Leiden University Data Management Plan[i] - Linguistics example[ii]

Please contact Leiden Research Data Office if you need help: datamanagement@library.leidenuniv.nl or http://www.library.leiden.edu/education-research/library-research/researchdata/support.html

Name and contact details	Name Surname
Name of project and group	Finding the external relations of language X
Description of your research	I am trying to find the linguistic relatives of language X through lexical comparison in domains indicated in the archaeological record such as metallurgy and exchange networks. I am also investigating the language of perception, and its relation to prehistoric metalworking in the region.
Project duration	Start: *01-0X-2013* End: *31-0X-2017*
Names of people and their responsibilities for data management	Name Surname of PhD researcher : collecting, storing, documenting, sharing and archiving the data. Local researcher : collecting and documenting fieldwork data obtained in XX.
Funding body(ies)	*European Research Council*
Grant number	*XX*
Partner organisations	*Not applicable*

About this Data Management Plan

Date written	*31-03-2015*
Date last update	*31-03-2015*
Version	*1.0*

Changes in this version of the Data Management Plan

Component	Progress / Execution *Please describe shortly what progress you have made, any questions or issues you have encountered and want to discuss, etc.*
1. Data collection	………
2. Data storage and back-up	………
3. Data documentation	………
4. Data access, sharing and reuse	………
5. Data preservation and archiving	………

<table>
<tr><td colspan="2">1. Data collection
Describing the data you will be creating/collecting</td></tr>
</table>

1.1	**Will the project use existing or third party data ?** ☐　No ☐　Own / group previous research ☐　Academic collaborators ☐　Commercial collaborators X　**Publicly available database / archive** ☐　Specialist commercial data provider ☐　Other (please specify) *Describe shortly provenance, type and format of this data. Are there any restrictions or requirements for use of third party data such as licensing conditions?*
	Wordlists or parts of wordlists from online lexical databases such as the Intercontinental Dictionary Series (IDS), World Loanword Database (WOLD), and CLiCS. All three are covered by creative commons licences: WOLD data can be shared and re-used but their use must be attributed, IDS similarly but any re-worked data cannot be re-distributed. CLiCs data can be shared and re-worked without citation/attribution.
1.2	**What type()s of data will you collect or create, in what file format(s)?[iii]** *Note that not all formats are long-lived. For sustainable access you best use the formats recommended by data archives, see for examples: https://datacentrum.3tu.nl/fileadmin/editor_upload/File_formats/Preferred_formats.pdf or http://www.dans.knaw.nl/sites/default/files/file/EASY/DANS%20preferred%20formats%20UK%20DEF.pdf*
	Lexical and typological data collected from dictionaries, grammars and existing wordlists will be collected in spreadsheets. Spoken data from experiments, interviews and structured language data collection in the field will be collected on an audio recorder and stored as wav files. Transcriptions of these recordings will be stored (and analysed) in spreadsheets (language of perception) and possibly also ELAN. Observations from the field will be stored in notebooks and later written up into word files, where extended notes on readings will also be collected. Photographs (for documenting fieldwork and the region) will be collected in jpeg format.
1.3	**How will you collect and/or create your data?** *Please describe shortly. Name any relevant protocols and/or standard in your area of expertise.*
	I will follow best practices as recommended in the MPI Psycholinguistics field manuals.
1.4	**What tools, instruments, equipment, hardware or software will you use to capture, produce, collect or create the data?** *Please give the names of the tools and state if they are already available. If not, state how you intend to acquire them. If applicable, describe whether you use a paper or electronic labjournal.*
	Audio recorder and microphone: project Laptops for experiment stimuli, data processing, work in the field: project Camera: own Software: QGIS, ELAN, R: project laptop
1.5	**What is the estimated size of the data?** *Please describe shortly. Stages to be adopted if relevant.*

Data stage	Specification of type of research data	Software choice and file format	Data size now	Data size when project is finished
Raw data	*Fieldwork experiments* *Lexical and typological lists*	*wav* *xls*	*365mb* *250kb*	*1000mb* *1mb??*
Processed data	*Fieldwork experiments* *Experiment transcriptions* *Lexical and typological lists*	*wav* *xls* *xls*	*365mb* *300kb* *250kb*	*1000mb* *1200kb* *1mb??*
Results	*Fieldwork experiments*	*Xls*	*N/A*	*1mb??*

	…….

	2. Data storage and security
	Ensuring that all research data are stored securely and backed up or copied regularly during your research

2.1	**Where will you store your data?**
	Please describe how safe storage is guaranteed. Specify your method if your data is collected and / or transported in different locations / countries.
	☐ On university departmental network storage (J:)
	x On university personal network storage (P:)
	☐ In a Virtual Research Environment (Sharepoint)
	x Physical storage (e.g. USB, external hard drive)
	☐ Cloud service (e.g. SURFdrive)
	☐ Other, namely: …
	….
2.2	**Will your data be backed up?**
	Please specify shortly for each storage device frequency, location of backups and who is responsible.
	Describe how you can restore your data in the event of data loss and who is responsible.
	Data will be backed up on an external hard disk and key files will also be stored in an online facility such as Google Drive or the university cloud service (when it becomes available).
2.3	**Are there any commercialisation, ethical or confidentiality restrictions about handling your data?**
	Please specify shortly.
	☐Contractual obligations
	☐Requirements by law : protection of personal data (e.g. privacy law) : specify in 4.1
	☐Requirements by law : copyright, intellectual property : specify in 4.1
	X Ethical restrictions (e.g. ethical review) : specify in 4.1
	☐ Commercial considerations (e.g. patentability)
	☐ Formal security standards
	☐ No requirements
	☐ Other, namely: ………
	…………
2.4	**How will access to the data be managed during the project?**
	Please specify for each storage device, from different locations / countries.
	In the office (i.e. in Leiden) access will be through the desktop, in other locations/countries the project laptop and external hard drive will be used.
2.5	**What are the main risks to data security?**
	Please list risks, e.g. accidental deletion, falling into the wrong hands.
	Please describe what would happen if the data get lost or become unusable.
	The main risks to data security in my project are: Loss or theft of laptop and/or external hard drive, accidental deletion and a computer/server crash. In the event of any of them the backed up data on the hard drive and cloud should avoid loss of data.
2.6	**What measures do you take to comply with the security requirements and to mitigate the risks?**
	Describe how you can restore your data in the event of data loss and who is responsible.
	If applicable, please describe procedures to ensure personal data are handled confidentially and who is responsible.
	☐ Access restrictions
	☐ Encryptions
	☐ Data processing
	X De-identification / Anonymisation
	X Regular back-ups
	X Master copy stored on university network storage
	☐Master copy stored elsewhere
	☐ Other, namely: …

	In the event of loss or theft, the backup locations should allow data to be restored. I will anonymise audio files from the field so the participants cannot be identified by a third party.	
2.7	**How do you differentiate between raw and processed data?** *Please explain shortly why you (do not) differentiate.*	
	☐ I will not differentiate X I will create a new file for processed data ☐ I will create a new file for processed data and I will lock raw data ☐ Other, namely: …	
	…………	
2.8	**Is there any non-digital data or outputs that the project will generate? Where will these outputs be stored?** *Please specify shortly and describe who is responsible for storage of these outputs.*	
	I am currently contemplating a photography exhibition, to be shown in a public or semi-public location (e.g. museum, embassy). Once the exhibition and project end, I would keep and store the photographs.	
2.9	***Do you expect to have any supplementary costs for storage not covered by the project budget?*** *Please specify*	
	None known	

	3. Data documentation Documenting your data to help future users to understand and reuse it	
3.1	**How will files be named?** *Please describe shortly.*	
	With short, meaningful names, including date and version number if appropriate, e.g. metallurgy_comparative_v1	
3.2	**How will folders be named and structured?** *You are invited to draw a folder structure and describe it shortly.*	
	Fieldwork - Summer 2014 - Language of perception - Audio files - Transcriptions - Corrections - Analysis - Summer 2015 - Etc - Etc Language XX - Data (sources) - External relations This structure keeps separate the different elements of the project, as well as extra-project activities, such as editorial work. Reading is also kept apart from data collection and analysis.	
3.3	**How do you handle version control to maintain all changes that are made to the data?** *Please explain your choice shortly. Remember to also document any deletion of data, if applicable.*	

		☐ No version control (e.g. original files are overwritten) ☐ Version control software, namely: … Ⅹ Data/version number in filename/folder Ⅹ 'Track changes' feature in software Ⅹ By saving the script with which I process my data ☐ Other, namely: …
		…………
3.4	**What metadata standard will be used, if any?**[iv] *Please explain why you use this standard (most used in my discipline, required by the data archive where I will deposit my data). Please outline how the metadata will be created (read me file, spreadsheet, in the data). If no standard exist, please specify which metadata is needed to understand the data.*	
	☐No metadata standard is used ☐ Generic metadata standard (e.g. Dublin Core) ☐Standard automatic Windows metadata (e.g. from Word, Excel) Ⅹ Specialised metadata standard, namely: IMDI ☐ Other metadata standard, namely: …	
	…………	
3.5	**What supporting information / documentation will you create to enhance understanding of the data ?** *Please describe shortly how peers should be able to understand the data. Examples are a readme.txt, lab journals, a codebook, survey questions etc. Is there a standard for documentation in your field? Describe at what moment in your research process you will add the documentation necessary to make sure the data is understandable for peers.*	
	Spreadsheet with metadata for experiment results as well as comparative lexical and typological lists; overview of experimental paradigms with example stimuli (e.g. language of perception).	

4. Data access, sharing and reuse
Managing access and security, sharing your data

4.1	**Are there any restrictions placed on sharing / reuse of some / all of your data?** *Please account for not sharing your data. Reasons may be ethical, commercial, security-related, protection of personal data rules, intellectual property, copyright,*
	No
4.2	**With whom will you share your data at which stage in your research? You can use the table below.** *Please state any sharing requirements, e.g. funder data sharing policy. Please describe shortly how you will share your data: on request, pro-actively, etc.. Please specify how your data can be accessed.*

	Would not share with anyone	Would share with my immediate collaborators	Would share with others in my research centre or at my institution	Would share with scientists in my field	Would share with scientists outside of my field	Would share with anyone
Immediately after the data has been generated		X	X			
After the data has been normalized and/or corrected for errors		x	x	x	x	X

After the data has been processed for analysis		X	x	x	x	x
After the data has been analysed		x	x	x	x	x
Immediately before publication		x	x	x	x	x
Immediately after the findings derived from this data have been published		X	x	x	x	x

Based on: Interview worksheet, Jake Carlson, Purdue University Libraries / Distributed Data Curation Center

Access will be provided through the university cloud service and The Language Archive.

4.3	**If intending to share any part of the data, do your participant consent forms include information about intentions for sharing, retention of data and steps taken to protect participants privacy and confidentiality?**
	☐ Not applicable. X Yes. *Please specify the relevant formula in the consent form.*
	Not currently, although forms are not a very appropriate medium for gaining consent in my fieldsite. However the consent is gained (written or oral), I need to build in a sharing and data retention clause.
4.4	**Who has authority to grant (additional) access to your data?** *Please describe shortly.*
	X Only you X A colleague from the project, namely: local researcher ☐ Supervisor ☐ Funder ☐ Collaborator / research partner organisation ☐ Other, namely: …
	…………
4.5	**How will you manage copyright and Intellectual Property Rights issues?** *Who owns the data? How will the data be licensed for reuse? Please describe shortly your choices and their consequences.*
	The data will be available to all for reuse, as long as the source is attributed. All participants will be anonymised.
4.6	**What is the audience for reuse?** *Please list possible audiences and purposes. Consider who might use it now and who might use it later.*
	Academic community (linguists, archaeologists, anthropologists): cross-linguistic/cultural studies, gaining deeper understanding into certain aspects of the lexicon and grammar (e.g. language of perception) Speaker communities: teaching/learning purposes, documenting certain aspects of the language, understanding the language in more depth.

	5. Data preservation and archiving Preserving your data	
5.1	**Which criteria will you use to decide which data has to be archived?** *Please shortly describe your choices.*	
	✗ Type of data (raw, processed) and how easy it is to reproduce it ✗ Relevance of content for others ✗ Usability of format for others ✗ Data underlying publications ☐ Verification of research ✗ Available time ☐ Available money ☐ Other, namely: …	
	…………	
5.2	**How long should your data be preserved? Are there any requirements regarding the disposal of data?** *State obligations you have by law, funder, university, etc. if any.* *Describe how you will dispose of the data, e.g. how you will get approval, what people and/or tools you need, etc.*	
	There are no specific obligations regarding how long my data should be preserved therefore they will be archived normally at TLA in Nijmegen, which is a relatively stable location. I won't need to dispose of data for any reason.	
5.3	**Which data repository is appropriate for archiving your data?** *Please describe shortly. Does this archive have a 'data seal of approval' or another form of certification?*	
	☐ Discipline specific (international) repository, namely … ☐ 3TU.Datacentrum ☐ SurfSara ☐ DANS ✗ Other (international) repository, namely : The Language Archive ☐ Other, namely: …	
	TLA data has recently been included in the DANS archive, so sending it to the former infers inclusion in the latter too now.	
5.4	**Does the archive have specific requirements concerning file formats, metadata etc.** *Provide relevant urls to the documentation on these requirements. Describe how you intend to meet those requirements, e.g. converting the file formats, providing supplementary documentation.* *Will there be extra costs to prepare your data for archiving? Please specify. See http://www.data-archive.ac.uk/media/247429/costingtool.pdf*	
	TLA requests IMDI metadata structure but for specific questions regarding formats, etc, it is necessary to contact them directly. Spreadsheets will need to be converted to csv files, but the audio files can remain as wavs. There will be no extra cost attached to doing this.	
5.5	**What costs (if any) will your selected repository charge? Who pays?** *Please state the costs in euro's and the institution that pays for it.*	
	There will be no costs associated with archiving the data.	
5.6	**Who is responsible for the data after the project ends?** *Please state a position and the current person in that position.*	
	Supervisor / Name Surname	

Think of the different stages (for instance : video recording, transcript, annotation, lists of typological features ....).

iv See http://www.dcc.ac.uk/resources/metadata-standards or http://en.wikipedia.org/wiki/Metadata_standards or the relevant repository.