data_Bank)

5.1.2.4. Microscopy:

- **Image Data Resource (IDR)**: https://idr.openmicroscopy.org/; a house for large and complete benchmark microscopy image datasets associated with a publication.
    - File Format: Any Bio-Formats, OME-TIFF preferred.
- **Electron Microscopy Public Image Archive (EMPIAR)**: https://www.ebi.ac.uk/pdbe/emdb/empiar/; deposits electron microscopy image data, accepts high-resolution images from subcellular compartments and biological structures.
    - File Format: TIFF, HDF5, MRC, MRCS, DM4, IMAGIC, SPIDER, FEI
- **BioImage Archive**: https://www.ebi.ac.uk/bioimage-archive/; enables a home for all other microscopy image datasets, often those of a smaller size.
    - File Format: Any Bio-Formats, OME-TIFF preferred
- **Cell Image Library**: http://www.cellimagelibrary.org; a repository for a wide variety of biological images and movies for research and education purposes.
    - File Format: Any Bio-Formats, OME-TIFF preferred
- **The Systems Science of Biological Dynamics (SSBD)**: http://ssbd.qbic.riken.jp; database of a variety of images, even computationally simulated microscopy images. It allows analysis of experimental and computationally simulated biological image data.
    - File Format: Any Bio-Formats, OME-TIFF preferred

5.1.2.5. Other subject-specific repositories:

- **Universal Protein Resource (Uniprot)**: https://www.uniprot.org/; a protein sequence database, providing comprehensive resources for protein sequences and functional information. Assigns a stable URL, and deposits metadata, which is machine-readable. The format uses standardised vocabularies and ontologies and links to over 150 other databases.

- **Metabolights**: https://www.ebi.ac.uk/metabolights/; a database for metabolomics experiments and derived information.
- **Metabolomics Workbench**: https://www.metabolomicsworkbench.org/data/; a repository for metabolomics data and metadata providing analysis tools and access to various resources. NIH grantees may upload data and general users can search metabolomics databases.
- **Accelerating Medicine Partnership Type 2 Diabetes**: https://www.kp4cd.org/datasets/t2d; portal and database of DNA sequence, functional and epigenomic information, and clinical data from studies on type 2 diabetes and analytic tools to analyse these data.
- **Flow repositories:** http://flowrepository.org/; a database of flow cytometry experiments where users can query and download data collected and annotated according to the MIFlowCyt data standard. (The website is temporarily unavailable)
- **Drugbank database**: https://go.drugbank.com/; provides comprehensive molecular information about drugs, their mechanisms, interactions, and targets.
- **Open PHACTS**: http://www.openphacts.org/index.php; this is a data integration platform for drug discovery.
- **Brain Image Library (BIL):** https://www.brainimagelibrary.org/; enabling researchers to deposit, analyse, mine, share and interact with large brain image dataset.
- **GlycoPOST**: https://glycopost.glycosmos.org/; a mass spectrometry data repository for glycomics.
- **Springer Nature Scientific Data's recommendations:** https://figshare.com/articles/dataset/Scientific_Data_recommended_repositories_June_2015/1434640; these repositories were checked by Springer Nature to ensure they meet their requirements for data access, preservation and stability.
- **ELIXIR:** https://elixir-europe.org/; an intergovernmental organisation that brings together life science resources from across Europe. This organisation aims to make it easier for scientists to find and share data, exchange expertise, and agree on best practices.
    - *ELIXIR Core Data Resources*: https://elixir-europe.org/platforms/data/core-data-resources
    - *ELIXIR Deposition Databases for Biomolecular Data*: https://elixir-europe.org/platforms/data/elixir-deposition-databases; includes a list of resources that it recommends for the deposition of experimental data.
    - *ELIXIR Interoperability and resources*: The goal of the Interoperability Platform is to help people and machines discover, access, integrate and analyse biological data. (https://elixir-europe.org/platforms/interoperability/rirs)

Sources and Relative Links:

Sharing biological data: why, when, and how:
https://febs.onlinelibrary.wiley.com/doi/10.1002/1873-3468.14067

A Guide to the FAIR Principles in Biopharma:
https://frontlinegenomics.com/a-guide-to-the-fair-principles-in-biopharma/

Murphy F, Bar-Sinai M, Martone ME (2021) *A tool for assessing alignment of biomedical data repositories with open, FAIR, citation and trustworthy principles.* PLoS ONE 16(7): e0253538. https://doi.org/10.1371/journal.pone.0253538

Where should I deposit my data?, Jisc: https://www.jisc.ac.uk/guides/rdm-toolkit/where-should-i-deposit-my-data

5.2. Metadata Standards and Ontologies

Most biological disciplines have specific metadata standards that define the information expected to accompany datasets.

Standard Metadata Schemas:

- DataCite Metadata Schema**:** https://schema.datacite.org/
- Dublin Core Metadata Element Set: https://www.dublincore.org/specifications/dublin-core/dces/

Disciplinary Metadata Schemas:

- Disciplinary metadata, DCC: https://www.dcc.ac.uk/guidance/standards/metadata
- Proteomics metadata standard: https://github.com/bigbio/proteomics-metadata-standard
- Metadata Standards Directory: http://rd-alliance.github.io/metadata-directory/standards/
- Fairshairng.org, standards: https://fairsharing.org/search?fairsharingRegistry=Standard
- Defining Our Research Methodology (DORy): https://doryworkspace.org/; a catalogue of 3D microscopy standards.
- Brain image library, new metadata model: https://www.brainimagelibrary.org/newmetadatamodel.html
- REMBI: Recommended Metadata for Biological Images; enabling reuse of microscopy data in biology: https://www.nature.com/articles/s41592-021-01166-8
    - Metadata pdf template: :https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-021-01166-8/MediaObjects/41592_2021_1166_MOESM1_ESM.pdf

Ontology Resources:

- The Gene Ontology: http://geneontology.org/
- Uberon Multi-Species Anatomy Ontology: http://obophenotype.github.io/uberon/
- Semantic Web technologies enable people to create data stores on the web, build vocabularies, and write rules for handling data: Semantic web: https://www.w3.org/standards/semanticweb/

5.3. Other Useful Platforms

- **FAIRDOM:** https://fair-dom.org/about; a consortium for systems biology to support researchers, students, trainers, funders and publishers by enabling collaborative projects to make their data, documents, operating procedures and models, FAIR.

- **FOSTER portal:** https://www.fosteropenscience.eu/; is an e-learning platform that brings together the best training resources addressed to those who need to know more about Open Science. Some useful and relative courses and resources in The FOSTER portal:
    - Introduction to Open Science: https://www.fosteropenscience.eu/node/2076
    - Open and FAIR research data: https://www.fosteropenscience.eu/node/2820
    - Open access publishing: https://www.fosteropenscience.eu/node/2331
    - Ontogene entity recognition OGER: https://www.fosteropenscience.eu/node/2311
    - Text Mining Neuroscience Literature using the OpenMinTeD Platform: https://www.fosteropenscience.eu/node/2300
    - Florilege, a new database of habitats and phenotypes of food microbe flora: https://www.fosteropenscience.eu/node/2292
    - BOLSTM classifying relations via long short term memory networks along biomedical ontologies: https://www.fosteropenscience.eu/node/2287
- **FAIR Metrics group:** http://fairmetrics.org; enables subjective and self-assessments of FAIRness, and defining ways to measure FAIRness.
- **FAIRassist**: https://fairassist.org/#!/; offers stakeholders a personalised guidance to discover standards and repositories in FAIRsharing to make their data FAIR.

6. FAIR Software

At the policy level, software is seen as part of FAIR, with the European Commission expert group on FAIR data stating that “Central to the realisation of FAIR are FAIR Digital Objects, which may represent data, software or other research resources.” (Lamprecht et al., 2020) Applying the FAIR principles to research software will provide similar benefits, such as enabling transparency, reproducibility and reusability of research that facilitate efficient access to software-based knowledge by industry, science, education and society. In particular, FAIR software should facilitate the creation of FAIR data.

Therefore, you should share software and code created to process data, or details of proprietary software used.

6.1. What is FAIR for Software? Translating FAIR Principles to A Software

Findability:

1. Create a description of your software with metadata and ontologies.
    - Codemeta is a set of keywords used to describe software and how to structure them in a machine readable way (https://codemeta.github.io/terms/)
    - Edam is an example of an ontology that provides terminology that can be used to describe bioinformatics software. (http://edamontology.org/page)
2. Register your software in a software registry. Some software registries were given:
    - Biotools: https://bio.tools/
    - GitHub: https://github.com/research-software-directory/research-software-directory
    - Zenodo: https://zenodo.org/

3. Get and use a unique and persistent identifier for your software.

Accessibility:

Make sure that people can download your software, so deposit your software on a reliable platform (*repositories are given in section 6.2.*)

Interoperability:

1. Explain the functionality of your software.
    - Use terms from a domain ontology like EDAM, if possible

2. Use standard (community agreed) formats for inputs and outputs that allows data exchange between different pieces of software.
    - FASTA format is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences. Filename extensions: .fasta, .fna, .ffn, .faa, .frn, .fa.
    - NetCDF is a standard file format used for sharing of array-oriented scientific data.
    - Avoid defining your own standards.

Reusability:

1. Document your software.
    - Write 'The Docs' page that describes how to install, run and use your software.
    - A beginner's guide to writing documentation: https://www.writethedocs.org/guide/writing/beginners-guide-to-docs/
2. Give your software a licence.
    - Visit the links below for software specific licences:
        - Choose an open source licence: https://choosealicense.com/
        - Adopt a licence and comply with the licence of third-party dependencies: https://softdev4research.github.io/4OSS-lesson/03-use-license/index.html
3. State how to cite your software to get credit for your work. For further information visit the links:
    - How to cite and describe software: https://www.software.ac.uk/how-cite-software
    - Software Citation Principles: https://force11.org/info/software-citation-principles-published-2016/
    - What is a CITATION.cff file?: https://citation-file-format.github.io/
    - Generate your citation metadata files with ease: https://citation-file-format.github.io/cff-initializer-javascript/#/
4. Follow best practices for software development to improve the quality of your software.
    - Make your code modular
    - Have a code level documentation
    - Provide tests
    - Follow code standards
    - Use version control

There are useful guidelines below:

- The eScience Centre Guide: https://guide.esciencecenter.nl/#/

- Best Practices for Scientific Computing: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001745
- Good enough practices in scientific computing: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005510

6.2. Software Repositories and Platforms

Repositories:

- **PyPI**: https://pypi.org/; a repository of software for the Python programming language.
- **The Comprehensive R Archive Network (CRAN)** : https://cran.r-project.org/; is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R.
- **Docker Hub**: https://hub.docker.com/; is a hosted repository service for finding and sharing container images.
- **Zenodo**: http://zenodo.org/; allows depositing research papers, data sets, research software, reports, and assigns persistent identifiers.

Version control and collaboration platforms:

- **GitHub**: https://github.com/; is a cloud-based Git platform that allows developers to host and monitor their code changes. It has also evolved to become a development platform. It gives developers the option to implement apps and integrations freely through the GitHub marketplace.
- **GitLab**: https://gitlab.com/users/sign_in; https://about.gitlab.com/; is a cloud-based Git and DevOps platform that helps developers monitor, test, and deploy their code.
- **Bitbucket**: https://bitbucket.org/; is a version control repository hosting platform. It is a cost-effective solution if you're looking for a safe hosting service for your private, proprietary code.
- How to choose a proper repository hosting service? Visit the links:
  - https://kinsta.com/blog/gitlab-vs-github/;
  - https://kinsta.com/blog/bitbucket-vs-github/;
  - https://software.ac.uk/choosing-repository-your-software-project
- Advantages of versioning:
  - Persistence of identifiers pointing to different/earlier versions
  - Maintaining previous versions of code, software, and data
  - Sharing various levels of processed data (primary, secondary, raw/clean/processed, etc.).
  - De-accessioning of data that has reached the end of its life cycle.

Sources and Relative Links:

TOP 10 FAIR DATA & SOFTWARE THINGS: Research Software:
https://librarycarpentry.org/Top-10-FAIR//2018/12/01/research-software/

TOP 10 FAIR DATA & SOFTWARE THINGS: Biomedical Data Producers, Stewards, and Funders: https://librarycarpentry.org/Top-10-FAIR//2018/12/01/biomedical-data-producers/

Lamprecht Anna-Lena et al. '*Towards FAIR Principles for Research Software'* 1.Jan 2020: 37-59. https://www.iospress.com/news/towards-fair-principles-for-research-software

From FAIR research data toward FAIR and open research software: https://www-degruyter-com.sheffield.idm.oclc.org/document/doi/10.1515/itit-2019-0040/html

Top 10 metrics for life science software good practices: https://f1000research.com/articles/5-2000/v1

Four simple recommendations to encourage best practices in research software: https://f1000research.com/articles/6-876/v1

File naming - 8-minute video with guidance on file naming and version control, especially relevant for software developers: https://www.youtube.com/watch?v=3MEJ38BO6Mo (taken from https://www.data.cam.ac.uk/support/external)

Good practice: *4 Simple recommendations for Open Source Software*: https://softdev4research.github.io/4OSS-lesson/05-use-registry/index.html

Good practice: *FOSTER, Open Source Software and Workflows*: https://www.fosteropenscience.eu/node/2329