Prepare your data

site.uit.no/dataverseno/deposit/prepare/

Before depositing your data in DataverseNO (including the different sub-archives, e.g. UiT Open Research Data, TROLLing, etc.) you have to make sure your dataset(s) comply with our guidelines below. DataverseNO accepts only research data in digital formats. In brief, good practice for preparing research data for archiving may be summarized as follows:

- Use consistent and comprehensible file names (see section 1 below).
- Save your data in a preferred file format(s) (see section 2 below).
- Describe your data in a ReadMe file (see section 3 below).

For more detailed guidelines, see below:

1 File naming

Following good practice for file naming and organizing makes it much easier to find the right data file, not just for you, but also for your collaborators, and later on for other researchers who may re-use your data. Please make sure your file names comply with the following fundamental file naming recommendations:

- Files must be named consistently.
- File names must be descriptive, but short (< 25 characters).
- Do not use spaces. Instead, use underscores (e.g. first_study), hyphens (e.g. first-study) or camel case (FirstStudy).
- Avoid characters like \ / ? : * ” > < | : # % ” { } | ^ [ ] ` ~ æÆ øØ åÅ äÄ öÖ ...
- Use the international dating convention YYYY-MM-DD (e.g. 2017-10-25).
- The name of a file in original file format must be identical with the name of the corresponding file in preferred file format (see below).

2 Preferred file formats

What are preferred file formats?

The choice of a preferred file format is crucial in order to ensure that your data will be readable also in the future. Some file formats are more likely to allow long-term readability than others are. Such formats are usually

- non-proprietary
- open, with documented international standards
- using standard character encoding, preferably Unicode (e.g. UTF-8)

- uncompressed (space permitting)

The table below gives an overview of preferred vs. non-preferred file formats for **a selection of** document types. The list of file formats in the column "Non-preferred file formats" is non-exhaustive and includes the formats considered the ones used most commonly. If your dataset contains file formats not listed here, please contact the support services of your home institution. When uploading your data to the archive, please make sure you add your files in a preferred format. Make also sure that all of your files contain a valid file extension, e.g. .txt, .pdf. If your data cannot be stored in a preferred format, they can still be published in their original format, but in that case, DataverseNO does not commit to preserve the data in the long term. If appropriate, the file may also be archived in their original file format **in addition to** preferred format(s).

File type	Preferred file formats (examples)	Non-preferred file formats (examples)
Audio	• Uncompressed and lossless Wav or AIFF (.wav/.aiff) • Compressed and lossless FLAC (.flac) • Compressed and lossy Mp3 (.mp3)	• AAC (.m4a) • Monkey's Audio (.ape) • Ogg Vorbis (.ogg) • Windows Media Audio (.wma)
Container file	Container files are automatically unpacked when uploaded and should only be used to keep the folder structure in your dataset; see more in section Upload data files.	In case container files need to be archived as container files, use .zip. Note! In this case, files must be packed twice. That way, the inner container will be preserved when uploaded.
Image	• Uncompressed TIFF (.tif or .tiff) • Compressed and lossless PNG (.png) • Compressed and lossy JPEG (.jpg)	• Adobe Photoshop (.psd) • Apple Picture File (.pct) • Graphics Interchange Format (.gif) • Raw Image Data File (.raw) • Windows Bitmap (.bmp)
Text (slides, illustrations)	PDF/A (.pdf) combined with original file	PowerPoint (.pptx)
Text (tables)	Tab separated Unicode plain text (.txt)	Excel (.xlsx)

File type	Preferred file formats (examples)	Non-preferred file formats (examples)
Text (text)	Plain text (.txt) **If formatting needed:** XML, PDF/A (.pdf) combined with original file	• Word (.docx) • HTML
Markup language	• XML (.xml) • HTML (.html) • Related files: .css, .xslt, .js, .es	• SGML (.sgml) • Markdown (.md)
Transcription	**File format:** • PDF/A (.pdf) combined with original file • PDF/A (.pdf) combined with Comma/Tab Separated Values (.csv/.txt) **Font:** Unicode IPA (e.g. Charis SIL, Doulos SIL, Gentium Plus, Andika), ASCII SAMPA	**File format:** • Word (.doxc) • Excel (.xlsx) **Font:** Transcription legacy fonts (SIL IPA(93))
Video	MPEG-4 (.mp4)	• AVI (.avi) • Flash Video (FLV) • Quicktime (.mov) • Windows Media Video (WMV)
Statistical analysis	• R (.R, .RData) • SPSS (.dat/.sps) • STATA (.dat/.DO)	• SPSS Portable (.por) • SPSS (.sav) • STATA (.dta) • SAS (.7dat, .sd2, .tpt)
Qualitative data analysis	• Basic data in preferred file format, e.g. PDF/A, plain text in Unicode (.txt) • Analysis dump/package as REFI-QDA Project (.qdpx)[1]	The different workspace dump formats, e.g. .nvp, .hpr

File type	Preferred file formats (examples)	Non-preferred file formats (examples)
Workspace dump formats for mass spectrometry	mzML (.mzML)[2]	• Agilent D (.D) • Bruker BAF (.BAF) • Bruker FID (.FID) • Chromtech DAT (.DAT) • …

[1 Read more about this format here.
[2] Read more about this format here.

How to save or convert your data into a preferred file format?

This section contains information on the following document types: Audio, container, image, text, transcription, and video. If your data contain types not listed here, please contact the support services of your home institution.

Audio

- **Recording:**
  The quality of your audio file depends on the purpose of your recording. If the recording is of such nature that acoustic details are irrelevant, the mp3 format is sufficient. Note however, that mp3 is a lossy compression format: Information in the speech signal is irreversibly discarded during recording and can therefore be considered less suited for speech analysis in the case of data reuse.Given that the mp3-format reduces the reusability of your data, we advise recording in an uncompressed format, .wav or .aiff.
- **Conversion:**
  If space is an issue, you can convert the uncompressed .wav and .aiff-files after recording. We recommend a format that does not remove information, like FLAC (Free Lossless Audio Codec). Conversion to FLAC is fully reversible, i.e. the original sound file is restored when decompressed.File conversion can easily be done in free software like Audacity (http://web.audacityteam.org/) or Praat (http://www.fon.hum.uva.nl/praat/).

Container files

Container files are automatically unpacked when uploaded and should only be used to keep the folder structure in your dataset; see more in section Upload data files. In case container files need to be archived as container files, use .zip. Note! In this case, files must be packed

twice. That way, the inner container will be preserved when uploaded. For packing, please follow the recommendations below:

- Use container files with extensions .zip (do not use .7z, tar.gz, .rar, and so on).
- Use one of the following tools to pack your files into a container:
  - 7-Zip (for Windows)
  - Keka (for Mac)
- Do not use compression or encryption when packing your files into containers.

Image

- **Compression:**
  Images are often compressed to reduce the amount of redundant or irrelevant data information. This does not mean that the quality reduction is visible to the human eye. For instance, PNG-files maintain all information in the image. As for JPEG-files – a widely used file format – the rate of compression can be manipulated: Depending on type of image and potential size issues, you must, in each case, determine how much compression is advisable, with regard to both reuse and sharing of your image files.
- **Conversion:**
  If your images are stored in a format considered non-preferred (see the section *What are preferred file formats?* above), they must be converted to JPEG, PNG or TIFF. Conversion can easily be done in the software Paint (Windows), Preview (Mac) or GIMP Image Editor (Linux). There are numerous free image converters.

Text

Plain text

If your data is represented in plain text, requiring little or no formatting, you are recommended to create and save your data as plain text files (.txt). You may use a simple text editor, e.g. gedit, TextEdit or WordPad. If you use a more advanced text editor when structuring your data, e.g. Microsoft Word or LibreOffice Writer, you must still save it in plain text format. To do so, select “Save as file type: Plain text (.txt)” in the menu *File* > *Save As*. Also, choose Unicode UTF-8 character encoding.

Formatted text

If your data contains formatted text, e.g. including essential line breaks, tabs, figures, we recommend you to convert your data file into a PDF/A file (.pdf). The original text file as well as the PDF/A file must be uploaded. The same procedure must be carried out if you use a text

editor like Microsoft Word or LibreOffice Writer when structuring your data, or a presentation editor like Microsoft PowerPoint or LibreOffice Impress.

To create a PDF/A file in Microsoft Word:
Mac (2011): *Print > PDF > Save as Adobe PDF > Adobe PDF Settings: PDF/A-1b: 2005 (CMYK)*. Note that this option requires Adobe Acrobat. If Adobe Acrobat is not available, save the file as plain PDF, and convert it using a tool like PDFTRON (see below).

Windows (2013): *Save as Adobe PDF > File type: PDF files > Options: Create PDF/A-1a: 2005 compatible file*

To create a PDF/A file in LibreOffice Writer:
Linux: Save as PDF > Check the PDF/A-1a box > Export.

To save/convert a PDF file as a PDF/A file in Adobe Acrobat (Pro or similar):
Save As Other > More Options > PDF/A.

To save/convert a PDF file as a PDF/A file in PDFTRON (eller similar):
Go to https://www.pdftron.com/pdf-tools/pdfa-converter/, scroll down to the *Drag and drop files* area, choose *PDF/A-1A* in field 1, and upload your PDF file in field 2.

Tabular text

Tabular text data must be provided as Unicode-encoded text files (.csv/.txt). If you have stored your data in a spreadsheet software like Microsoft Excel or LibreOffice Calc, the following instructions show you how to convert it to a recommended format:

Microsoft Excel (Mac, Windows):

- (On a laptop: Click **More options** below the file type field displaying Excel Workbook (*.xslx))
- Choose File > **Save as** > Choose folder
- In the option Save as type, choose **Text (Tab delimited) (*.txt)**
  (Note! Do **not** choose Unicode Text (*.txt))
- In Tools, choose **Web options**
- Choose the tab **Encoding**
- In the field Save this document as, choose **Unicode (UTF-8)**, and then click **OK**
- Choose the tab **Fonts**
- In the Character set window, choose **Multilingual/Unicode/Other script**, and click **OK**
- Click **Save**
- Confirm by clicking **Yes**
- Note: This process has to be repeated for each sheet in the Excel workbook

LibreOffice Calc (Linux, Mac, Windows):

- Click *File* > *Save As*
- For each sheet in the LibreOffice Calc workbook, proceed as follows:
    - Linux and Windows: In the data export dialogue window, select
        - Text encoding/Character set: Unicode (UTF-8)
        - Field delimiter: {Tabulator} (= recommended)
        - Text delimiter: none (erase the prefilled one from the field)
    - Mac: In the field *File type*, select "Text CSV (.csv)". In the data export dialogue window, select
        - Character set: Unicode (UTF-8)
        - Field delimiter: {Tab}
        - Text delimiter: " (double quotation mark)

If the very graphical layout of your tabular data is essential in order to understand them, you must also upload a PDF/A version of the document. Also, if your tabular text data contain figures, charts or other kinds of graphical elements that are essential for understanding your data, it is recommended that you convert these elements into PDF/A documents. See conversion procedure for formatted text above.

Transcription

- **Font:**
  All transcriptions must be made using Unicode-encoded fonts, e.g. IPA Doulos SIL.[1] For phonetic transcriptions, SAMPA (Speech Assessment Methods Phonetic Alphabet, ASCII characters)[2] is an alternative to IPA. If the recommended font is not available for the type of transcription your dataset requires, it is imperative to include a separate ReadMe file in your dataset with instructions about how to read the transcriptions.[3] Note that the font package itself must *not* be uploaded, given copyright restrictions.
  [1] To download SIL Fonts, cf. http://scripts.sil.org/cms/scripts/page.php?cat_id=FontDownloads.
  [2] For an overview of SAMPA symbols, cf. https://www.phon.ucl.ac.uk/home/sampa/.
  [3] Cf. for instance an example in the file "To read the Church Slavonic transcriptions.pdf" in Eckhoff (2015), cf. http://hdl.handle.net/10037.1/10190.
- **Conversion:**
  If your videos are stored in a format considered non-preferred (see the section *What are preferred file formats?* above), these must be converted to the MPEG-4 format. If you do not have license to any professional conversion software, we advise you to use the VLC Media Player (standard application on both Mac and Windows), or an online free image converter.

Workspace/analysis space

- **Statistical analysis software, e.g. Matlab, R, S-Plus, SPSS:**
  Most softwares for statistical analysis allow you to save the basic data as (or export them to) a plain text format (.txt). In addition, you must copy the script, and save it as plain text in a text editor.
- **Qualitative analysis software, e.g. ATLAS.ti, NVivo:**
  Some software packages for qualitative analysis allow you to save the basic data (or export them to) a preferred file format, e.g. PDF/A or plain text format (.txt). In addition, you can export the analysis package as a so-called REFI-QDA Project (.qdpx). In NVivo, this may be done in the following way: Click the menu tab *Share*, and then click *Export Project*. In the pop-up window, select *REFI-QDA Project*, and choose *Location*, i.e. where you want to save the file, and enter the filename.
- **Software for mass spectrometry:**
  Guidelines on how to convert .mid files to .mzML can be found here. If you are unfamiliar with the command line in Windows, please contact user support at your home institution.

3 How to describe your data

In order for other researchers to be able to understand and reuse your data, it is essential that you describe them in a comprehensible and consistent manner before they are published. In DataverseNO, this kind of documentation must be provided in two ways, in the **metadata fields**, and in a separate **ReadMe file** which must be uploaded together with your data files:

Metadata

Metadata is information about your data which makes them findable in discovery services. When creating a dataset, it is therefore important to fill in as much information as possible in the **metadata schema** (see the sections Enter metadata and Enter more metadata in the Deposit Guidelines.

ReadMe file

A **ReadMe file** is a more detailed user guide to your dataset so that other researchers are able to interpret, understand, and reuse your data, including information about how the dataset was created, how complete it is, and what kind of restrictions it has. The ReadMe file must minimally contain the following:

- Title of the dataset, DOI, contact information
- Methods

- Data and file overview
- Data-specific information
- Terms of Reuse

We recommend you building your ReadMe file based on this **general template**. For dataset containing software code or code-based data, you may use this **template for software code**.

The ReadMe file should be in plain text format with Unicode UTF-8 character encoding (.txt). If you need to illustrate or format your description, you may save your ReadMe file as PDF/A (see the section What are preferred file formats? for more information). We also recommend you to add "00_" in front of the ReadMe file name (e.g. "00_ReadMe.txt"), which will make the file appear on the top of the file overview.

Here are some sample ReadMe files: sample 1 (Social Sciences); sample 2 (Life Sciences).

4 File size

The size of each individual file upload must not exceed 8 Gb. If you want to upload files that are larger than 8 Gb in total, you have to upload them in several uploads. You do this by saving the dataset after each upload. As of today, there is **no upper limit** to the size of a dataset, but we recommend that you contact the support services of your home institution if you wish to add a dataset with a total file size of more than 50 Gb.

5 References

Parts of the guidelines above have been adapted from several sources, including

Data Management General Guidance. Curation Center of the California Digital Library, University of California. https://dmptool.org/dm_guidance#types.

Praat beginners' manual by Sidney Wood.
http://www.fon.hum.uva.nl/praat/manualsByOthers.html

Preparing tabular data for description and archiving. Research Data Management Group, Cornell University. http://data.research.cornell.edu/content/tabular-data.

Recommendations for uploading data. ETH-Bibliothek.
http://www.library.ethz.ch/en/content/download/17058/442689/version/2/file/Empfehlungen_Datenupload_en.pdf

Sustainable Formats and Conversion Strategies at the Bentley Historical Library. Version 1.0, November 9th, 2011.
http://bentley.umich.edu/dchome/resources/BHL_PreservationStrategies_v01.pdf.

For questions, comments or suggestions, see our support page.