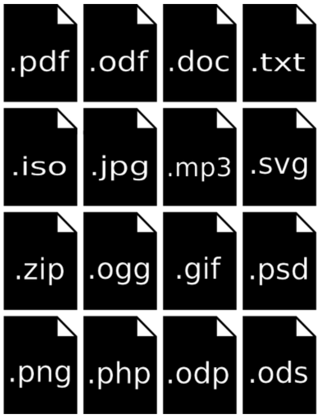
File formats for research data



In planning a research project, it is important to consider which file formats you will use to capture and store your data. In some cases, this will be dictated by the software you are using or the conventions of your discipline. In other cases, you may have to make a choice between several options.

The file format you choose may affect who you can share your data with and how easy your data will be to access in the future. Ideally, a format that is open and sustainable will be available for your data; however it may be necessary to use a proprietary format depending on the equipment or software you are using while you gather and analyse data. You could consider converting to an open or sustainable format when you share with collaborators or at the end of the project. In some cases, it may be appropriate to preserve the original file format, with either a copy of the software or a note of the software version, along with the open and sustainable format.

Factors to consider

- Which software and formats you or your colleagues have used in the past
- Which discipline-specific norms and/or peer support which are available in your field
- Which software is compatible with hardware you already have
- Whether you have funding for new software or equipment
- Your plans for data analysis and storage
- Which formats will be most useful for data sharing and for future reuse (your intended repository may require specific formats)
- Which formats are at risk of obsolescence, either due to lack of compatibility or dependence on specific software
- Which formats will be easiest to annotate with metadata
- All other factors being equal, try to choose a format which is non-proprietary unencrypted and uncompressed

In some cases, it may be necessary or desirable to use one format for data collection, manipulation and analysis, but convert your data to another format for deposit in a repository once your project is completed.

When it is necessary to save files in a proprietary format, consider including a readme.txt file in your directory that documents the name and version of the software used to generate the file, as well as the company who made the software. This could help you down the road if you need to figure out how to open these files again.

Formats for preservation

File formats can affect long-term preservation and reuse. While researchers may use proprietary file formats for analysis, converting data to open and/or standard formats will help ensure the data can be rendered and accessed in the future. Researchers can also choose to make data available in both preservation-friendly formats and original file formats.

Best practice suggests selecting formats that are open/documented standards, non-proprietary, unencrypted, uncompressed, and commonly used by your research community. For example, when you have spreadsheet-based (aka tabular) data save the file as Comma-separated values (.csv) instead of Excel (.xls, .xlsx), unless you are using functionality in the spreadsheet which is only available to .xls(x). For text files, use Plain text (.txt) or PDF/A (.pdf) instead of Microsoft Word (.doc, .docx).

Repositories may provide a list of preferred file formats.

Common (preferred) file formats

TYPE OF DATA	PREFERRED FILE FORMATS FOR SHARING, RE-USE AND PRESERVATION	Other Acceptable formats
Quantitative tabular data with extensive metadata • a dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data	• SPSS portable format (.por) • delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information • structured text or mark-up file containing metadata information, e.g. DDI XML file	• MS Access (.mdb/.accdb)
Quantitative tabular data with minimal metadata	• comma-separated values (CSV) file (.csv) • tab-delimited file (.tab)	• delimited text of given character set -- only characters not present

- a matrix of data with or without column headings or variable names, but no other metadata or labelling	- including delimited text of given character set with SQL data definition statements where appropriate	in the data should be used as delimiters (.txt) - widely-used formats, e.g. MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf) and OpenDocument Spreadsheet (.ods)
Geospatial data vector and raster data	- ESRI Shapefile - (essential: .shp, .shx, .dbf ; optional: .prj, .sbx, .sbn) - geo-referenced TIFF (.tif, .tfw) - CAD data (.dwg) - tabular GIS attribute data	- ESRI Geodatabase format (.mdb) - MapInfo Interchange Format (.mif) for vector data
Qualitative data textual	- Rich Text Format (.rtf) - plain text data, UTF-8 (Unicode; .txt) - eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml)	- plain text data, ASCII (.txt) - widely-used proprietary formats, e.g. MS Word (.doc/.docx) - LaTeX (.tex) - Hypertext Mark-up Language (HTML) (.html)
Digital image data	- TIFF version 6 uncompressed (.tif)	- JPEG (.jpeg, .jpg) - TIFF (other versions; .tif, .tiff) - JPEG 2000 (.jp2) - Adobe Portable Document Format (PDF/A, PDF) (.pdf)

Digital audio data	- Free Lossless Audio Codec (FLAC) (.flac) - Waveform Audio Format (WAV) (.wav) - MPEG-1 Audio Layer 3 (.mp3) - spoken word audio only	- MPEG-1 Audio Layer 3 (.mp3) - Audio Interchange File Format (AIFF) (.aif)
Digital video data	- MPEG-4 High Profile (.mp4) - motion JPEG 2000 (.jp2)	- JPEG 2000 (.mj2)
Documentation & Scripts	- Rich Text Format (.rtf) - Open Document Text (.odt) - HTML (.htm, .html)	- plain text (.txt) - widely-used proprietary formats, e.g. MS Word (.doc/.docx) or MS Excel (.xls/ .xlsx) - XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0 - PDF/A or PDF (.pdf)
Chemistry data spectroscopy data and other plots which require the capability of representing contours as well as peak position and intensity	- Convert NMR, IR, Raman, UV and Mass Spectrometry files to JCAMP format for ease in sharing. - JCAMP file viewers: JSpecView, ChemDoodle	

Useful resources

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

University of Glasgow
RDM Service September 2021

Library of Congress Recommended formats statement:
https://www.loc.gov/preservation/resources/rfs/

Dryad guidance on preferred file formats for deposit:
https://datadryad.org/stash/best_practices

Stanford Libraries Case study on file formats:
https://library.stanford.edu/research/data-management-services/case-studies/case-study-file-formats

University of Edinburgh DataShare – Recommended file formats guide:
https://www.ed.ac.uk/files/atoms/files/recommended_file_formats-apr2015.pdf

Sources of information used in the compilation of this guide

University of Cambridge Research Data Management Team
University of Edinburgh DataShare
UK Data Service
Duke University Libraries
Princeton University Library
Oregon State University Libraries
University of Washington Libraries
Stanford Libraries