The FAIR Checklist for the School of English

1. Introduction 2

1.1 What is research data? 2

1.2 What is FAIR and why is it important? 2

2. Plan (at the start of your research) 3

3. Collect, Organise, and Deposit (during the research process) 4

3.1 Finding and reusing existing data 4

3.2 Documenting and Organising 4

3.3 Storing and archiving 6

4. Share (at the end of your research) 6

4.1 Benefits of sharing 6

4.2 What to share? 7

4.3 Prepare your data to share 7

4.4 Where and how to share? 8

5. Resources (in alphabetical order) 9

5.1 Archives, Repositories, and Databases 9

5.2 Data tools 12

5.3 Corpora 13

(Author: Yi Liu, July 2022)

1. Introduction

1.1 What is research data?

Research data is 'the evidence used to inform or support research conclusions' (University of Sheffield n.d.). It can be digital or non-digital, and includes but is not limited to:

- audio, video, images
- archival materials and handwritten notes
- questionnaires and interview transcripts
- spreadsheets, field notebooks, and diaries
- code and software

So, the FAIR principles not only apply to STEM (Science, technology, engineering, and mathematics), but they are also highly relevant to the **Arts and Humanities**.

Sources:
Research Data Management from The University Library:
https://www.sheffield.ac.uk/library/rdm/whatisrdm
Research data from the Consortium of European Social Science Data Archives (CESSDA)
https://dmeg.cessda.eu/Data-Management-Expert-Guide/1.-Plan/Research-data

1.2 What is FAIR and why is it important?

- **F**indable: your data, metadata[1], and datasets should be easy to find by both humans and machines, with globally unique and persistent identifiers.
- **A**ccessible: how your data, metadata, and datasets can be accessed, including limitations on the use of data and protocols for querying or copying data. Metadata should remain accessible, even when the data are no longer available.
- **I**nteroperable: your (meta)data should use standardised terms, and can be integrated with other data and interoperated with various applications or workflows for analysis, storage, and processing.
- **R**eusable: your (meta)data can be replicated and/or combined with other data in different settings with clear accessible data usage licences.

FAIR data and data tools enable the research community to:

- Maximise the value and usefulness of research data
- Increase accountability of your research and verifiability of the results
- Enhance the potential for collaboration
- Speed up the progress of academic research

[1] Metadata is "data that provides information about other data" (Merriam-Webster, retrieved June 2022).

Examples and relevant links:
FAIR data principles and their application to speech and oral archives:
https://www.tandfonline.com/doi/abs/10.1080/09298215.2018.1473449?journalCode=nnmr20
Applying FAIR principles in digital cultural heritage:
https://pro.europeana.eu/post/europeana-and-the-fair-principles-for-research-data
Sustainable and FAIR data sharing in the Humanities from ALLEA:
https://allea.org/portfolio-item/sustainable-and-fair-data-sharing-in-the-humanities/
The FAIR Guiding Principles for scientific data management and stewardship:
https://www.nature.com/articles/sdata201618

2. Plan (at the start of your research)

- Think about what data you are going to use or collect, and what can be shared at the end of your research:
  - Can it be reused for **future research** by yourself? Can it be reused for other research aims **by other researchers**?
  - Can your **experiment design, workflow, or code and software** be reused by others?
  - If your research involves **human participants**, how much consent is required, and for how long should you obtain permission for the storage and sharing of data?
  - If your research involves **materials archived by yourself** or your research team, can they be published for future reference and be reused by others?
- Make a data management plan (DMP):
  - Use DMPonline to create your plan: https://dmponline.sheffield.ac.uk
  - Contact the team at rdm@sheffield.ac.uk for questions and further information.

Relevant links:
(**Data management**)
Data management planning from the University Library:
https://www.sheffield.ac.uk/library/rdm/dmp
How to Develop a Data Management and Sharing Plan from Digital Curation Centre:
https://www.dcc.ac.uk/guidance/how-guides/develop-data-plan
Research data management from UK Data Service:
https://ukdataservice.ac.uk/learning-hub/research-data-management/
Research data management toolkit from Jisc: https://www.jisc.ac.uk/full-guide/rdm-toolkit
(**Ethical considerations**)
Ethics and integrity from Research Services at the University of Sheffield:
https://www.sheffield.ac.uk/research-services/ethics-integrity
Consent for data sharing from UK Data Service: https://ukdataservice.ac.uk/learning-hub/research-data-management/ethical-issues/consent-for-data-sharing/

3. Collect, Organise, and Deposit (during the research process)

3.1 Finding and reusing existing data

- Use data archives, repositories, and databases (see Section 5 for resources)
- Keep a record of data and where it was found
- Check terms and conditions:
  - Are the data publicly available and will they continue to be?
  - Are there restrictions on sharing the data and data derived from them?
  - Do arrangements have to be made with the original researcher(s)?
  - Don't assume you can share data just because they are available online
- Guidance on *Finding and reusing data* from the University Library: https://www.sheffield.ac.uk/library/rdm/finddata

3.2 Documenting and Organising

- Make data clear to understand and easy to use. Check out the **existing data formats and standards** in your research area, which will prepare you for data sharing (For more details, see Section 4.2).
- Unambiguous file naming and folder structure:
  - Are there specific standards that you want to implement, such as naming conventions or standardised coding structures?
  - Create and follow the file naming conventions. For example:

    (i) use consistent punctuation, spelling, version numbers, and abbreviations

    (ii) use YYYYMMDD date format

    (iii) make file names concise but meaningful.
  - Create a logical file structure with consistent filenames and make sure that everyone involved in the research works within the structure. Avoid repetition within filenames and folders.
  - **Good practice:**
    How best to keep track of your data file from the UK Data Archive: https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/organising/
    Processing Qualitative Data Files from the Finnish Social Science Data Archive: https://www.fsd.tuni.fi/en/services/data-management-guidelines/processing-qualitative-data-files/

- Comprehensive data documentation and metadata:
  - Create clear documentation for your data, ideally in a **README** file, which normally includes your methodology, file structure, and any other information that will help you and others to understand and use the data.
  - **Project-level documentation** explains "the aims of the study, what the research questions/hypotheses are, what methodologies were being used, what instruments and measures were being used". **Data-level documentation** provides "the information at the level of individual objects such as pictures or interview transcripts or variables in a database" (CESSDA, retrieved June 2022). The examples of data-level documentation for **qualitative vs. quantitative data**, see: https://dmeg.cessda.eu/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata
  - For the **metadata**, you should think about:

    (i) What metadata will you use? In case metadata standards do not exist in your discipline, outline what type of metadata will be created and how.

    (ii) If you already know in which data repository you will publish your data, what metadata standard do they use?

    (iii) Can metadata be added directly into the files or will the metadata be produced in another program or document?
  - **Examples and tools:**
    Metadata Template from the Georgia Tech Library:
    https://www.library.gatech.edu/smartech-metadata
    Dublin Core Metadata Generator: http://nsteffel.github.io/dublin_core_generator/
    Guide to writing "readme" style metadata from Cornell University:
    https://data.research.cornell.edu/content/readme

Relevant links:
Organising data from the University Library:
https://www.sheffield.ac.uk/library/rdm/organising#tab00
Organise & Document from CESSDA:
https://dmeg.cessda.eu/Data-Management-Expert-Guide/2.-Organise-Document
Documentation resources from UK Data Service: https://ukdataservice.ac.uk/learning-hub/research-data-management/document-your-data/documentation-resources/
What Is Metadata and How Do I Document My Data? by Dr Alexander Jedinger at the CESSDA Training Days 2019 (CTD2019): http://doi.org/10.5281/zenodo.3923956

3.3 Storing and archiving

- Store your digital data securely, preferably in University research data storage or University Google drive. (https://students.sheffield.ac.uk/it-services/research/storage)
- **Backups**: 3-2-1 (3 copies, 2 different media, 1 off-site). This means that you should have 3 copies of your data on two different types of media with one copy off-site for disaster recovery.
- If you collect or create physical data, digitise them if possible. See *How to deal with non-digital data: the benefits of digitising data* from OpenAIRE: https://www.openaire.eu/non-digital-data-guide

  Planning for digitisation provided by UCL Library Service: https://www.ucl.ac.uk/library/digital-collections/digitisation
- You may need to make your own archives as online data may disappear. You might also need to re-digitise existing archive materials and (in doing so) create a new archive, as the existing digitalisations sometimes can be low quality (e.g., pixelated, missing pages, no metadata).
- What else can or should be archived:
  - Data with potential for reuse
  - Data that validates a publication
  - Data that must be archived because the funder requires it
  - Software and code
  - Data that can't be reproduced, once-only events, expensive processes
- Things to consider when archiving data:
  - Any legal or ethical obligations to delete or retain specific data
  - How long data needs to be archived (usually minimum 10 years)
  - Possible financial implications of long-term storage

4. Share (at the end of your research)

4.1 Benefits of sharing

Share research data to enable:

- Greater transparency and accountability of your research
- Greater impact and visibility of research
- Credit for research outputs
- Potential for increased citation rates

- New collaborations between data users and data creators
- The improvement and validation of research methods
- Cost savings by not duplicating data collection
- Promotion of innovation and potential new data uses
- Great resources for education, training, and Knowledge Exchange
  (Source: UK Data Service – https://ukdataservice.ac.uk/learning-hub/new-to-using-data/#why-share-research-data)

4.2 What to share?

- Share data that validates your research, especially if it has potential for reuse. Share software and code created to process data, or details of proprietary software used. See the guidelines for data sharing provided by the University Library: https://www.sheffield.ac.uk/library/rdm/publish
- Check if you have **permission to share** the data. 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.
- Observe all terms of **participant consent** regarding data sharing. Make sure data are fully anonymised - with direct and indirect identifiers removed, if this is a condition of sharing. Decide if your dataset will need to be restricted or embargoed, e.g., sensitive data or extremely large files:
  - Can you share some of your data as fully anonymised analysed or sample data through repositories?
- When sharing data outside the UK, check if Export Control Legislation applies: https://staff.sheffield.ac.uk/research-services/about/contact/export-control
- Discuss options for data sharing with external research partners.

4.3 Prepare your data to share

- Transfer your data to **an open or more widely accessible format** if they are in a specialised or proprietary format:
  - Recommended formats by the UK Data Service: https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/

- Check out **the existing data formats and standards** in your research area. For example:
  - Cross-Linguistic Data Formats (CLDF): https://cldf.clld.org/
  - Recommended data formats by CLARIN: https://repository.clarin.dk/repository/xmlui/page/formats
- Place **a data availability statement** in these and other publications, including a DOI[2] for your data where possible, or contact details for access requests.
  - The University Library has a contract with the DataCite consortium, through the British Library, to enable us to allocate DOIs. If you are building or running a data archive and would like to add a DOI, please contact rdm@sheffield.ac.uk.
- Select **an appropriate licence or conditions** for reuse of your data and software.
  - What and why? – A licence is "a mechanism through which the owner of a copyright can authorise other parties to make use of a work". When used properly, they "offer an extremely low risk and unambiguous way to use a work". See *Licences* from the University Library: https://www.sheffield.ac.uk/library/copyright/licences
  - Licence recommended by the University Library, for example:
    Creative Commons Licence Terms: including 6 different licence types, which become active simply by stating them clearly alongside your data, see: https://creativecommons.org/about/cclicenses
  - **Guidance on choosing a licence:**
    How to License Research Data: https://www.dcc.ac.uk/guidance/how-guides/license-research-data
    Licensing your data: https://dmeg.cessda.eu/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives/Licensing-your-data
    Choose an open source licence: https://choosealicense.com/
    EUDAT provides a wizard to help you choose an appropriate licence: https://ufal.github.io/public-license-selector/

4.4 Where and how to share?

- Share data openly through **archives, repositories, and databases** (see Section 5). Check whether individual repositories charge costs for depositing data, and if these will be covered by your funding.

[2] A Digital Object Identifier (DOI) is "a string of numbers, letters and symbols used to permanently identify an article or document and link to it on the web. A DOI will help your reader easily locate a document from your citation" (Source: https://researchguides.uic.edu/doi, retrieved June 2022).

- How to choose a repository, see the guide from the Digital Curation Centre: https://www.dcc.ac.uk/guidance/how-guides/where-keep-research-data
- Guidance to authors on data sharing and research data policy offered by publishers: Oxford University Press (OUP): https://academic.oup.com/journals/pages/authors/preparing_your_manuscript/research-data-policy?login=false
  Cambridge University Press (CUP): https://www.cambridge.org/authorhub/resources/guides
  Elsevier: https://www.elsevier.com/authors/tools-and-resources/research-data/data-base-linking and https://www.elsevier.com/about/policies/research-data
  Springer: https://www.springeropen.com/get-published/editorial-policies
  Taylor & Francis: https://authorservices.taylorandfrancis.com/data-sharing/share-your-data/repositories/
  FAIRsharing.org provides a database of standards, policies, and databases: https://fairsharing.org/

- If you have used data in your research that is publicly and permanently available, share a link rather than sharing the actual data.
- Store and share data for a minimum of 10 years after the end of the project, or in line with funder requirements.
- Make any delay to the release of data as short as possible and within funder requirements.
- Make access arrangements for physical data if they are important for the validation or reproduction of your research and cannot be digitised.

5. Resources[3] (in alphabetical order)

5.1 Archives, Repositories, and Databases

(Language and Linguistics:)

AILLA – Archive of the indigenous languages of Latin America, a digital archive of recordings and texts in and about the indigenous languages of Latin America.
https://www.ailla.utexas.org/

CHILDES – Child Language Data Exchange System. https://childes.talkbank.org/

[3] All links in this section were last accessed in June 2022.

CLARIN – Depositing data and data tools to make sure language resources (e.g., corpora, lexica, grammars) can be archived and made available to the community reliably and sustainably. https://www.clarin.eu/content/depositing-services

CLICS – database of cross-linguistic colexifications. http://clics.lingpy.org/

ComparaLex – an online lexical database developed by the Canada Institute of Linguistics. http://comparalex.org/

Early English Books Online (EEBO) TCP – a partnership with ProQuest and with more than 150 libraries to generate highly accurate, fully-searchable, SGML/XML-encoded texts corresponding to books from the EEBO Database.
https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/

EUROM – a spoken language resource for the EU with comparable speech recordings available in 7 different European languages. https://www.phon.ucl.ac.uk/shop/eurom1.php

IDEA – International Dialects of English Archive, recordings of native speakers of English from various parts of the world, and English spoken in various non-native accents. https://www.dialectsarchive.com/

IRIS – a digital repository of data collection instruments for research into second language learning and teaching. https://www.iris-database.org/iris/app/home/index

Language Goldmine – a list of linguistic databases and datasets.
http://languagegoldmine.com/

LingBuzz – an article archive and a community space for Linguistics.
https://ling.auf.net/lingbuzz

Linguistic Data Consortium (LDC) – is an open consortium of universities, libraries, corporations, and government research laboratories that creates and distributes a wide array of language resources including corpora. https://catalog.ldc.upenn.edu/

LSA – an online repository of semantics articles. https://www.linguisticsociety.org/

OLAC – this catalogue provides access to information about thousands of languages, including details of text collections, audio recordings, dictionaries, and software, sourced from digital and traditional archives. http://dla.library.upenn.edu/dla/olac/index.html

Oxford Text Archive – a repository of full-text literary and linguistic resources.
https://ota.bodleian.ox.ac.uk/repository/xmlui/

PHOIBLE – a repository of cross-linguistic phonological inventory data. http://phoible.org

Speech Error Database – for Psycholinguistics
https://www.mpi.nl/dbmpi/sedb/sperco_form4.pl

TalkBank – Shared databases of recordings and coded transcripts within subfields studying communication, including aphasia, audiology, bilingualism, CHILDES, conversational analysis, dementia, phonological and phonetic analysis, second language acquisition, and traumatic brain injury. https://www.talkbank.org/

The English Lexicon Project – a database containing a variety of lexical characteristics and experimental measurement data for over 40,000 English words. https://elexicon.wustl.edu/

The Rosetta Project – a global collaboration of language specialists and native speakers working to build a publicly accessible digital library of human languages.
https://rosettaproject.org/

TROLLing – an archive of linguistic data and statistical code.
http://site.uit.no/trolling/about/

UCL Dysfluency Database – recordings of 61 speakers who have been studied by the UCL Psychology department speech group as part of their research into stammering.
https://www.phon.ucl.ac.uk/shop/ucldysfluency.php

UCL Speech Data database – recordings of a wide range of speech materials for 45 speakers of South-Eastern British English. https://www.phon.ucl.ac.uk/shop/uclspeaker.php

UCLA Phonetics Lab Archive – recordings of hundreds of languages and source materials for phonetic and phonological research. http://archive.phonetics.ucla.edu/

WALS Online – World Atlas of Language Structures Online is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors. https://wals.info/

(General Arts & Humanities:)

ADP – Social science data archives. https://www.adp.fdv.uni-lj.si/

Europeana – a web portal created by the European Union containing digitised cultural heritage collections of more than 3,000 institutions across Europe.
https://www.europeana.eu/en

Internet Archive – a digital library of Internet sites and other cultural artifacts in digital form.
https://archive.org/

ORDA – the hub for managing and sharing research data at The University of Sheffield:
https://orda.shef.ac.uk/ with instructions: https://www.sheffield.ac.uk/library/rdm/orda

Qualitative Data Repository – an archive for storing and sharing digital data and accompanying documentation generated or collected through qualitative and multi-method research in the social sciences and related disciplines. https://qdr.syr.edu/

Registry for Research Data Repositories – a comprehensive registry of research data repositories from different academic disciplines including Linguistics.
https://www.re3data.org/

5.2 Data tools

AntConc – a freeware corpus analysis toolkit for concordancing and text analysis.
https://www.laurenceanthony.net/software/antconc/

Audiamus – a tool for building corpora of linked transcripts and digitised media.
https://www.nthieberger.net/audiamus.htm

CorpusSearch2 – a tool for linguistic research. http://corpussearch.sourceforge.net/

DataCite – it gathers metadata for each DOI assigned to a research object.
https://search.datacite.org/

ELAN – an annotation tool for audio and video recordings. https://archive.mpi.nl/tla/elan

Gorilla Experiment Builder – a cloud-based research platform that allows researchers and students to quickly and easily create and deploy behavioural (reaction-time) experiments online. The School of English has a departmental licence for Gorilla, please contact Dr. Robyn Orfitelli (r.orfitelli@sheffield.ac.uk) for a subscription. Publications that cite Gorilla can also request to be added to the list on their website to make them more discoverable.
https://app.gorilla.sc/

LaTeX – a software system for document preparation. https://www.latex-project.org/

Nvivo – a qualitative data analysis computer software, which can be used for text markup and analysis. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Optical Character Recognition (ORC) in Adobe Acrobat – it helps you extract text and convert scanned documents into editable, searchable PDF files.

Phon – a software program that supports the building of textual and phonological data corpora.
https://www.phon.ca/phon-manual/getting_started.html

Praat – a free computer software package for speech analysis in phonetics.
https://www.fon.hum.uva.nl/praat/

R – a free software environment for statistical computing and graphics.
https://www.r-project.org/

SPSS – a statistical software platform. https://www.ibm.com/products/spss-statistics

Stanford Parser – A natural language parser is a program that works out the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb.
https://nlp.stanford.edu/software/lex-parser.shtml#About

Treeform – syntax tree drawing software. https://sourceforge.net/projects/treeform/

Universal Dependencies – a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages.
https://universaldependencies.org/

5.3 Corpora

AMC/IHD Corpora and Atlases – a list of linguistic atlases and corpora produced at the Institute for Historical Dialectology. https://www.amc.lel.ed.ac.uk/amc-projects-hub/project/

BASE – British Academic Spoken English corpus
https://warwick.ac.uk/fac/soc/al/research/collections/base/

BAWE – British Academic Written English Corpus
https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/

British National Corpus – Information about the BNC, and links to other English corpora sites. http://www.natcorp.ox.ac.uk/

CEEC – Corpus of Early English Correspondence
https://varieng.helsinki.fi/series/volumes/14/nevala_nurmi/

COHA – the Corpus of Historical American English (COHA) is the largest structured corpus of historical English. https://www.english-corpora.org/coha/

Corpora: Legal Language – Information on and links to corpora made up of legal texts.
http://korpus.uib.no/icame/corpora/1998-4/0160.html

Corpus of late 18c Prose – 300,000 words of north-western English letters on practical subjects (1761-89), collected by the University of Manchester.
https://personalpages.manchester.ac.uk/staff/david.denison/late18c

EUSTACE – Edinburgh University Speech Timing Archive and Corpus of English: 4608 sentences of spoken English provided online by Edinburgh's Centre for Speech Technology Research. https://www.cstr.ed.ac.uk/projects/eustace/

Penn Parsed Corpora of Historical English – texts and text samples of British English prose across its history - from the earliest Middle English documents up to the First World War.

SCRIBE – a pilot corpus of spoken British English produced in a collaboration between UCL, Cambridge University, and Edinburgh University.
https://www.phon.ucl.ac.uk/resource/scribe/

Salamanca Corpus – a digital archive of English dialect texts.
http://www.thesalamancacorpus.com/

SCOTS – Scottish Corpus of Texts & Speech. https://www.scottishcorpus.ac.uk/

The IViE corpus – contains recordings of nine urban dialects of English spoken in the British Isles. Recordings of male and female speakers were made in London, Cambridge, Cardiff, Liverpool, Bradford, Leeds, Newcastle, Belfast in Northern Ireland, and Dublin in the Republic of Ireland. http://www.phon.ox.ac.uk/files/apps/IViE/

UCREL Corpus Holding – A large selection of links to corpora of written and spoken languages (chiefly English). https://ucrel.lancs.ac.uk/corpora.html

WebCorp – a suite of tools that allows access to the World Wide Web as a corpus.
https://www.webcorp.org.uk/live/

W3 Corpora – Web access to linguistic corpora provided by the University of Essex (no longer maintained). https://www1.essex.ac.uk/linguistics/external/clmt/w3c/

York-Toronto-Helsinki Parsed Corpus of Old English – a selection of poetic texts from the Old English Section of the Helsinki Corpus of English Texts (henceforth the Helsinki Corpus), annotated to facilitate searches on lexical items and syntactic structure.