Datasharing addendum - additional considerations due to marginalized groups
(This is an addendum to the datasharing ground rules - please start there)

Types of data

Not anonymous

- raw audio - unvetted
- raw audio - vetted
- transcripts - un-anonymized
- audio-based derivative files which could now (or *reasonably* in the future) be recovered back to the original audio with sufficient fidelity to identify individuals by voice or content
- transcript-based derivative files which which could now (or *reasonably* in the future) be recovered back to the original source transcript in a manner that might identify individuals (e.g. based on the names)
- Individual metadata (except as noted below)

Individual level anonymous - not group level anonymous

- transcripts - altered/selected for anonymity (***N.B. These are treated separately***)
- segmentation/annotation files (no transcription)
- audio-based, transcript-based, or annotation-based derivative data for which no backwards recovery to individuals is possible/likely, but for which the data may be grouped in meaningful ways (e.g. by language community and/or SES). This would include, for example, a list of words ordered by frequency of use, ratios of IDS:ADS of utterances, organized by group.

Individual AND group level anonymous

- audio-based, transcript-based, or annotation-based derivative data for which no backwards recovery to individuals is possible/likely, and for which it is not possible for the data to be grouped in meaningful ways (e.g. by language community and/or SES). This would include, for example, a list of words ordered by frequency of use, ratios of IDS:ADS of utterances, organized in based on other kinds of features (e.g. infant age, infant gender), but NOT by group.
- Metadata spreadsheet ("ACLEW List of Corpora") ONLY:
    - corpus level tab: all fields
    - recording level tab: columns a-n, q-v, y-AB, AE, AF
        - I.e. all current columns as of March 23, 2018 **except**:
            - age (m;d), age_exact (Y/N),
            - mat_ed, fat_ed,
            - Mother DOB, Father DOB - these fields will be modified to allow sharing in a manner that is more anonymized.

Possible levels of access

- Level 0 - fully public, no restrictions
- Level 1 - available to researchers passing a minimal bar (Databrary/Homebank users)
- Level 2 - available upon request after Level 1 bar AND signing a statement of principles on the ethical use of group level data with marginalized populations
- Level 3 - available to the ACLEW team
- Level 4 - available only by direct approval of the PI responsible for that sub-corpus

Minimal Access Agreed to by all Datasets PIs - DURING the ACLEW project (i.e., until all planned publications are completed).

	Access?				
Type of data	**Level 0**	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Non-anonymous	No	No	No	Yes	Yes
Vetted/Anonymized transcripts	No	No	No	Yes	Yes
Individual but not group-level anonymous *involves a marginalized group AND **a potential ethical concern is articulated AND *** the data are not part of a publication pipeline	No	No	No	Yes	Yes
Individual but not group-level anonymous *involves a marginalized group AND **a potential ethical concern is articulated BUT *** the data are part of a publication pipeline	No	No	Yes	Yes	Yes

Individual but not group-level anonymous BUT any one (or both) of the following applies: *DOES NOT involve a marginalized group **there is no potential ethical concern articulated	Yes	Yes	Yes	Yes	Yes
Fully individual AND group level anonymous	Yes	Yes	Yes	Yes	Yes

Minimal Access Agreed to by all Datasets PIs as of April 5th 2018 - At project end (defined as all agreed to publications are complete) - TO BE FURTHER DISCUSSED IN AUGUST

- All derived data that form part of an analysis pipeline will be available for review by other researchers upon request (Level 2 or below, depending on data type)
- Raw audio and transcripts may be withdrawn to Level 4 - HOWEVER, all datasets PIs agree that should a question arise regarding a published paper that can only be resolved by examining the source data, any needed access to audio and transcripts will be provided.

Principles of datasharing for ACLEW

- The levels of access for the different data types described above are, by default, what is minimally agreed-upon by all ACLEW PIs - less stringent restrictions are always possible (and strongly encouraged) for particular subcorpora/data with individual PI approval. For clarity, PIs setting a less restrictive access policy for their particular sub-corpus should create a table parallel to this one for their sub-corpus
- Data collected (or derived from ACLEW data) with ACLEW funds should be shared at the most open level possible and should be only restricted if there is a *compelling* ethical reason (e.g. a legal custody battle leads to a court order removing parental status from the caretaker who signed the data sharing consent, or child requests data removal upon reaching legal adulthood age) - our project adheres to principles of open science
- The fact that data are shared at any level (including publicly) does not absolve researchers of appropriate attributions and/or authorships as described in the datasharing groundrules document.

- When considering the possibility of re-identification for a given data file, it is important to consider the potential for combining *across* data sources/pipelines
- Group level data ONLY create a source of ethical concern for data-sharing (assuming it is otherwise anonymous) if
    - It is from marginalized, vulnerable, or Indigenous groups **AND**
    - It has the potential to be interpreted in a manner that could have a negative impact on that group
- The PI of the given sub-corpus is acknowledged to be the best suited to make decisions regarding negative impact. Any concerns regarding access to particular aspects of the data should be raised early in the analytic process (see below). Concerns regarding the framing of findings in a manuscript should be raised as soon as the issue emerges.
- All datasets PIs agree that basic group-level quantitative data (e.g. amount of speech of a certain type heard by infants across groups) is necessary for the ACLEW project and have agreed to allow this data to be published.

Timeline for datasharing

1. **Stage 0:** Data are made available to the ACLEW team as needed to conduct research for the ACLEW project, as described in the main datasharing document.
2. **Stage 1:** At the point when a specific set of analyses for a manuscript is being determined, the relevant data pipeline is also determined. Data at this point should be organized and stored in a restricted location so that it is clear what data form part of the analysis and in what form. Any necessary discussions about level of access can then take place meaningfully and should be fully resolved before proceeding. External PIs (for the Warlaumont and LuCiD corpora) may need to be consulted at this stage.
3. **Stage 2**: At the point of peer review, the pre-existing dataset is made accessible according to the levels determined at Stage 1, in the locations described in the main datasharing document.

Examples

Please add suggestions for examples here that you think would be helpful to work through.