# Data Management Handbook

## *Project name*

Table of Contents

[How to use this handbook 2](#_Toc68786070)

[Introduction 3](#_Toc68786071)

[Data Management Plan 4](#_Toc68786072)

[Data collection 5](#_Toc68786073)

[1. Data storage 5](#_Toc68786074)

[2. Folder structure 5](#_Toc68786075)

[3. File naming 6](#_Toc68786076)

[4. Versioning 6](#_Toc68786077)

[5. File formats 7](#_Toc68786078)

[6. Quality assurance 7](#_Toc68786079)

[7. Metadata/documentation 8](#_Toc68786080)

[Data processing 9](#_Toc68786081)

[1. Processing personal data 9](#_Toc68786082)

[2. Transcription 9](#_Toc68786083)

[3. Anonymisation/pseudonymisation 9](#_Toc68786084)

[4. Data cleaning 9](#_Toc68786085)

[Data sharing and archiving 10](#_Toc68786086)

[1. Selection and appraisal 10](#_Toc68786087)

[2. File formats 10](#_Toc68786088)

[3. Metadata and documentation 11](#_Toc68786089)

[4. Licences 12](#_Toc68786090)

## How to use this handbook

This handbook template has been designed to assist PIs in creating a handbook for research data management to share with research teams at the beginning of a project to ensure that data management procedures are agreed, understood and adhered to throughout the project.

* All text in italics should be replaced with project specific information.
* Text boxes contain further information which may prove useful when completing the handbook.
* Not all sections will apply to all projects and should only be completed as necessary.

The Library Research Support Team are available to help with completing this template and to provide project-specific guidance and training, please feel free to contact us at [library-research-support@open.ac.uk](mailto:library-research-support@open.ac.uk)

## Introduction

* *A brief pre-amble from the PI about the importance of Data Management, including funder specific requirements.*

In keeping with OU principles of openness, it is expected that research data will be open and accessible to other researchers, as soon as appropriate and verifiable, subject to the application of appropriate safeguards relating to the sensitivity of the data and legal and commercial requirements. Research data must be managed to the highest standards throughout their lifecycle, in order to support excellence in research practice

* *Roles and responsibilities of the team members*

Indicate who within your research team will be responsible for data management, metadata production, dealing with quality issues and the final delivery of data for sharing or archiving (and any other tasks related to data management)

## Data Management Plan

*Include your Data Management Plan here*

## Data collection

### Data storage

*Include details of your chosen storage solution.*

*When choosing storage consider:*

* *Who needs access?*
* *Do you need to restrict access to certain files?*
* *Are the data personal/sensitive? (see* [*our guidance on Data Protection*](http://www.open.ac.uk/library-research-support/research-data-management/ethics-and-data-protection)*)*

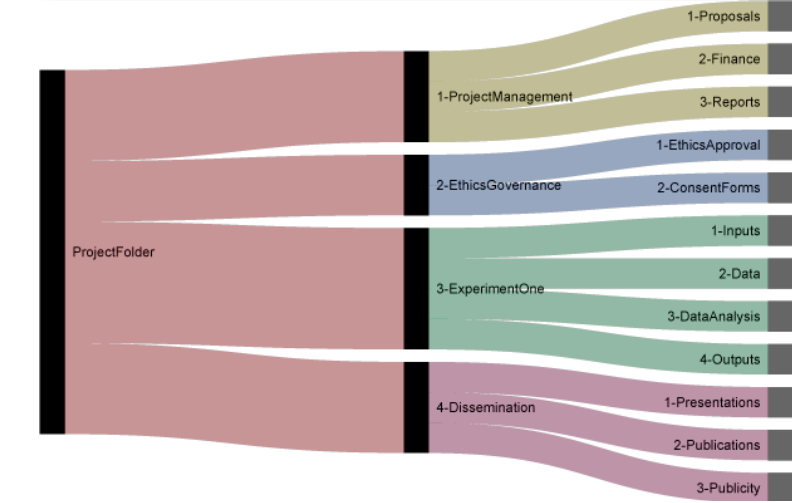
*Consult* [*our comparison table of data storage for projects*](http://www.open.ac.uk/library-research-support/sites/www.open.ac.uk.library-research-support/files/files/RDM-data-storage-options-AccessibleWord.docx) *to help choose storage.*

### Folder structure

***I****nclude details of your chosen folder structure.*

**Tips for creating a folder directory:**

* Apply logical structuring of files within folders relating to projects or issues. Don’t leave files unsorted, hanging under top level folders
* Design a hierarchy with higher level broader topics, with more specific folders within these.  Avoid very tall, and/or labyrinthine structures
* Folders should be named after projects and research issues, with clear meaning.  Avoid names which are meaningless, are excessively long, or relate to individuals
* It may help to separate current and completed work or versions, e.g. where a document will have many versions and multiple contributors consider a “current version” folder
* You may want to follow the example below (from <http://nikola.me/folder_structure.html>):



[Download example template](http://nikola.me/assets/Folder_Template.zip)

### File naming

*Specify your file naming convention.*

**File naming good practice**

* Files should be named consistently
* File names should be short but descriptive (<25 characters)
* Avoid special characters or spaces in a file name
* Use date format: YYYYMMDD (this ensures files appear in chronological order)
* Include a version number (see below)

Example: 20210104\_ProjectA\_MeetingNotes\_SmithE\_v2.docx

### Versioning

*Include details of version control procedures.*

**Example of good practice – version control**

1. Save the document according to file naming guidance/good practice.
2. Create Document/File. Identify on the document e.g. in header or footer, the author, filename, page number and date the document is created/revised.
3. Document Identification. Versions and changes should be documented in a Version Control Table where significant/formal/project based.
4. Version Control Table.
   1. Current version number identified on the first page and where appropriate, incorporated into the header or footer of the document.
   2. Version number is included as part of the file name.
5. Version Number
   1. Named as version “01”
   2. Subsequent draft versions 02, 03, 04 …
6. First Draft Version
   1. When document is final/approved it becomes version 1.0
7. First Final/Approved Version
   1. Changed/revised final version becomes 1.1
   2. Subsequent drafts to Final version become e.g. 1.2, 1.3 etc.
8. Changes to Final Version
   1. Version number increased by “1.0” e.g. 1.0, 2.0, 3.0 etc.

### File formats

*Specify which file formats you will be saving your data in. This will ensure a consistent approach across the team/project. It is prudent at this point to think ahead to formats for preservation – see the section on Data Sharing and Archiving in this handbook for guidance.*

### Quality assurance

*Detail the processes the team will undertake to ensure data quality*

Data quality processes can include:

* calibration of instruments to check the precision, bias and/or scale of measurement
* taking multiple measurements, observations or samples
* checking the truth of the record with an expert
* using standardised methods and protocols for capturing observations, alongside recording forms with clear instructions
* computer-assisted interview software to: standardise interviews, verify response consistency, route and customise questions so that only appropriate questions are asked, confirm responses against previous answers where appropriate and detect inadmissible responses

**Digitisation and data entry: quality assurance**

When data are digitised, transcribed, entered in a database or spreadsheet, or coded, quality can be ensured by standardised and consistent procedures for data entry with clear instructions.

This may include:

* setting up validation rules or input masks in data entry software
* using data entry screens
* using controlled vocabularies, code lists and choice lists to minimise manual data entry
* detailed labelling of variable and record names to avoid confusion
* designing a purpose-built database structure to organise data and data files
* accompanying notes and documentation about the data

### Metadata/documentation

*Define which metadata should be captured during data collection and how this should be stored.*

**Metadata should be captured at study level and data level.**

**Study level metadata** usually takes the form of a README file and should include:

* The aims of your project
* The methods used to collect data
* The contents of your data
* The folder structure and file naming conventions
* The data processing techniques used
* The modifications made to the initial data throughout the project
* Data validation and other quality assurance processes
* Roles and responsibilities within the project
* Details on identifiers, licensing, and sensitive information

Note that much of this information will be captured in this handbook.

**Data level metadata** is usually embedded within data files (e.g. as a header, summary page, code values or in the document’s properties) or could saved in a data table/list and may include (depending on the data type):

* variable names, labels and descriptions (maximum 80 characters)
* units of measurement for variables
* reference to the question number of a survey or questionnaire
* information relating to the date/time/location/participant ID

For more information, consult the comprehensive [guidance on metadata and documentation](https://www.ukdataservice.ac.uk/manage-data/document.aspx) on the UK Data Service website, including a useful [Data List Template](https://ukdataservice.ac.uk/media/622362/uk_data_archive_data_listing_template.xlsx)

## Data processing

### Processing personal data

*Where applicable, include information about procedures team members need to undertake to comply with GDPR (or the relevant Data Protection legislation for your project).*

Actions required to comply with GDPR may include:

* Documenting consent
* Ensuring data are stored securely
* Encrypting personal data before saving or transferring
* Deleting personal data when it is no longer required
* Anonymising or pseudonymising data (see below)
* Restricting access to some files to certain members of the team
* Understanding and acting upon the participants’ right to withdraw

For more information, see the [OU’s Guide to GDPR and research data](http://www.open.ac.uk/library-research-support/sites/www.open.ac.uk.library-research-support/files/files/Guide_RDM_GDPR-HowDoesThisAffectRDM.pdf)

### Transcription

*Where applicable, include details of the processes for transcribing data.*

*If this involves using an external company, include how documents should be securely submitted.*

Further [guidance on creating quality transcripts](https://www.ukdataservice.ac.uk/manage-data/format/transcription.aspx) is available on the UK Data Archive website.

### Anonymisation/pseudonymisation

*Where applicable, describe the process for anonymising/pseudonymising personal data.*

Consult [guidance on anonymising qualitative and quantitative data](https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation.aspx) on the UK Data Service website.

### Data cleaning

*Describe any processes which will be undertaken to clean data, including any software which will be used to do this.*

Options for data cleaning tools include:

* [QAMyData](https://www.ukdataservice.ac.uk/about-us/our-rd/qamydata.aspx) (good for social sciences)
* [OpenRefine](https://openrefine.org/) (good for big, messy data)

## Data sharing and archiving

### Selection and appraisal

*Define which data will be deposited in a data repository (or how they will be selected if this is unknown at this point)*.

For guidance, refer to the [Archiving Data page](http://www.open.ac.uk/library-research-support/research-data-management/archiving-data) on the OU Library Research Support website.

### File formats

Provide a list of the file formats that data which will be deposited should be saved in.

*Below is a table of recommended formats for digital preservation (*[*from UK Data Service*](https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats)*)*

|  |  |  |
| --- | --- | --- |
| **Type of data** | **Recommended formats** | **Acceptable formats** |
| **Tabular data with extensive metadata**  variable labels, code labels, and defined missing values | SPSS portable format (.por)  delimited text and command ('setup') file (SPSS, Stata, SAS, etc.)  structured text or mark-up file of metadata information, e.g. DDI XML file | proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb) |
| **Tabular data with minimal metadata**  column headings, variable names | comma-separated values (.csv)  tab-delimited file (.tab)  delimited text with SQL data definition statements | delimited text (.txt) with characters not present in data used as delimiters  widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods) |
| **Geospatial data**  vector and raster data | ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx, .sbn optional)  geo-referenced TIFF (.tif, .tfw)  CAD data (.dwg)  tabular GIS attribute data  Geography Markup Language (.gml) | ESRI Geodatabase format (.mdb)  MapInfo Interchange Format (.mif) for vector data  Keyhole Mark-up Language (.kml)  Adobe Illustrator (.ai), CAD data (.dxf or .svg)  binary formats of GIS and CAD packages |
| **Textual data** | Rich Text Format (.rtf)  plain text, ASCII (.txt)  eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema | Hypertext Mark-up Language (.html)  widely-used formats: MS Word (.doc/.docx)  some software-specific formats: NUD\*IST, NVivo and ATLAS.ti |
| **Image data** | TIFF 6.0 uncompressed (.tif) | JPEG (.jpeg, .jpg, .jp2) if original created in this format  GIF (.gif)  TIFF other versions (.tif, .tiff)  RAW image format (.raw)  Photoshop files (.psd)  BMP (.bmp)  PNG (.png)  Adobe Portable Document Format (PDF/A, PDF) (.pdf) |
| **Audio data** | Free Lossless Audio Codec (FLAC) (.flac) | MPEG-1 Audio Layer 3 (.mp3) if original created in this format  Audio Interchange File Format (.aif)  Waveform Audio Format (.wav) |
| **Video data** | MPEG-4 (.mp4)  OGG video (.ogv, .ogg)  motion JPEG 2000 (.mj2) | AVCHD video (.avchd) |
| **Documentation and scripts** | Rich Text Format (.rtf)  PDF/UA, PDF/A or PDF (.pdf)  XHTML or HTML (.xhtml, .htm)  OpenDocument Text (.odt) | plain text (.txt)  widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx)  XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0 |

### Metadata and documentation

*Describe which metadata will be deposited alongside the data in your chosen repository.*

See the OU Library’s [guide to creating a README file for research data.](http://www.open.ac.uk/library-research-support/research-data-management/research-data-management-guide-writing-%E2%80%9Creadme%E2%80%9D-type-metadata)

### Licences

*Give details of which licence(s) will be applied to the data you share in a repository*.

Usually you will use Creative Commons licences for research data, although there are other options, for more information see the Library’s page on [licensing research data](http://www.open.ac.uk/library-research-support/research-data-management/licensing-research-data).

Below is a table detailing the various “flavours” of Creative Commons licences:

