ELAR: instructions for depositors

As a requirement of your ELDP grant, you must deposit your data with the Endangered Languages Archive (ELAR) at SOAS on an annual basis at the same time when you hand in your annual report. The payment of the next tranche of funds depends on the approval of your annual report and your deposit. To ensure swift processing and upload to the catalogue, ELDP and ELAR require you to prepare your data according to a set of guidelines that cover file naming, file formats and metadata preparation.

These guidelines are likely to meet the needs of most projects. If you feel the needs of your project are not met, please get in touch with us so we can work out a solution which meets both your and our requirements. In this case, get in touch with us by sending an email to elararchive@soas.ac.uk and cc eldp@soas.ac.uk. Also, we are always happy to hear back from you and grateful for any feedback which helps us to improve our guidelines.

File names

ELAR requires that you systematically name your files.

File names are important to minimise problems when being processed by different computer systems. We will outline file naming conventions and additional considerations for working in teams of researchers. In addition, there is some background information to help you understand the guidelines.

Getting started: clearing up file naming terminology

A file name consists of three parts.

1) base name of the file
2) dot/full stop
3) file extension:

base name	dot	extension
howto-file_names	.	pdf

howto-file_names.pdf

If the file extension is not visible in your computer view, there is a guide at the end of this section that shows you how to make file extensions visible.

Each file name consists of a base name followed by a dot followed by the file extension, e.g. `ikaan253.wav` or `ikaan253.mp4`. The dot and the file extension are given automatically by the computer; you only have to decide on a system for the base names. Base names may contain the lower case letters a-z, the numbers 0-9, the underscore symbol _ and the hyphen symbol - . Please note that base names may not contain any blank spaces. Instead of blank spaces please use an underscore _ .

1. For base names, first pick a keyword from your project description and write this in lower case letters, separated by an underscore if needed (do not use any blank spaces), e.g.

```
uruan
uruan_dirges
```

2. Then for each file add a sequential number to the word or phrase, e.g.

```
uruan001,uruan002, ..., uruan359, ...
uruan_dirges001, uruan_dirges002, ..., uruan_dirges339, ...
```

3. Sometimes files belong together thematically (e.g. a series of photographs of for example women; boat building; workshop…). In this case,
    a. Give all files the same beginning of the base name.
    b. Then add a hyphen and another sequential number.
    c. The dot and file extension will be automatically created by the computer.

```
uruan003-1.jpg, uruan003-2.jpg, uruan003-3.jpg, ...
uruan_dirges003-1.jpg, uruan_dirges003-2.jpg, ...
```

For series like this, please add a summary file with metadata. This file has the same base name, minus the hyphen and the second set of sequential numbers, e.g.

```
uruan003.txt
uruan_dirges003.txt
```

If you follow this file naming convention, we know that everything that has the same name up to the hyphen (if there is a hyphen) belongs together.

We will provide instructions for automatically renaming and numbering picture files as well as for printing out file listings and providing short file-by-file metadata for series of files. These will be downloadable from the ELAR website.

Managing file names in projects with more than one researcher

When two or more people are collecting and managing data at the same time, file naming may lead to clashes. Imagine that you have recorded the whole day, your colleague has recorded too, and in the evening you arrive at home only to realise that both you and your colleague have named their recordings `ikaan253.mp4`. Unfortunately, the two files called `ikaan253.mp4` are of course two different recordings which now ended up having the same file name and might end up overwriting each other when you share files or copy to your backup. The solution to this problem involves coordinating with your colleague(s) while sticking to the same basic principles outlined above.

Option 1

Use the same kind of base name for all researchers. Make sure that this base name follows the guidelines set out above. However, instead of using the numbers purely sequentially, let Researcher A assign numbers 001 to 499. Researcher B assigns the numbers 500 to 999, e.g.

Researcher A: `uruan001, uruan002, ..., uruan498, uruan499`
Researcher B: `uruan500, uruan501, ..., uruan989, uruan999`

Using this solution means that recording uruan300 will probably be recorded months after ikaan502. This may strike you as wrong, but it is in fact nothing to worry about because in the metadata it will be clear when each recording was done. File naming is not done for aesthetic or sentimental reasons, it is purely practical.

Option 2

Use different base names for all the researchers, e.g. the surname or initials of the researchers:

Researcher A: `salffner001, salffner002, salffner003, ...`
Researcher B: `adekanye001, adekanye002, adekanye003, ...`

Just make sure that the file names follow the guidelines outlined above.

Grouping files

Files that are derived from each other belong together and must receive the same name.

If you extract an audio file from a video file, the audio file and the video file belong together. If you annotate the video and audio, the ELAN annotation file also belongs to this set. These files that belong together must have the same base name but will have different file extensions, e.g.

video file	`uruan001.mp4`
extracted audio file	`uruan001.wav`
ELAN annotation file	`uruan001.eaf`

Files that are created in the same recording session but that are not derived from each other receive different names.

For example, you are video recording a session with many people and at the same time, you are also one particular speaker within that crowd with a separate audio recorder and a lavalier mic. In this case you create two bundles, one for the video recording of the entire group and one for the audio recording of the single speaker.

All files you derive from the main recording file (video or audio) will be named the same but the two bundles have two different names, e.g.

video file	`uruan001.mp4`
extracted audio file	`uruan001.wav`
ELAN annotation file	`uruan001.eaf`
separate parallel audio file	`uruan002.wav`
ELAN annotation file	`uruan002.eaf`

Note that the corresponding annotation files may also be quite different: the annotation based on the video includes all speakers, whereas annotation based on the parallel audio includes only one speaker.

File naming summary

1. File names consist of a base name only using letters (without special characters like accents), numbers, hyphen and underscores and **no spaces**; a dot and an extension determining the file type (wav, doc, pdf) will be given by the computer automatically when saving the file.
2. Files that are derived from each other get the same base file name and different extensions, e.g.

   `uruan001.mp4, uruan001.wav, uruan001.eaf`

3. Files that are not derived from each other get a new base file name, e.g.

   `uruan001.mp4, uruan002.wav`

4. Series of files that conceptually belong together have the same beginning of the base name followed by a hyphen followed by a sequential number, e.g.

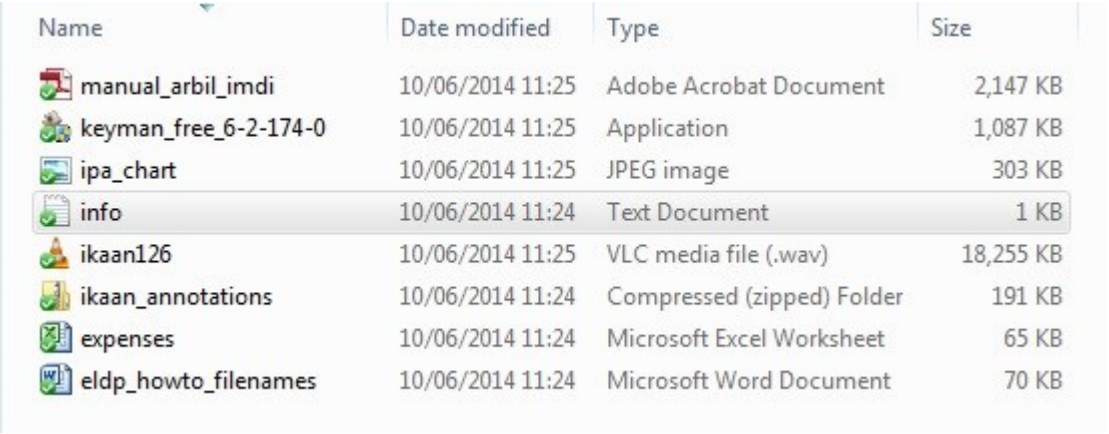
   `uruan003-1.jpg, uruan003-2.jpg, uruan003-3.jpg, ...`

Trouble shooting – I can't see the file extensions

Both on Windows and on Apple Mac machines, file extension are by default set to be invisible. Mostly this is okay, the computer is just trying to be helpful and prevent you from accidentally deleting the extension, which would make it impossible for the computer to know what kind of file you are dealing with. However, sometimes you may want to see the extensions and both on Windows and on Macs you can make extensions visible.

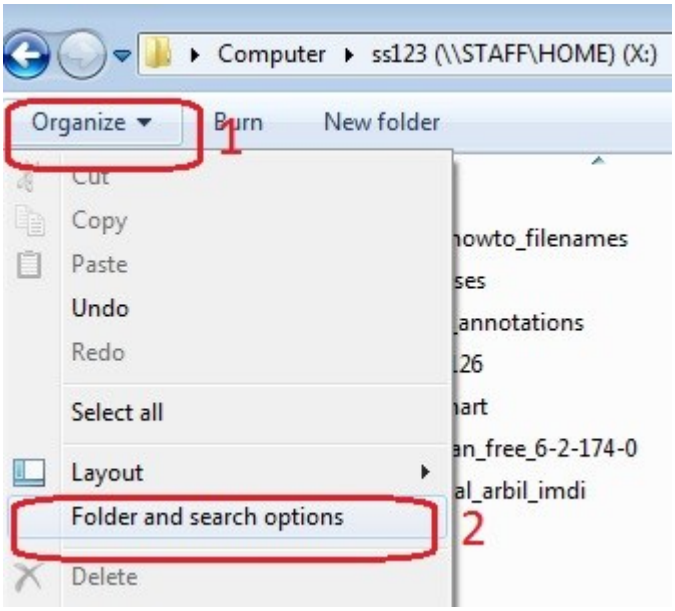
Windows

When the file extensions are hidden, your Windows Explorer looks like this. Windows knows what each file type is, and it can tell you in the type column and with the little icons next to the files, but file extensions are nowhere to be seen.

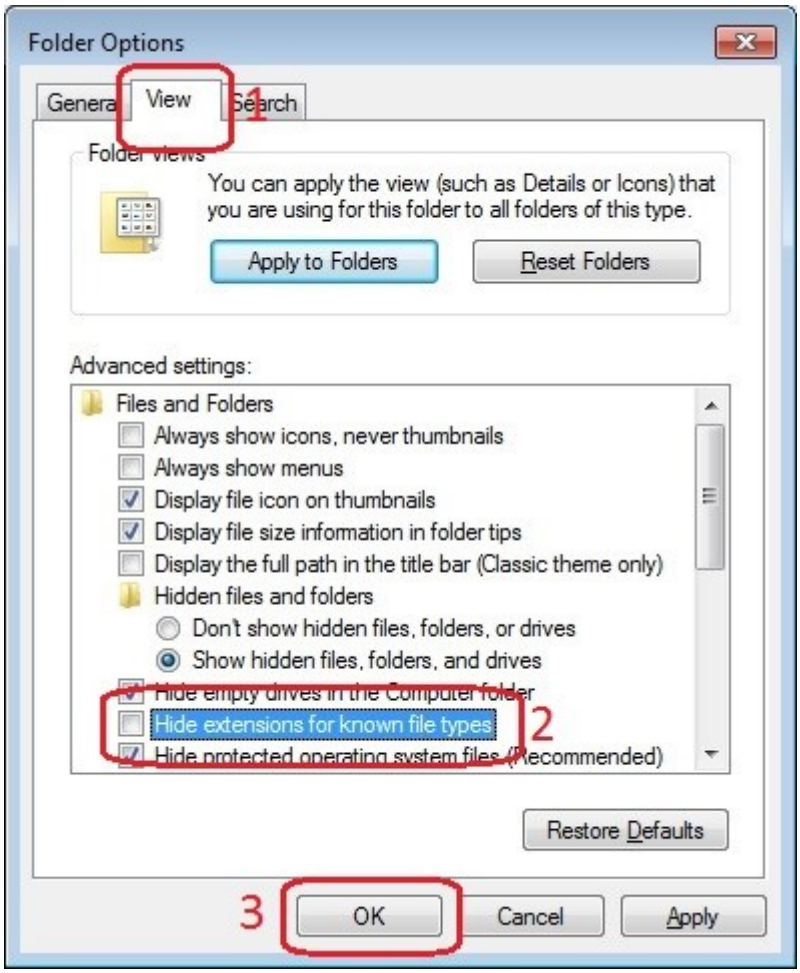


Name	Date modified	Type	Size
manual_arbil_imdi	10/06/2014 11:25	Adobe Acrobat Document	2,147 KB
keyman_free_6-2-174-0	10/06/2014 11:25	Application	1,087 KB
ipa_chart	10/06/2014 11:25	JPEG image	303 KB
info	10/06/2014 11:24	Text Document	1 KB
ikaan126	10/06/2014 11:25	VLC media file (.wav)	18,255 KB
ikaan_annotations	10/06/2014 11:24	Compressed (zipped) Folder	191 KB
expenses	10/06/2014 11:24	Microsoft Excel Worksheet	65 KB
eldp_howto_filenames	10/06/2014 11:24	Microsoft Word Document	70 KB

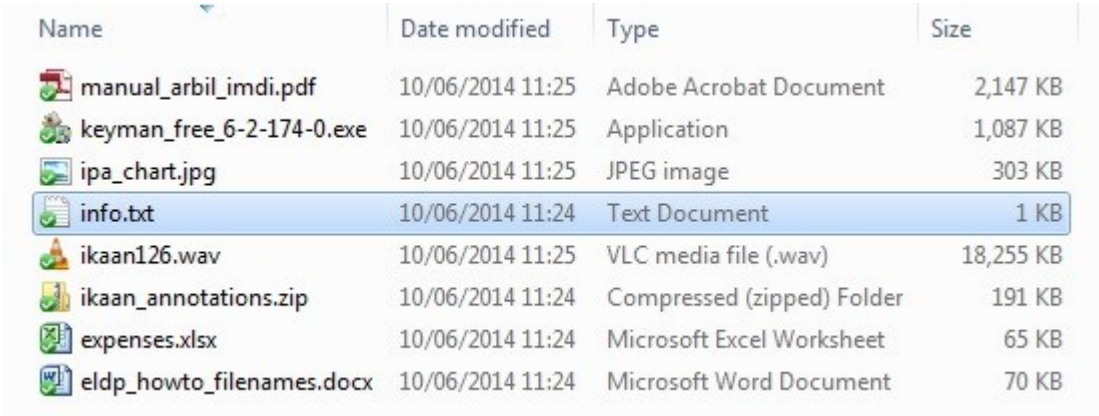
To make the file names visible, click on the *Organize* tab. A drop-down menu opens up. In this menu, click on *Folder and search options*.



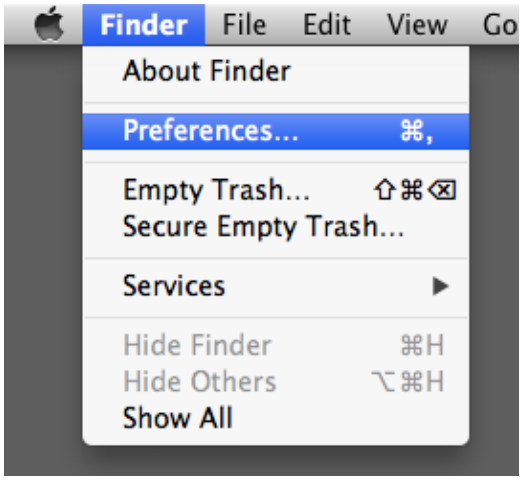
This opens the *Folder Options* window. In this window, click on the tab called *View*. Untick the box next to *Hide extensions for known file types*. Then click on *OK*.



You will now be able to see all dots and the file extensions.



Macs also hide most file extensions by default. To insure that all file extensions are visible open the Finder. Then go to the Finder menu and select *Preferences*.



The *Finder Preferences* window opens. In this window, click on the *Advanced* tab to select this tab. Then, tick the box next to *Show all filename extensions*. All file extensions are now visible.

When you are done, simply close the Finder Preferences window.

Format overview

Audio

Audio formats are now fairly stable and standardised. Please use the following format and settings:

type	**wav**
settings	**48kHz, 16bit, stereo**

If your recorder offers you different options for encoding wav files, choose PCM or LPCM.
If you use a mono microphone and have the option to change the recorder from stereo to mono you can do so, but don't forget to switch back to stereo for the stereo microphone the next time you use the recorder. There is no harm in using stereo settings for working with a mono microphone, but if you use mono settings with a stereo microphone you will lose information.
Please do not process audio data. You can cut out beginning and end or sections in between, but please do not filter for noise, change mono and stereo, or similar things.

Video

Please provide two versions of your video data: the original recordings and a compressed version in mp4 as a working copy.
First, do your original video recordings with the best quality and highest resolution your camera can produce. Archive these recordings as material that researchers later on can work with if they want to build on your data.
Second, compress your original recordings into smaller mp4 files that you can work with in ELAN and that you can stream over the internet. For compression, use the settings below and keep the remaining settings as close to the original settings as possible.

type	**mp4**
settings	**H264**

We are currently working on teaching materials for video conversion that explains which tools you can use with which settings. These materials will be downloadable from the ELAR webpage.

Photographs

Consumer-grade photo cameras save photos in JPEG format. Professional SLR cameras save photos in raw formats as well as JPEG. For both types of camera, set your camera to the highest resolution and quality the camera offers. Submit JPEGs, and if you have raw formats, please also submit these but make sure you have embedded the camera metadata in the file.

type	**jpg, raw with embedded metadata**
settings	**the highest resolution and best quality your camera offers**

Scans

For scans, use high colour settings (even if the writing is black on white paper) and a resolution of at least 300dpi.

type	**tiff; pdf**
settings	**high colour, high quality, 300dpi; PDF-A**

Annotations

Annotations produced by ELAN and by Praat are accepted. Note that Praat annotations can also be imported into ELAN and then saved as ELAN annotations.

type **eaf, pfsx, TextGrid**

Written documents/PDFs/Plain text files

Please convert all written documents to either archivable PDF formats or plain text files in Unicode UTF-8 encoding. Electronic written documents come in a wide range of file formats. The most common of these, e.g. doc and docx, are not suitable for archiving purposes.

type **pdf, txt**
settings **PDF-A, Unicode UTF-8**

We are currently working on teaching materials for conversion to PDF-A and UTF-8. These materials will be downloadable from the ELAR webpage.

Dictionary data and interlinearised texts

Dictionaries and interlinearised texts produced in Toolbox and FLEx are accepted. Please also give us a printable version in pdf.
Toolbox projects should be submitted with all data files and all associated settings files, i.e. the whole Toolbox folder. All files should be encoded as Unicode UTF8. FLEx files have to be exported to XML files.
We are working on other ways to make dictionaries available through the archive and will post updates on the website.

Toolbox

type **txt, typ, lng**
settings **Unicode UTF-8**

FLEx

type **xml**
settings **Unicode UTF-8**

Metadata

Metadata should be prepared in Arbil or the CMDI Maker as CMDI metadata files. We have prepared some guidelines for metadata management and are developing teaching more material as we go along. These materials will be downloadable from the ELAR webpage.

type **cmdi**

Metadata

Metadata have to be provided according to the ELDP metadata standard (based on CMDI)

ELAR has reorganised metadata management. We have changed from spread sheets to metadata management in dedicated metadata tools such as the CMDI Maker and Arbil. Both tools run on all operating systems. You can choose between one of the two editors:

CMDI maker

CMDI maker is a user-friendly tool which allows you to compile metadata in a systematic and straightforward way. The CMDI Maker and a video tutorial are available here:

http://cmdi-maker.uni-koeln.de/

https://www.youtube.com/watch?v=1YE2RLyTj-Q

Arbil

Arbil is a corpus management tool which is more complex and allows you to hierarchically organise your materials according to your needs. The software is available for download from the MPI LAT website here:

http://tla.mpi.nl/tools/tla-tools/arbil/

Please use the ELDP Project and ELDP Bundle CMDI profiles for your metadata. If you find that these do not suit your needs please get in touch with us.