FAIR data and software checklist for the School of Biosciences
Ecology & Evolutionary Biology and Plants, Photosynthesis & Soil clusters

author: Zuzanna Zagrodzka

The main aim of FAIR guiding principles proposed by Wilkinson et al. in 2016 was to identify the difficulties of discovering and reusing data and to call for infrastructures to improve the machine-actionability of their services. In 2018 the source code of software has been recognised as a digital object in a 'FAIR ecosystem' by the European Commission. Since the publication of the FAIR principles, many funders and journals now require scientists to follow them and to mandatorily deposit their data and software in line with the FAIR principles. The main reason for this is that FAIR data and software have the potential to revolutionise science by improving its transparency, reusability and applicability.

What means FAIR for data?

The principles refer to three types of entities: data (or any digital object), metadata (information about that digital object), and infrastructure (searchable resources). FAIR (Findable, Accessible, Interoperable and Reusable) data is easily findable high-quality data that is machine-actionable, so computational systems can find, access, interoperate, and reuse data with no or minimal human intervention). The aim of this principle is to make data as open as possible and as closed as necessary.

Findable means that data should be easy to find for both humans and computers. Machine-readable metadata are essential for the automatic discovery of datasets and services.

Accessible means that once the user finds the required data, they need to know how they can be accessed, possibly including authentication and authorisation.

Interoperable means that the data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

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.

Useful resources:

"The FAIR Guiding Principles for scientific data management and stewardship" Wilkinson et al., 2016
https://doi.org/10.1038/sdata.2016.18

"Towards open, reliable, and transparent ecology and evolutionary biology" O'Dea et al., 2021
https://doi.org/10.1186/s12915-021-01006-3

	FAIR checklist for data	Y/N/na
Findable		
F1	(Meta)data are assigned a globally unique and persistent identifier (e.g. DOI) *3.4. Persistent identifiers*	
F2	Data are described with rich metadata *2.13. Data documentation ("readme" file/metadata)*	
F3	Metadata clearly and explicitly include the identifier of the data it describes *2.4. Citing data*	
F4	(Meta)data are registered or indexed in a searchable resource *3. Data archiving*	
Accessible		
A1	(Meta)data are retrievable by their identifier using a standardized communications protocol *2.3. Finding and reusing data, 2.4. Citing data*	
A1.1	The protocol is open, free, and universally implementable *2.13. Data documentation ("readme" file/metadata)*	
A1.2	The protocol allows for an authentication and authorization procedure, where necessary *2.9. Workflow documentation, 2.13. Data documentation ("readme" file/metadata)*	
A2	Metadata are accessible, even when the data are no longer available	
Interoperable		
I1	(Meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation *2.10. Data formats, 2.11. Data organisation and data standards, 2.12. Data dictionary*	
I2	(Meta)data use vocabularies that follow FAIR principles *2.11. Data organisation and data standards*	
I3	(Meta)data include qualified references to other (meta)data *2.4. Citing data*	
Reusable		
R1	(Meta)data are richly described with a plurality of accurate and relevant attributes *2.13. Data documentation ("readme"file/metadata)*	
R1.1	(Meta)data are released with a clear and accessible data usage license *3.5. Licensing and Data Access Statement*	
R1.2	(Meta)data are associated with detailed provenance *2.1. Ownership of data, 2.2. Authorship of data, 2.4. Citing data*	
R1.3	(Meta)data meet domain-relevant community standards *2.10. Data formats, 2.13.2. Standards-based metadata*	

	FAIR checklist for code and software	Y/N/na
Findable		
F1	Software and its associated metadata have a global, unique and persistent identifier for each released version. *4.8. Unique and persistent identifier for software*	
F2	Software is described with rich metadata. *4.1. Best practices for software development, 4.2. Software description (metadata)*	
F3	Metadata clearly and explicitly include identifiers for all the versions of the software it describes. *4.2. Software description (metadata), 4.6. Version control, 4.11. Cite software*	
F4	Software and its associated metadata are included in a searchable software registry. *4.7. Software registry/repository*	
Accessible		
A1	Software and its associated metadata are accessible by their identifier using a standardized communications protocol. *4.8. Unique and persistent identifier for software*	
A1.1	The protocol is open, free, and universally implementable. *4.7. Software registry/repository, 4.9. Executable version*	
A1.2	The protocol allows for an authentication and authorization procedure, where necessary. *4.3. Software documentation*	
A2	Software metadata are accessible, even when the software is no longer available. *4.7. Software registry/repository*	
Interoperable		
I1	Software and its associated metadata use a formal, accessible, shared and broadly applicable language to facilitate machine readability and data exchange. *4.2. Software description (metadata), 4.3. Software documentation*	
I2S.1	Software and its associated metadata are formally described using controlled vocabularies that follow the FAIR principles. *4.2. Software description (metadata), 4.3. Software documentation*	
I2S.2	Software use and produce data in types and formats that are formally described using controlled vocabularies that follow the FAIR principles. *4.4. Software functionality, 4.5. Standard formats for inputs and outputs*	
I4S	Software dependencies are documented and mechanisms to access them exist. *4.3. Software documentation*	
Reusable		
R1	Software and its associated metadata are richly described with a plurality of accurate and relevant attributes. *4.2. Software description (metadata)*	
R1.1	Software and its associated metadata have independent, clear and accessible usage licenses compatible with the software dependencies.	

	4.10. Software license	
R1.2	Software metadata include detailed provenance, detail level should be community agreed. *4.1. Best practices for software development*	
R1.3	Software metadata and documentation meet domain-relevant community standards. *4.2. Software description (metadata), 4.3. Software documentation*	

FAIR (meta)data

The checklist below should help you prepare your (meta)data in a format that complies with the FAIR principles. It is "Findable" and "Accessible" by its globally unique and persistent identifier (DOI), described with rich metadata and registered in the right repository, "Interoperable" as it uses a broadly applicable standard and vocabularies, "Reusable" as it is associated with detailed provenance and released with a correct license. Literature discussing a need for improving research practices focusing on FAIR data is booming in recent years. Below you can find some interesting titles.

Learn more:

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

Useful resources:

FAIR 'cookbook' for the life sciences: https://faircookbook.elixir-europe.org/content/home.html

"Making forest data fair and open" de Lima et al., 2022 https://doi.org/10.1038/s41559-022-01738-7

"How FAIR are plant sciences in the twenty-first century? The pressing need for reproducibility in plant ecology and evolution" Manzano and Julier 2021 https://doi.org/10.1098/rspb.2020.2597

"Managing data locally to answer questions globally: The role of collaborative science in ecology" Aubin et al., 2020 https://doi.org/10.1111/jvs.12864

"A checklist recipe: making species data open and FAIR" Reyserhove et al., 2020 https://doi.org/10.1093/database/baaa084

"Six Simple Steps to Share Your Data When Publishing Research Articles" Serrano, 2019 https://doi.org/10.1002/lob.10303

"Navigating the unfolding open data landscape in ecology and evolution" Culina et al., 2018 https://doi.org/10.1038/s41559-017-0458-2

1. Pre-data collection steps

The easiest and least time-consuming approach to ensure that your data is FAIR is to start thinking about how to make it FAIR before data collection begins. Pre-data collection steps are not directly related to the FAIR principles but it will help you to comply with them later.

Learn more:

https://www.sheffield.ac.uk/library/rdm

1.1. Data management plan (DMP)

Creating a DMP at the start of your research will help you organise your data and think about choosing the right repository to archive it. DMPs are required by most research funders and recommended as part of the University's Research Data Management Policy.

Learn more:

https://www.sheffield.ac.uk/library/rdm/dmp

Useful resources:

https://www.sheffield.ac.uk/polopoly_fs/1.553350!/file/GRIPPolicyextractRDM.pdf

https://www.britishecologicalsociety.org/wp-content/uploads/Publ_Data-Management-Booklet.pdf

1.2. Ethics approval

All University of Sheffield research involving human participants, personal data and human tissue need ethics approval. If your research involves human participants, make sure that consent includes your plans for long-term storage and sharing of data. To check if your research requires ethics approval visit this website: https://students.sheffield.ac.uk/research-services/ethics-integrity/research-ethics

2. Data collection and sample processing

At this stage, it is not too early to start thinking about the FAIR data. Later during the process of archiving data, you will be asked to provide metadata/read.me file. By providing detailed information about how your data were collected you will be able to maintain a reproducible workflow, avoid potential misuse of your data by other researchers and save time during the data archiving process.

2.1. Ownership of data

If you collect raw data check the funders' requirements. Most researchers funded by the research council will have to publish their data accordingly with the FAIR principles at the end of the project. In case of collaboration with researchers from different institutions, charities or industrial partners, you will have to discuss and sign the co-owning agreements. It is important to have a conversation about data ownership at the beginning of the project. Research services or the Library should be able to advise you in case of any doubts.

Research services https://www.sheffield.ac.uk/research-services

Library https://www.sheffield.ac.uk/library/libstaff/scholarlycomms

If you reuse data you need to check the licence it comes with. Sometimes you will have to contact the data creators and make sure you can use their data. You should cite all datasets you have used in your research.

Learn more:

https://www.sheffield.ac.uk/library/rdm/citation

2.2. Authorship of data

Often data repositories can have a different list of contributors therefore, data providers or collectors might be listed as authors on the dataset but not the publication.

2.3. Finding and reusing data

Published datasets can be a valuable resource to use and cite in your research. Datasets across many disciplines can be found in data centres and repositories, which are generally established by research funders and researcher communities. Go to “3.1. Registry of Research Data Repositories” for more information on where to look.

Learn more:

https://www.sheffield.ac.uk/library/rdm/finddata

When using published data, you should always check and comply with the copyright and licensing terms specified by the data depositor. If there is no licence you might have to contact the creator or depositor directly.

Learn more:

https://creativecommons.org/about/cclicenses/

Useful resources:

If you collect data to conduct a meta-analysis, familiarise yourself with the PRISMA checklist for reporting in systematic reviews and meta-analyses.

https://prisma-statement.org/Extensions/EcoEvo

“Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: a PRISMA extension” O’Dea 2021 https://doi.org/10.1111/brv.12721

2.4. Citing data

Citing data properly is equally as important as citing journal articles and other papers. Cite your and others' data properly in your metadata and paper. In general, a data citation should include author/creator, date of publication, title of dataset, version, publisher/organisation, and unique identifier (preferably DOI).

Learn more:

https://www.sheffield.ac.uk/library/rdm/finddata

2.5. Organising and naming files

Create a logical file structure with consistent filenames, and make sure that everyone involved in the research works within the structure. It will help you later during the data archiving process.

Learn more:

https://www.sheffield.ac.uk/library/rdm/organising#tab01

2.6. Version control

The ability to consistently track and retrieve a specific version of a file can lead to more efficient collaboration and increased accuracy of research results. This is most effectively accomplished through the use of version control systems that automate various portions of the storage and record-keeping, e.g. GitHub.

Keep in mind that GitHub is good for version control but not for archiving. Data and/or code in GitHub can be deleted or modified at a later date meaning it does not meet the criteria for data availability. You should additionally deposit this code/data in a permanent repository (like ORDA) which will assign a DOI.

Learn more:

https://www.sheffield.ac.uk/library/rdm/organising#tab02

Useful resources:

https://help.osf.io/article/149-version-control

https://srse-git-github-zero2hero.netlify.app/00-intro-to-version-control/

2.7. Data storage

Regularly back up your digital data to several secure locations, preferably in University research data storage or University Google drive. If you collect or create physical data, digitise them if possible.

Learn more:

https://www.sheffield.ac.uk/library/rdm/storage

Useful resources:

https://www.protocols.io/

"How to pick an electronic laboratory notebook" Kwok 2018 https://doi.org/10.1038/d41586-018-05895-3

2.8. Data processing

You should always favour using open-source software to process and analyse your data. Keep detailed information on how you analysed the data and convert the output file into an open format where necessary.

Raw data vs. processed data

Keep in mind that journals don't always require submitting raw data collected during an investigation, especially, if the standard in the field is to share data that have been processed (e.g. CSV files recording response to stimuli rather than the electrical signals on which they were based). You should still consider long-term storage of all your data, for example, in case you would like to re-analyse it using a different, perhaps novel method.
Go to "4. FAIR Software" for more detailed information.

2.9. Workflow documentation

Good scientific practice includes good record keeping, which ensures not only transparency and reproducibility but also accountability. Keeping all records such as detailed protocols, methodologies, equipment used, code used to process the data, and software settings (microscope software, imaging software). Consider using an electronic lab notebook, keeping your notes on Google docs or regularly digitising your lab book (scan your records). If you know where your data will be archived at the end of your project familiarise yourself with the repository metadata requirements.

Markowetz (2015) identified five selfish reasons why you should make your work reproducible:

- reproducibility helps to avoid disaster
- reproducibility makes it easier to write papers
- reproducibility helps reviewers see it your way
- reproducibility enables continuity of your work
- reproducibility helps to build your reputation

Learn more:

https://www.sheffield.ac.uk/library/rdm/organising#tab03

Useful resources:

"Five selfish reasons to work reproducibly" Markowetz, 2015 https://doi.org/10.1186/s13059-015-0850-7

"The Turning Way: Guide for Reproducible" https://the-turing-way.netlify.app/reproducible-research/reproducible-research.html

https://nceas.github.io/oss-2017/lessons.html

2.10. Data formats

Data should be shared and archived in standard and open formats (.txt, .csv). For example, Excel spreadsheets (.xls) should be converted to a plain text format, such as comma-separated values (.csv) or tab-delimited ASCII text (.txt), FASTA is a format for representing molecular sequences like DNA, RNA and protein that most sequence analysis tools can handle and NetCDF is a standard file format used sharing of array oriented scientific data.

Avoid proprietary formats that are owned by a company that claims intellectual property rights for the use of the software by granting licenses.

Contact Research Software Engineering if you need any advice on how to handle non-open file formats, and make your code or software FAIR.

Research Software Engineering https://rse.shef.ac.uk/

If you used specialised equipment that saves data in proprietary file formats you should consider changing their formats at the early stage of your analysis. Some research groups are developing workflows that do it and can be integrated with a wider range of open source tools.

Learn more:

https://www.sheffield.ac.uk/library/rdm/organising#tab00

https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/file-formats/

Example from the School of Biosciences

Untargeted metabolomics workflow for the biOMICS facility is a step-by-step guide for each stage of the untargeted metabolomics workflow including refactored code for each stage of the workflow that is conducted using our own software. It has a number of useful features and sanity checks to the code so that the workflow is customisable and can be integrated with a wider range of open-source metabolomics tools.

You can find the workflow here:
https://github.com/LizzyParkerPannell/Untargeted_metabolomics_workflow.git

Useful resources:

"Chapter Six - A practical guide to implementing metabolomics in plant ecology and biodiversity research" Uthe et al., 2021 https://doi.org/10.1016/bs.abr.2020.09.019

2.11. Data organisation and data standards

Spreadsheet programs are used for data entry, storage, analysis, and visualization, but they are best suited only for data entry and storage. Analysis and data visualization should ideally happen in a separate program to reduce the risk of contaminating or destroying the raw data in the spreadsheet program. It is important to follow certain rules to organise your data in spreadsheets.

Useful resources:

"Data Organization in Spreadsheets" Broman and Woo 2018
https://www.tandfonline.com/doi/pdf/10.1080/00031305.2017.1375989?needAccess=true

"The Turning Way. Data Organisation in Spreadsheets" https://the-turing-way.netlify.app/reproducible-research/rdm/rdm-spreadsheets.html

"Towards an ecological trait-data standard" Schneider et al., 2019 https://doi.org/10.1111/2041-210X.13288

Ontologies and thesaurus

Many projects produce a huge amount of data, but the diversity in methodologies used to collect data, scales and topics covered, result in large numbers of small datasets using heterogeneous terminologies. Recently we can see a trend toward defining standards for acquiring, organising and describing data.

Thesaurus is a controlled vocabulary that provides key terms with their associated concepts and relations for a specific field or domain of interest. Ontology includes a set of standardized terms

(each with a numerical code) describing features of structure and development as well as logical relationships between them; commonly published in OWL format for machine readability.

Finding and applying the most suited thesauri and ontologies is not an easy task. Their definitions in some domains are better covered than others. Also, often different curation strategies and measures for peer-review and community building are employed.

You can look for suitable for you biological ontologies on these platforms:

- Ontobee http://www.ontobee.org/
- GFBio Terminology Service https://terminologies.gfbio.org/
- OBO Foundry https://obofoundry.org/

You can find thesaurus and normative term names, and definitions:

- eLTER Vocabularies https://vocabs.lter-europe.net/envthes/en/
- Darwin Core https://dwc.tdwg.org/terms/
- Ecological Trait-data Standard https://terminologies.gfbio.org/terms/ets/pages/

Useful resources:

"Ocean Data Product Integration Through Innovation-The Next Level of Data Interoperability" Buck et al., 2019 https://doi.org/10.3389/fmars.2019.00032

"Ecological Data Should Not Be So Hard to Find and Reuse" Poisot 2019 https://doi.org/10.1016/j.tree.2019.04.005

"The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics" Cooper et al., 2018 https://doi.org/10.1093/nar/gkx1152

"Towards a thesaurus of plant characteristics: an ecological contribution" Garnier et al., 2016 https://doi.org/10.1111/1365-2745.12698

"The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation" Buttigieg et al., 2016 https://doi.org/10.1186/s13326-016-0097-6

"The environment ontology: contextualising biological and biomedical entities" Buttigieg et al., 2013 https://doi.org/10.1186/2041-1480-4-43

"Ontologies as integrative tools for plant science" Walls et al., 2012 https://doi.org/10.3732/ajb.1200222

"Advancing ecological research with ontologies" Madin et al., 2007 https://doi.org/10.1016/j.tree.2007.11.007

You can download a free "The gene ontology handbook" by Dessimoz and Skunca (2017) from https://oapen.org/

An example of well-organised datasets (Schneider et al., 2019):

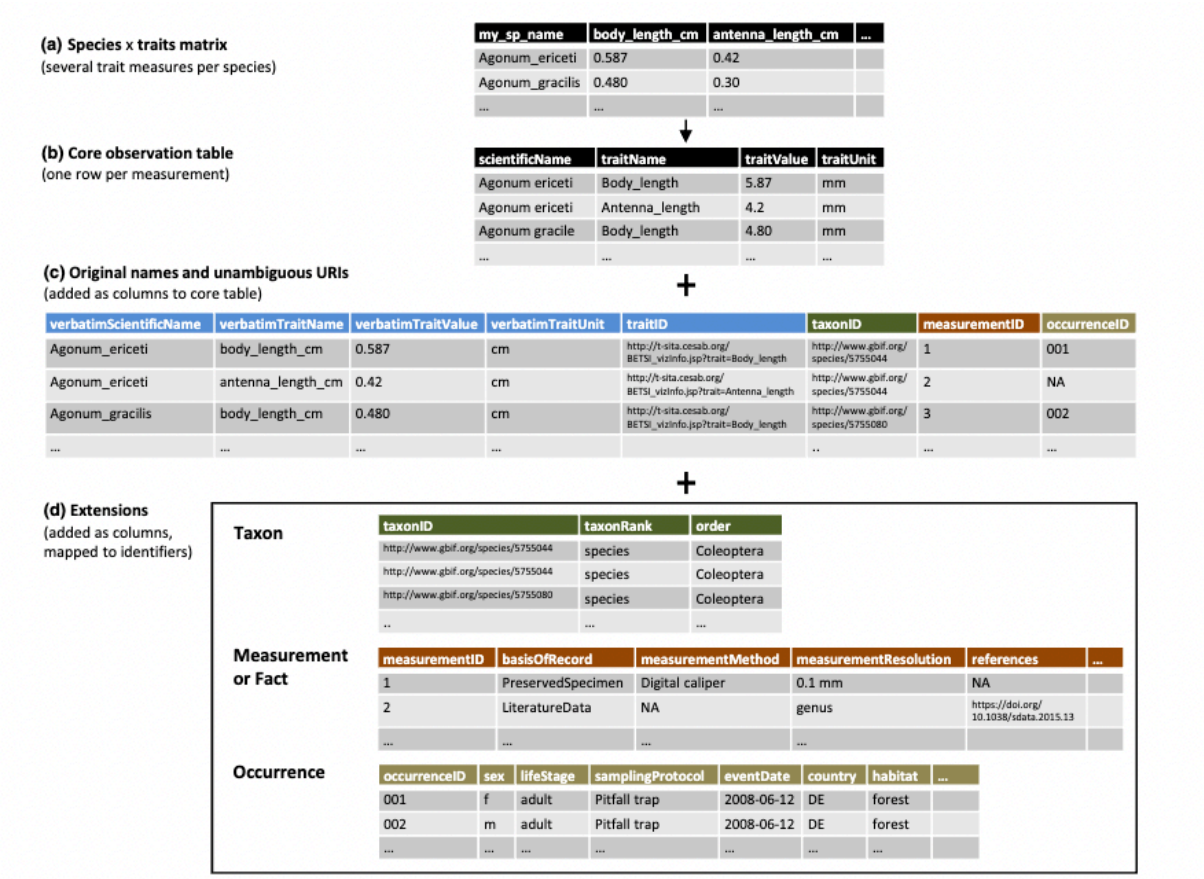


FIGURE 2 Formats used for trait datasets: (a) taxon-level trait data compiled from literature or aggregated from measurements are often published as a compiled species × traits wide-table; (b) observation long-tables are a well-defined and tidy data format, reporting one single measurement per row and relating it to a standard trait definition and accepted taxon name; (c) additional columns may provide original names for maintaining author-side continuity, identifiers reference to taxa and trait concepts via unambiguous URI pointers. Additional identifiers relate each row to other layers of information on (d) the taxon resolution, the individual organism (i.e. occurrence), or the origin of or confidence in the reported measurement or fact

2.12. Data dictionary

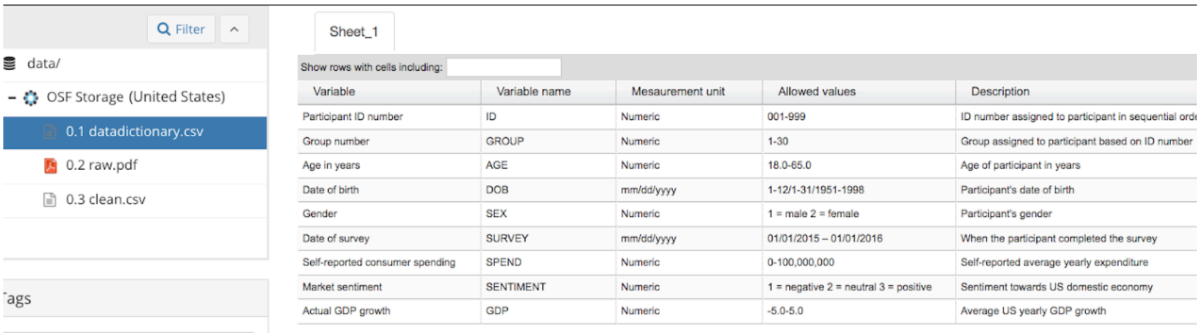
Create a data dictionary, it can be used to fill in entity and attribute section or feature catalogue of formal metadata. The purpose of a data dictionary is to explain what all the variable names and values in your spreadsheet mean so others can understand your data. Instead of coming up with your own terms, you might want to familiarise yourself with subject-specific terms and ontologies (See: 2.11. Data organisation and data standards).

Useful resources:

https://help.osf.io/article/217-how-to-make-a-data-dictionary

https://www.makeareadme.com/

Image source: https://help.osf.io/article/217-how-to-make-a-data-dictionary



Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential ord
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth

2.13. Data documentation ("readme" file/metadata)

There are different ways in which you can document your data depending on the context within which it is being collected:

in a 'readme' file: any information that cannot be recorded in a structured way (i.e. as the values of fields in a data or metadata file) can be recorded as free text within a readme file.

in a separate metadata file: some disciplines have developed special file formats or data structures for recording supporting information. High-quality metadata includes information about the data, such as creator, keywords, units, flags, how the data were collected, why, where, machines/instruments used, such as make, model, and manufacturer. Process metadata includes details about any cleaning/analysis steps and scripts used to create data products.

in a file mimicking a web form: in some cases, archives generate specialist metadata files from their submission forms. Find out the fields of the submission form of the archive to which you are planning to submit your data, copy these fields into your data documentation and fill these in as you go through your project.

Documenting your data only in a published journal article is not recommended.

Useful resources:

https://uwyo.libguides.com/data-mgmt/metadata

2.13.1. "Readme" style metadata

Standards-based metadata is generally preferable, but where no appropriate standard exists, writing “readme” style metadata is an appropriate strategy. A readme file provides information about a data file and is intended to help ensure that the data can be correctly interpreted.

Try to provide a clear and concise description of all relevant details about data collection, processing, and analysis so others can easily interpret and reanalyse your dataset.

It is best practice to create a README file for each dataset regardless of whether it is being deposited in a repository because the document might become necessary at a later point.

Learn more:

https://data.research.cornell.edu/content/readme

2.13.2. Standards-based metadata

If possible your metadata should be logically organised and correspond to the metadata disciplinary standards. Often when submitting to the subject-specific repository you will be informed about repository metadata requirements. The aim is to avoid datasets being stored in variable tabular structures and labelled following self-defined terms, which makes extraction and further reuse or synthesis unnecessarily tedious. Standards improve the quality and share-ability of data by increasing data compatibility, improving the consistency and efficiency of data collection, and reducing data redundancy.

Before the submission, you can familiarise yourself with the standards used by other researchers in your discipline. There are multiple websites you can use to search for them:

Disciplinary Metadata guide https://www.dcc.ac.uk/guidance/standards/metadata

RDA Metadata Standards Catalog https://rdamsc.bath.ac.uk/subject-index

FAIRsharing registry https://fairsharing.org/search?fairsharingRegistry=Standard

Open directory of metadata standards (via Research Data Alliance) http://rd-alliance.github.io/metadata-directory/standards/

Biodiversity Information Standards (TDWG) https://www.tdwg.org/standards/

Some of the repositories require the datasets to follow certain standards. They should guide you and assist you during the data archiving process. For example, Global Biodiversity Information Facility (GBIF) requires standards such as Darwin Core Standard (DwC), Ecological Metadata Language (EML), and BioCASe/ABCD.

Darwin Core (DwC): A metadata specification for information about the geographic occurrence of species and the existence of specimens in collections. More details: http://rs.tdwg.org/dwc/, https://github.com/tdwg/dwc

“Darwin Core: An Evolving Community-Developed Biodiversity Data Standard” Wieczorek et al., 2012 https://doi.org/10.1371/journal.pone.0029715

Access to Biological Collection Data (ABCD): A standard for the access to and exchange of data about specimens and observations (a.k.a. primary biodiversity data). More details: http://www.tdwg.org/standards/115/

“The ABCD of primary biodiversity data access” Holetschek et al., 2012 https://doi.org/10.1080/11263504.2012.740085

Ecological Metadata Language (EML): A metadata specification developed by the ecology discipline for the ecology discipline. EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document ecological data.
More details: http://knb.ecoinformatics.org/software/eml/

"Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation" Fegraus et al., 2005
https://www.jstor.org/stable/bullecosociamer.86.3.158

Some data types should refer to specific standards. For example, microarray data should refer to the MIAME standard, metabolomics to MSI standard, and proteomics to the MIAPE recommendations.

MIAME: "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data" Brazma et al., 2001 https://doi.org/10.1038/ng1201-365

MSI: "The metabolomics standards initiative (MSI)" Fiehn et al., 2007
https://doi.org/10.1007/s11306-007-0070-6

MIAPE: "The minimum information about a proteomics experiment (MIAPE)" Taylor et al., 2007
https://doi.org/10.1038/nbt1329

3. Data archiving

Archiving research data (as well as metadata files and any derived data products) is a crucial step in any research that is required by many funding agencies and journals' policies. It secures data for future use, makes access easier and increases the efficiency of collecting and storing research data. Archiving good quality data complying with the FAIR principles increases transparency in research, promotes critical evaluation and increases the credibility of your research. The biggest challenge in ecology and evolutionary biology is a wide variety of data formats, file sizes, and states of data completeness.

Remember: samples or data that are collected, but never archived, analysed or reported lead to wasted research resources.

Data should be logically and consistently formatted. Formats used to create and collect data may vary according to discipline and practical requirements and may include proprietary formats readable only using specific software. For sharing and long-term preservation, however, data should be stored using standard or open formats. This will help to ensure the data remains accessible as technology progresses and changes. Planning what these formats will be at the beginning of a research project will reduce the risk of data being locked into a proprietary format, and the formats chosen should be detailed in your data management plan.

3.1. Registry of Research Data Repositories

Finding the "right" repository for your data can be overwhelming, but there are resources available to help you pick the best location for your data. You can start by going to the registry of research

data repositories such as re3data and FAIRsharing and search for a data repository related to your research subject area.

Re3Data: https://www.re3data.org/
FAIRsharing: https://fairsharing.org/search?fairsharingRegistry=Database

Also, you can check what repositories are recommended by the journals you want to publish your research.

Remember that when choosing the right repository, you should archive your data only if it meets the following requirements:

- ensure long-term persistence and preservation of datasets in their published form (you should store and share data for a minimum of 10 years after the end of the project or in line with funder requirements),
- provide metadata,
- provide stable persistent identifiers for submitted datasets, e.g. DOIs (Digital Object Identifiers),
- allow open public access to data, and support open licences (CC0 and CC-BY, or their equivalents).

Check whether individual repositories charge costs for depositing data and if these will be covered by your funding.

Learn more:

https://www.sheffield.ac.uk/library/rdm/repositories#tab00

3.2. Subject-specific repositories

Below you can find some recommended subject-specific repositories.

Ecology, taxonomy and species diversity:
- Global Biodiversity Information Facility (GBIF) https://www.gbif.org/
- The Knowledge Network for Biocomplexity (KNB) https://knb.ecoinformatics.org/
- Morphobank https://morphobank.org/
- Movebank Data Repository https://www.movebank.org/cms/movebank-main
- Broad scope Earth and environmental sciences:
- NERC Data Centres https://www.ceh.ac.uk/data/nerc-data-centre
- PANGAEA https://www.pangaea.de/
- HydroShare (CUAHSI) https://www.hydroshare.org/

Climate sciences:
- World Data Center for Climate (WDCC) https://www.dkrz.de/up/systems/wdcc

Ocean sciences:
- Marine Data Archive https://mda.vliz.be/
- SEANOE https://www.seanoe.org/

Mathematical and modelling resources:

- The Network Data Exchange (NDEx) https://home.ndexbio.org/index/

Raw sequencing data (reads or traces), genome assemblies, annotated sequences:

- INSDC repositories https://www.insdc.org/
- Genome Sequence Archive (GSA) https://ngdc.cncb.ac.cn/gsa/

Protein sequence:

- UniProtKB https://www.uniprot.org/
- Protein Data Bank Japan (PDBj) https://pdbj.org/

Proteomics:

- PRoteomics IDEntifications Database (PRIDE) https://wwwdev.ebi.ac.uk/pride/

3.3. Generalist repositories

If you cannot locate a repository for your discipline or the type of data you should store your data in a generalist repository. Generalist repositories accept data regardless of data type (e.g., videos, recordings, images), format (e.g., .jpg, .mp3), content, or disciplinary focus.

Researchers at the University of Sheffield are strongly encouraged to deposit their data in ORDA (if there is no subject-specific repository). It is provided by FigShare and enables University research data to be preserved, discovered and accessed.

Other generalist repositories:

- Dryad Digital Repository https://datadryad.org/
- Zenodo https://zenodo.org/
- Harvard Dataverse https://dataverse.harvard.edu/
- Open Science Framework https://osf.io/dashboard

Learn more:

https://www.sheffield.ac.uk/library/rdm/orda

3.4. Persistent identifiers

Getting a unique and persistent identifier (PID) is a necessary step in data archiving. A permanent identifier like a digital object identifier (DOI) is a unique ID assigned to a dataset to ensure that properly managed data does not get lost or misidentified. Additionally, a DOI makes it easier to cite and track the impact of datasets like cited journal articles. You should archive your metadata and data in repositories that provide stable persistent identifiers for submitted datasets.

Learn more:

https://datacite.org/value.html

3.5. Licensing and Data Access Statement

You should allow public access to your data (CC0 and CC-BY) unless your data contain sensitive human information, contain confidential commercial information, or in situations where there are sound public good reasons for restricting data (e.g. protection of endangered species). However, you should openly share protocols and metadata, and provide a detailed description of the methods you used to analyse data.

The UKRI-funded research articles (from 1st April 2022) must be openly available by the publication date. Moreover, all articles must include a Data Access Statement.

Examples of Data Availability Statements can be found here:
https://www.sheffield.ac.uk/library/rdm/publish
https://www.sheffield.ac.uk/library/copyright/licences

FAIR code and software

4. FAIR software

What the FAIR principles mean for (scientific) code and software is an ongoing discussion. The most recent summary of the current status of FAIR and software is described in Lamprecht et al. (2020) paper. To keep it simple, FAIR software is a software with sufficiently rich metadata and unique persistent identifier (Findable), its metadata is in machine and human readable format both deposited in trusted, community approved repository (Accessible), uses community accepted standards and platforms, making it possible for users to run the software (Interoperable) and has clear licence and documentation (Reusable).

The checklist below follows FAIR principles suggested by Lamprecht et al., 2020 "Towards FAIR principles for research software" https://doi.org/10.3233/DS-190026

In your readme file or metadata, you should have a dedicated section that lists all software with their version numbers and properly cited.

Note about this section

To limit the scope of this section, "software" here mainly refers to scripts and packages in open source languages like R and Python, but not to other kinds of software frequently used in research, such as advanced software, web services or web platforms.

Note about programming languages

There is a number of free open source high-level, dynamic programming languages that researchers use for numerical computing. The most commonly used are R, Python and Julia. MATLAB is not free and not open source. If you are looking for an open source alternative that is highly compatible with MATLAB you should check out GNU Octave https://octave.org/ . Using open-source programming languages and software make it easier to collaborate with researchers from all around the world. Moreover, you don't need to pay for a licence once you will not be affiliated with the institution that paid for it.

Note about the importance of the section

Most journals require now to provide (complete) code used to generate statistics and generate figures. Analysis code (such as R scripts) should be made available at the point of submission.

Useful resources:

"Low availability of code in ecology: A call for urgent action" Culina et al., 2020
https://doi.org/10.1371/journal.pbio.3000763

4.1. Best practices for software development and code writing

There are a number of actions you can take to improve the quality of your software: annotate your code, save session information, make your code modular, have code level documentation, provide tests, follow code standards, use version control, etc.

Make sure that you share all your code files including simulation code, analysis code and code used for the creation of figures/plots.

Useful resources:

"Guides to better science: Reproducible Code" by British Ecological Society:

https://www.britishecologicalsociety.org/wp-content/uploads/2019/06/BES-Guide-Reproducible-Code-2019.pdf

"Reproducible Data Science with Python" https://valdanchev.github.io/reproducible-data-science-python/intro.html

eScience Center Guide software development guide: https://zenodo.org/record/4020565#.YsS_ZOzMKqA

4 Simple recommendations for Open Source Software: https://softdev4research.github.io/4OSS-lesson/

4.2. Software description (metadata)

A good description of your software will enable people to learn what it does and if they can use it in their research. It should include short descriptive text and meaningful keywords. Moreover, for reproducibility and reusability purposes, state clearly which version of the software is described by the metadata.

Edam is an example of an ontology that provides terminology that can be used to describe bioinformatics software https://edamontology.org/page#

4.3. Software documentation

Your software should include sufficient documentation: instructions on how to install, run and use your software. All dependencies of your software should be clearly stated. Provide sufficient examples on how to execute the different operations your software offers, ideally along with example data.

4.4. Software functionality

Your software performs one or more operations that take an input and transform it into the output. You should provide a clear and concise description of the operations along with the corresponding input and output data types.

List all operations that your software provides, and describe them along with corresponding input and output data types. If possible, use terms from a domain ontology like EDAM.

4.5. Standard formats for inputs and outputs

In order for people to use your software, they need to know how to feed data to it – standard and open formats are easy ways to exchange data between different pieces of software. By sticking to standards, it is possible to use the output from another piece of software as an input to your software (or the other way around).

4.6. Version control

Using a version control system allows you to easily track changes in your software, both your own changes as well as those made by collaborators. By using GitHub, GitLab or Bitbucket, you will have backups of every version of your software.

Useful resources:

https://srse-git-github-zero2hero.netlify.app/

4.7. Software registry/repository

Registering your software in a dedicated registry will make it findable by search engines like Google. The registries will usually ask you to provide descriptions (metadata). Generalist repositories can be used to archive your code. For example, Zenodo promises metadata, and a snapshot of the software release, to be available for the upcoming 20 years, even when the versioned source code on GitHub may not be accessible anymore.

Examples of research software registries:

- bio.tools https://bio.tools/
- ORDA https://orda.shef.ac.uk/
- Zenodo https://zenodo.org/

A list of other research software registries: https://github.com/NLeSC/awesome-research-software-registries

Learn more:

https://www.sheffield.ac.uk/library/research/openresearch

4.8. Unique and persistent identifier for software

It will help others find and access the particular version of your software. For example, ORCID or Zenodo provides you with a DOI. Recent initiatives, such as software Heritage, propose to associate a permalink as intrinsic SHA1 identifier to software.

4.9. Executable version

In order for anyone to use your software, they need to be able to download an executable version along with documentation. For interpreted languages like Python and R, the code is also the executable version. Downloading the software and documentation should be possible from a project website, a git repository or from a software registry.

4.10. Software license

A license tells your (potential) users what they are allowed to do with your software (and what not to do), and can protect your intellectual property. Repository such as ORDA will let you choose a license under which to make your data available. “Choose a license” website provides a simple guide for picking the right license for your software.

Useful resources:

https://choosealicense.com/licenses/

https://tldrlegal.com/

How to add a license to your GitHub repository: https://docs.github.com/en/communities/setting-up-your-project-for-healthy-contributions/adding-a-license-to-a-repository

4.11. Cite software

By providing the citation guideline you will help users of your software to cite your work properly. A proper citation includes author name(s), software title, version number, and unique identifier/locator. The software must also be cited in the text and in references.

Useful resources:

https://www.software.ac.uk/how-cite-software

https://www.r-bloggers.com/2018/08/how-to-cite-packages/

https://ropensci.org/blog/2021/11/16/how-to-cite-r-and-r-packages/