Purpose: To clearly describe the expectations and limitations around use of datasets within the ACLEW team
N.B. Additional details in addendum here.
N.B. 2: March 25th Addendum for all PIs to sign here

Term Definitions:
Data: *The audio recordings and any derived data that are not fully anonymized. Fully anonymized data that describe group characteristics are also subject to constraints as specified below and in the addendum.*
Trainee: *Research Assistant (undergrad, staff, paid or volunteer), Graduate Student, Postdoc, or Lab Technician who is supervised by an ACLEW PI.*

Background:

1. The various datasets PIs have agreed, by signing onto the DiD grant, to share their data for the purpose of the ACLEW project.
2. ***Each*** of the datasets involve specific agreements with the recorded participants regarding how they will be used and shared. ***Some*** datasets involve recordings of third parties whose agreement to be recorded may be questionable or not fully documented. ***Some*** of the datasets involve vulnerable populations for which extra care is needed to ensure their rights as participants are safe-guarded.
3. In order to effectively share the datasets, access is granted to a rather large number of people (the PIs), who will be granting access to their trainees. Dataset "owners" nonetheless need to have knowledge of, and control over, who has access to their data and for what purpose.
4. It is therefore important to set out specific guidelines about how and when data can be shared, and what is considered legitimate usage, as well as a system for tracking who has access.

Ground rules:

Access Responsibility and monitoring of usage by trainees. Each contributing laboratory will have a primary PI responsible for data usage and access for that laboratory. The primary PI must sign off on this document. Each PI should personally sign on to these rules, and is ultimately responsible for the data usage of any trainee under their supervision, although they may delegate day-to-day monitoring to a single individual (e.g. a postdoc) if desired. All laboratory members with access (including PIs) must be listed in the ACLEW Datasets Current Access List.

Ethics training. Any trainee with access to the ACLEW data must have ethics training. Ideally, trainees should have comprehensive ethics training such as CORE or CITI. At the minimum, trainees should complete the ACLEW ethics training on Qualtrics here.

Data-sharing and access. Data should only be shared with trainees under the PI's direct supervision who have a specific ACLEW-related task. Data should not be shared with anyone outside of the ACLEW team without permission from relevant data owners. Access to ACLEW files (whether via direct access to an ACLEW repository like Github or Databrary, or via local files downloaded from an ACLEW repository) should not be granted by a PI to anyone other than their direct trainee. If a need arises to grant access to someone new that is not a direct trainee, permission should be obtained first from the datasets PIs as a group. Trainees' access should be removed upon the completion of their work or when they leave the PI's direct supervision. Trainees should be made aware that they do NOT have permission to share the data with anyone not already approved (where technically possible, trainees should not have the ability to share data). PIs may grant access via the ACLEW repositories to their own subcorpora at their discretion. Such sharing should not compromise the confidentiality of the other sub-corpora and is otherwise not covered by this agreement..

Storage. Local storage of non-anonymous data files is permitted *where necessary*, but a) storage should be password protected and restricted to PIs and trainees, b) copies should be kept to a minimum, c) storage should be constrained and documented.
Any local copies held by trainees who leave the project should be deleted; this will be verified semesterly (by Bergelson and Soderstrom).
Audio files will be stored on Databrary; derivative data that is identifiable (or that has the potential to be re-identified) will be stored on a private github repository (with local mirrors as needed by tools people); anonymized derivative data must all be posted on an appropriate ACLEW github at the end of the project (though may also be kept locally as writeup and analysis continue). ***Summary of storage locations, access levels, and data types (see addendum for more details:***

Data type	***Location***	***Access/Restrictions (During project period)***
Audio	Databrary	ACLEW members, others by PI permission*
Annotation files	Github	ACLEW members, others by PI permission, may be made public at PI's discretion
Tools/Code	Github	Fully public
MetaData**	Various locations	Individual metadata restricted to ACLEW members (or public at the PI's discretion), group level meta-data necessary to describe the samples is fully public.
Derivative data***	Github	Fully public in anonymized form unless there is a specific concern regarding impact on a vulnerable, Indigenous, and/or traditionally marginalized group. All anonymized derivative data that form part of a pipeline to a published manuscript must be public.

"by PI permission" refers to the PI of the particular sub-corpus

Metadata here refers to characteristics of the sample necessary to characterize it with respect to the populations under study (e.g. gender, education level) or for practical reasons (e.g. filenames, recording dates).
****Derivative data here refers to summary or processed data from the audio and/or annotation files that feeds into an analysis.***

Usage and Dissemination. Access is granted for ACLEW-related purposes as described in the overarching goals of the grant proposal (including analyses and data wrangling pipelines that further these goals). Usage for other (non-ACLEW) purposes is permitted only with the direct permission of the given datasets PI. If you are uncertain, please consult the datasets PIs.

Authorship. Authorship will be decided on a per-paper basis, based on this spreadsheet. An initial set of manuscripts and lead authors was explicitly listed in the grant proposal. Any subsequent conference submissions, proceedings, or journal articles using the ACLEW datasets should be known to all ACLEW PIs (well in advance where possible, but at the minimum 7 days for manuscripts, 48 hours for proceedings as described elsewhere). Authorship should be generally generously inclusive, with roles of each author delineated, where space allows (e.g. not in a 500 word abstract, yes in a journal article). More details on authorship guidelines are found here.

A special note regarding sensitive group data. Our corpora include groups that are vulnerable, Indigenous, and/or that have been traditionally marginalized. Data that are technically “anonymous” at the individual level may still contain information that has the potential to cause harm at the group level (e.g. group mean measures of caregiver characteristics that may vary across cultural groups and be interpreted according to our cultural norms/out of context). The datasets PI for each dataset ultimately has the authority within ACLEW (on behalf of the group their dataset originated) to determine allowable dissemination of group data of this type (and/or how it is discussed if dissemination is allowed), and the datasets PI in question should be consulted before dissemination. Group data with no conceivable value judgment possible (e.g. sample size) is not subject to such restrictions.

PIs sign and date your name below:
Melanie Soderstrom February 27, 2018, April 17, 2018, again
Okko Räsänen, March 6, 2018, April 23, 2018, again
Marisa Casillas Tice, March 6, 2018; April 20, 2018 again
Celia Renata Rosemberg March 8, 2018, April 13, 2018 again
Elika Bergelson March 12 2018; April 10 2018 again
Frank Rudzicz 13 March 2018; 23 April 2018 again
Björn Schuller 20 April 2018
Florian Metze - April 20, 2018
Alejandrina Cristia - April 26, 2018

Amendment to these datasharing ground rules, agreed by PIs on 13–14 March 2019

During the time of the project (until all planned and other mutually agreed-upon publications are completed, as defined in April 2018, see above) the audio files will be stored on servers owned by the Max Planck Institute of Psycholinguistics in Nijmegen, The Netherlands (henceforth "MPI"; storage referred to as "MPI secure storage" below). Media files will remain in Databrary until we verify that the MPI secure storage system fulfils ACLEW's needs. After that, each datasets PI may decide to remove their data from Databrary to avoid multiple copies if they wish.

It is forbidden to process any of the audio data on MPI resources without the express written consent of the PI responsible for the dataset(s) being processed—by default data may only be stored on MPI servers (basic upload/download). In the unlikely event that data are *processed* on the MPI server in error or without the permission of the dataset(s) PIs, that processed data is not *required* to be archived at the Language Archive (see details on archiving below).

MPI secure storage
The Max Planck Institute of Psycholinguistics in Nijmegen, The Netherlands has a private network of computers on which the data from all projects produced within the MPI is securely stored.

Who has access to the network? Only members of the MPI (employees and guests) have access to this network; in other words, files can only be directly accessed through an MPI account (e.g., jane.doe@mpi.nl + password). All MPI accounts are term-limited, which means that access expires on a pre-defined date. There is, however the possibility to apply for an extension of access. All initial MPI accounts and account extensions must be approved by a Max Planck director.

How do ACLEW members access the files? Marisa Casillas, a current MPI employee, will set up a project directory called "ACLEW_casillas" in her own workspace. Then read/write access to that directory will be granted to ACLEW members via an OwnCloud, a secure cloud service which can be linked to the ACLEW_casillas workspace on the MPI's servers. ACLEW members with current permission (see the Access List) will be able to log in with a password and securely upload/download files from the folder as needed, without access to the rest of the MPI internal network. Access to this folder is limited to an expiration date specified upon account creation (proposed end: 1 August 2020), but account extensions can be requested as needed to comply with the project-end needs (see the first sentence of this amendment).

Who will have access to our files? Only those with permissions access to the ACLEW workspace will be able to access our files. Directly, that includes (a) MPI employees affiliated with the project (Marisa Casillas, Caroline Rowland, and their trainees) and (b) network administrators at the MPI who are in charge of network technical upkeep. Indirectly, it includes ACLEW members who are approved on the access list.

Other details? The data are backed up in two other remote locations, also secure. Our memory limitation starts at 500GB but can be increased. The folder structure used for workspaces is set up to facilitate long-term data storage in the MPI's Language Archive (see below), though archival via the Language Archive is **not required** for any dataset unless that dataset's PI voluntarily decides to do so.

The Language Archive

The Language Archive (TLA) at the MPI is one optional long-term storage solution for post-ACLEW data archival. We are strongly encouraged, but not obligated, to store our data post-ACLEW at TLA. Note: ONLY in the case that a dataset PI willingly processes their data using MPI resources are they then obligated to archive with TLA**. In the case that a datasets PI *does* choose to archive at TLA, they can choose from multiple tiers of access permission for each file in the folder including:

- "Public": materials can be accessed by anyone without having to log in.
- "Authenticated Users": any user with a valid account for the archive can access them.
- "Academic Users": users that log in with an academic account or whose academic status has been verified can access them.
- "Private": means that the materials are only accessible to the depositor.

Access policies can also be refined later in consultation with the TLA archival team and MPI directors.

** In other words, simply ***storing*** files on the server comes with no strings attached. Anything else does, and constitutes a basis for file archival at the Language Archive.

PIs sign and date your name below:

Marisa Casillas Tice, 13 March 2019
Melanie Soderstrom, 13 March 2019
Celia Rosemberg, 13 March 2019
Björn Schuller, 14 March 2019
Elika Bergelson 14 March 2019
Florian Metze, March 13, 2019
Okko Räsänen, March 14, 2019
Alejandrina Cristia, March 15, 2019

Amendment to these datasharing ground rules, agreed by PIs on 25 March 2020

CSC storage

During the remaining time of the project (until all planned and other mutually agreed-upon publications are completed, as defined in April 2018, see above) the audio files will be stored on

servers owned by the CSC. CSC, an IT Center for Science, is a Finnish center of expertise in information technology owned by the Finnish state and higher education institutions. CSC provides data storage and processing services in terms of multiple servers, data warehouses, and computing clusters. CSC is a reliable partner whose data centres have been granted an ISO/IEC 27001 certificate for their information security management systems. The servers are located in Kajaani, Finland (visiting address: CSC - IT Center for Science Ltd / Kajaani datacenter, Renforsin Ranta business area, Tehdaskatu 15, Kajaani, Finland). For capacity services (ICT platforms) and certain customer services CSC complies with the raised information security level as defined by the Finnish Government (see https://www.csc.fi/en/web/guest/info/security).

Data will be stored in a folder specifically for the ACLEW project of the Puhti high-performance computing cluster. Access to the folder will be limited to researchers who have access to ACLEW project resources on CSC servers, as managed by project supervisor (Okko Räsänen). That includes the supervisor himself (Okko) and dedicated ACLEW project members who are helping to run data processing (e.g., Marvin Lavechin, Alex Cristia, Björn Schuller, Najla Al Futaisi)—all these people are already on the data access list. In addition, administrators of the computing environment have access to all data on the server.

Access to the CSC Puhti server will require registration either through Finnish university Haka identity federation system (Finnish researchers) or through a separate email-based registration (external users). A separate access right to ACLEW project data will be required from project supervisor (Okko Räsänen) to those other ACLEW members who will help run the data processing.

The storage areas utilized for the project are intended only for data that is in active use, which means that files that have not been modified for longer than 90 days will be automatically and permanently removed.

The ACLEW data processing plan involves making new, short audio files. The short audio files will be stored on the same network disk area than the original audio recordings (i.e., the ACLEW project folder on Puhti). The same security considerations apply to these short audio files as to the original data.

All data, the original audio recordings and the short audio files, will be removed from the server once the required analyses are conducted, but no later than August 21, 2020.

If a data contributor wants to completely and permanently remove their files, they can contact project supervisor Okko Räsänen (okko.rasanen@tuni.fi) for data removal at any time. The specified data will be removed from the CSC server within the shortest possible reasonable time frame.

PIs sign and date your name below:

Marisa Casillas Tice, 25 March 2020
Celia Renata Rosemberg, 25 March 2020
Melanie Soderstrom, 25 March 2020
Alejandrina Cristia, 25 March 2020
Elika Bergelson, 25 March 2020

Okko Räsänen, 27 March 2020
Björn Schuller, 1 April 2020