UiT Open Research Dataset Guidelines

Contents:

I. Summary

Before uploading your data to UiT Open Research Data (including Arctic Language Technology, TROLLing, and other sub-archives) we urge you to make sure your dataset(s) comply with our guidelines below. In brief, good practice for preparing research datasets for archiving may be summarized as follows:

- Use consistent and comprehensible file names.
- Add your data in a persistent file format in addition to the original file(s).
- Describe your data in (a) ReadMe file(s).

II. 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 should be named consistently
- File names should be descriptive, but short (< 25 characters)
- Use underscores (_) instead of spaces
- Avoid characters like " / \ : * . ? ' < > [ ] ( ) & $ æÆ øØ åÅ äÄ öÖ ...
- Use the international dating convention YYYY-MM-DD

III. Persistent file formats

The choice of a persistent 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
- in common usage by the research community
- using standard character encodings, preferably Unicode (e.g. UTF-8)
- uncompressed (space permitting)

The table below gives an overview of persistent vs. non-persistent file formats for a selection of document types. When uploading your data to the archive, please make sure you add your files in a persistent format **in addition to** the original file format. Make also sure that all of your files contain a valid file extension, e.g. .txt, .pdf.

Table 1: Persistent vs. non-persistent file formats for various document types[1]

File type	Persistent file formats	Non-persistent 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 files	Container files are not recommended. If necessary, us the following formats: • zip • tar	• 7z • tar.gz • rar
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)
Text (text)	• Plain text (.txt) **If formatting needed** • XML, PDF/A (.pdf) combined with original file	• Word (.docx) • HTML
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)
Workspace dump for Matlab, R, S-Plus, SPSS or similar	Include • Basic data as tab separated Unicode plain text (.txt) • Script(s) as Unicode plain text (.txt)	• The different workspace dump formats, e.g. .mat, RData, .R

[1] The list of file formats in the column "Non-persistent 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 us at research-data@support.uit.no.

IV. Saving or converting your data into a consistent 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 us at research-data@support.uit.no.

A. Audio

1. Recording

The quality of your audio file depends on the purpose of your dataset. If the dataset 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.

2. 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/).

B. Container files

We do not recommend to use container files. By default, ZIP container files containing up to 1 000 files will be automatically unpacked when uploaded to the TROLLing archive. If you want to retain the original folder structure, you have to tag the files with the respective folder names. If you for some reasons have to use container files, please follow the recommendations below:

- Use container files with extensions .zip or .tar (do not use .7z, tar.gz, .rar, and so on). The tar format is preferred for long-term archiving because it is openly-documented.
- Use one of the following tools to pack your files into a container:
    - 7-Zip (for Windows)
    - Keka (for Mac, or use function “tar” on command line)
- Do not use compression or encryption when packing your files into containers.

C. Image

1. 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 should, in each case, determine how much compression is advisable, with regard to both reuse and sharing of your image files.

2. Conversion

If your images are stored in a format considered unacceptable (cf. Table 1), these 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. However, before using one of these, it is advisable to read any terms of use.

D. Text

1. 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 should still save it in plain text format. To do so, select "Save as file type: Plain text (.txt)" in the menu *File > Save As*.

2. 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 should 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 this is not available, save the file as plain PDF.

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.

3. Tabular text

Tabular text data should 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 should 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 in Section D2 above.

E. Transcription

1. Font

All transcriptions should be made using Unicode-encoded fonts, e.g. IPA Doulos SIL.[2] For phonetic transcriptions, SAMPA (Speech Assessment Methods Phonetic Alphabet, ASCII characters)[3] is an alternative to IPA. If the recommended font is not available for the type of transcription your dataset requires, it is imperative to upload, under Data & Analysis in your TROLLing dataset, a separate ReadMe file with instructions about how to read the transcriptions.[4] Note that the font package itself should *not* be uploaded, given copyright restrictions.

2. File format

Transcriptions can be orthographic or phonetic, and in both cases, one is likely to use non-standard symbols (e.g. Cyrillic letters or the IPA alphabet). Regardless of the nature of the transcriptions, if these are presented in a Word- or Excel-file, or as plain text, you are recommended to convert the file into a PDF/A file (see conversion procedure in Section 3.2 above). The original text file as well as the PDF/A file must be uploaded.

If the transcriptions are presented in a Praat TextGrid-file, for which we at this point cannot ensure future readability, the following steps should be taken:

- Upload the original TextGrid-file as is.
- Convert the original TextGrid-file to a CSV-file, then upload. By following the procedure on the linked web page, the range of each interval is displayed next to the transcription itself, making future linking of the transcription and the sound file possible without resorting to the TextGrid: http://wwwhomes.uni-bielefeld.de/gibbon/Forms/Python/PHONETICS/textgrid2csv.html

F. Video

1. Format

The highest quality video format is the one in which the movie has been recorded. The size of an uncompressed video file is however problematic for sharing, thus conversion, with a certain loss in quality, is necessary. Remember, however, to keep a copy of the master file in the original format. If later editing or conversion is required, this should be done using the master file: Editing or conversion of an already converted file will increase loss in quality.

2. Conversion

If your videos are stored in a format considered unacceptable (cf. Table 1), these must be converted to the MPEG-4 format. If you do not have license to any professional

[2] To download SIL Fonts, cf. http://scripts.sil.org/cms/scripts/page.php?cat_id=FontDownloads.
[3] For an overview of SAMPA symbols, cf. https://www.phon.ucl.ac.uk/home/sampa/.
[4] Cf. for instance an example in the file entitled "To read the Church Slavonic transcriptions.pdf" in Eckhoff (2015), cf. http://hdl.handle.net/10037.1/10190.

conversion software, we advise you to use the VLC Media Player (standard application on both Mac and Windows), or an online free image converter. However, before using any free converter, it is advisable to read any terms of use.

V. How to describe your data

In order for users to be able to understand and reuse your data, it is essential that you describe it in a comprehensible and consistent manner. Data come in many different forms, and for most types, there is no common standard of description. In this section, we present guidelines on how you should prepare and describe data for archiving in TROLLing.

Your data description should be provided in a file named "ReadMe" together with your data files. You should save your ReadMe file(s) as a Unicode UTF-8 plain text file (.txt). In case you need to use illustrations or special characters, you may save your ReadMe file(s) as PDF/A (see Sections III and IV above for more information about these file formats).

First in your ReadMe file(s) you should give an overview and short description of the files contained in your dataset. The remaining contents of your ReadMe file(s) will vary according to what kind of data you are going to archive. Below we give some recommendations for ReadMe files for two common types of data, tabular data and computer scripts.

A. Tabular data

It is advisable to upload a separate ReadMe file with a comprehensive description of the data file, including the data in each column, the data format and the standard(s) used. This can alternatively, or additionally, be inserted into the Description field in the Citation Metadata tab.

1. Column and column headings

For each column in your tabular text file (.csv or .txt; see above) you should indicate what kind of data it contains, and what data format the values have. Column headings should be meaningful and not too long. Make sure you do not use duplicate column headings within a file. Use only alphanumeric characters, underscores, or hyphens in column headings. It is good practice to have column headings start with a letter. If possible, indicate units of measurement in the column headings.

Use only the first row for column headings, otherwise rows may be missed when your data is imported to spreadsheet software or other utilities.

Examples of good column headings

vowel_length_ms
record_time
language_name
pos

2. Data values and formatting

Use standard codes or names when possible, e.g. ISO code for language names http://www.loc.gov/standards/iso639-2/php/code_list.php) and established tag sets for POS/parts of speech (e.g. http://ucrel.lancs.ac.uk/claws2tags.html, CLAWS2 Tagset).

Avoid using special characters, such as commas, semicolons, or tabs, in the data itself. This might cause trouble when the data file is imported into a spreadsheet, or read by other software. If such characters are nevertheless necessary in the presentation of your data, please specify their use in the ReadMe file.

3. Examples of tabular data description

The column "vowel_length_ms" contains values for the vowel length in milliseconds of the analyzed items in the dataset. Only integer numbers are used, e.g. 45, 32, 11.

The column "record_time" contains values for the time when the record was made. The time format used is YYYY-MM-DD hh:mm, e.g. 2014-03-15 17:21.

The column "lang_name" contains values for the name of the analyzed languages. The ISO 639-2 Code format is applied.

dan	Danish
nob	Norwegian Bokmål
swe	Swedish
…	

The column "pos" contains values for the part of speech of the analyzed items. The applied tag set is the CLAWS2 Tagset.

NP	proper noun, neutral for number (Indies, Andes)
NP1	singular proper noun (London, Jane, Frederick)
NP2	plural proper noun (Browns, Reagans, Koreas)
…	

B. Scripts

Another common data type are scripts used in statistical analysis. Before archiving, make sure you add a description for each step used in the script. Below, we present an example, taken from TROLLing[5]:

[5] Janda et al. (2014), cf. http://hdl.handle.net/10037.1/10121

```
#Here we input the data on mannel from Table 2 in the article:
mannel=matrix(c(2535,1231,90,261),byrow=TRUE,ncol=2)

#Now we print out the table so that we can see what it looks like:
print("mannel+ in newspapers and literature")
print(mannel)

#Now we run a chi-square test:
test=chisq.test(mannel)
print(test)

#Now we sum all the numbers in our table:
print("This is the sum for the mannel+-table:")
print(sum(mannel))

#Now we calculate the effect size (Cramer's V). We do this by taking the square root
 of the chi-squared value and dividing it by the sum for the table:
cramer=sqrt(test$statistic/sum(mannel))
print(cramer)
print("This is the effect size")
```

VI. File size

The size of the individual data file should not exceed 2 Gb. If you have files exceeding this limit, please contact us at research-data@support.uit.no.

VII. 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