Plan Overview

A Data Management Plan created using DMPTool

Title: CSSI AnnotationBank

Creator: Marisa Casillas

Affiliation: University of Chicago (uchicago.edu)

Principal Investigator: Marisa Casillas, Michael C. Frank

Funder: National Science Foundation (nsf.gov)

Funding opportunity number: 20-592

Grant: https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505505

Template: NSF-SBE: Social, Behavioral, Economic Sciences

Project abstract:

The AnnotationBank is a resource for engaging international researchers in large-scale, multi-site campaigns to add annotations to the wealth of underutilized daylong child-centric recordings already available to the research community. This proposed cyberinfrastructure builds on HomeBank, a RIDIR-funded repository for daylong recordings, relying also on increasing field practices of large-scale and collaborative open science projects. The data garnered via AnnotationBank campaigns will shed light on previously invisible facets of children's developmental experience: The temporal breadth, sample size, and cultural diversity of existing daylong recordings, once annotations are added, offer critical information about what children do, when, where, and with whom, and will thereby re-shape our understanding of learning and development in the home environment. The same annotations would naturally serve as training and test data for new machine learning systems that automate many of these annotation tasks in the future for both applied and scientific use.

Start date: 09-01-2021

End date: 08-31-2024

Last modified: 10-07-2020

CSSI AnnotationBank

PIs Casillas and Frank, both experienced and early Open Science adopters, will lead the management and retention of research data in this project; they will supervise a dedicated software developer and project coordinator to aid in data management. Because this project builds on an existing community data repository (HomeBank), PIs Casillas and Frank are obliged to respect and preserve the access standards set by each sub-dataset contributor, which ranges from free public access to fully embargoed data. Because this project involves the creation of new annotations of the existing data, PIs Casillas, Frank, and their team are responsible for individually coordinating with HomeBank data contributors to ensure that plans for new annotation collection are aligned with the spirit of the access standards set by that contributor before any annotation campaign can proceed. Alignment with data contributor access standards includes: guarantees to secure storage, moderated user access, and long-term data maintenance, as outlined in the HomeBank rules at the point of data contribution; AnnotationBank must uphold these standards from top to bottom. To ensure that these responsibilities are adequately fulfilled, PIs Casillas and Frank must also work closely with the HomeBank leadership team in the construction and management of AnnotationBank. All analysis scripts, pre-prints, and derived, anonymized data used for scientific discovery, must be open to public access by other researchers for re-use and re-analysis. Both PIs Casillas and Frank are responsible for these tasks and, should one of them leave the institution or project, the other will take over the tasks entirely.

Two types of data are produced in the AnnotationBank system: Temporary HomeBank media storage following data requests and a comprehensive database of the cumulative annotations collected via AnnotationBank campaigns. "Media" here includes: short (< 5 min) clips of audio/video or brief streams of (<100) photos and/or image frames from video. Following a data request from a HomeBank user, media files are extracted from HomeBank using the user's credentials, prepared as requested (e.g., snipped into clips). Each media clip is securely stored on AnnotationBank until the user completes its online annotation or downloads it to their local computer, at which point AnnotationBank deletes its copy of the clip. The second type of data, "Annotations", begin with those already associated with HomeBank datasets and are iteratively added to via annotation campaigns using AnnotationBank. Annotations take the form of a database with time-aligned, text-based information relating to the original media (e.g., transcribed speech, descriptions of a scene, etc.); the size and scope of the database grows over the course of the project as new annotations are cumulatively incorporated. Annotation metadata includes the source of the annotation, including the date it was added and the unique and anonymized HomeBank member ID with whom the contribution is associated (ID keys are stored securely in a shared table hosted by HomeBank but accessible to AnnotationBank). Annotations are retained long-term as a corollary data source to the original HomeBank recordings (also long-term), in line with the wishes of the contributor for each dataset annotated. Similarly, HomeBank users' rights to data access extend to accessing AnnotationBank data relating to those same datasets they have access to on HomeBank. Long-term storage of these data is ultimately guaranteed through HomeBank via its host, Carnegie Mellon University. All other products from AnnotationBank, including software, pre-prints, analysis scripts, and derived and anonymized data from scientific analysis will be stored long-term in public repositories (GitHub, OSF).

As outlined above, all media storage following data requests is temporary, with each stored clip/image stream deleted permanently upon user download and/or online annotation contribution. All annotation data storage is long-term, cumulating in the AnnotationBank and kept as part of the datasets HomeBank users gain rights to upon gaining membership. Far long-term storage of HomeBank materials (which may include an archived version of the AnnotationBank database) is guaranteed via its host, Carnegie Mellon University. All other products including software, analysis scripts, pre-prints and derived and anonymized data from scientific analysis will be stored long-term for public access on OSF and GitHub. Contributed annotations following directed annotation campaigns (such

as those outlined in the proposal) shall have a negotiable default embargo time of 1.5 years from the point of data request before HomeBank users can freely access any contributed data; at this point any "checked out" clips to researchers will also be released for additional annotation by others. Requested media files stored temporarily on AnnotationBank will be stored for a negotiable default maximum of two weeks (offline annotation) or six months (online annotation) before deletion.

Media data formats include widely-used audio, video, and image types (e.g., .wav, .mp3, .mp4, .png, .jpg). Annotation data will be kept in a multi-tiered, structured database from which specific data tables can be queried (via a combination of SQL and our permissions-checking function). All other resulting scripts and scientific outputs will be shared in standard formats accepted by the scientific community for long-term re-use and re-analysis (e.g., with R/Rmarkdown, Python, and Bash with data tables in .csv and .txt UTF-8 formats). Dissemination of the annotation data will occur via HomeBank, as access to AnnotationBank data requires HomeBank credentials; with a growing body of >50 PI-level members, access to the resulting datasets will be widespread. By limiting access to those with HomeBank credentials, we can guarantee smooth continuation of the access privileges granted by the original data contributors. Dissemination of the other resulting scripts and scientific outputs (which include no personal identifying information) will be completely public, via repositories on OSF and GitHub. These research products (system software, analysis scripts, pre-prints, de-identified derived data for analysis) will be rendered under an Attribution-NonCommercial Creative Commons license as they are considered property of the research community at large; part and parcel of the project’s long-term sustainability plan.

AnnotationBank data and associated software will be hosted on hardware within the University of Chicago's Research Computing Center and affiliated secure data enclave, which provides top-notch tools for secure data archival and access (e.g., secure storage and encryption with MidwayR, a HIPAA-compliant secure computing enclave, multi-site redundant archival systems). Nightly AnnotationBank back-ups will ensure minimal data loss in the case of system failures. All other shared products (e.g., software, scripts, pre-prints, and derived data) will be shared via existing stable repositories with their own archival systems for public access (GitHub, OSF); frequent (>1 per day) commits to these shared project repositories also ensures a record of progress and project contributions, as well as a way to recover previous versions if necessary; all part of the record publicly accessible to future re-users and re-analyzers.

There are no further data management requirements from the NSF solicitation nor from local policies and best practices at the PIs' home institutions.
