Collaborative Research: Designing Expressive CS Learning Environments for Learners who are Visually Impaired

Data Management Plan

Roles and responsibilities: The PI and Co-PIs on this project will systematically manage the data regarding this research project. Georgia Tech PI Magerko is ultimately responsible for data management and will consult regularly with the project team to assure data is handled and stored appropriately. PI Magerko, co-PI Freeman, co-PI Ludi, and co-PI Pardo will oversee collection, analysis, storage, and dissemination of student research data. The research data will be accessible only via keycard and housed in Georgia Tech's Expressive Machinery Lab in the Technology Square Research Building.

Types of data or products: This project will produce a new version of the EarSketch product (EarSketch v3), as well as professional development (PD) materials for teachers. The types of data produced by this project are quantitative and qualitative data primarily focused on: (1) EarSketch v3 software development, (2) the student products and outcomes associated with EarSketch v3. Quantitative data will include student artifacts such as EarSketch scripts and surveys. Qualitative data includes students' worksheets and narrative artifacts, focus groups and interviews with students and teachers (to be audio recorded and transcribed), as well as notes from classroom observations. Qualitative data will be coded and analyzed according to best practices in qualitative and case study research. All coding schemes developed by the research team to categorize qualitative data will be captured. External evaluation data will be used to monitor internal research activities and to report on the project's progress.

Data format, storage, and retention: Staff time has been allocated in the proposed budget to cover the costs of preparing data and documentation for archiving. Data collection, access, and storage will follow protocols designed to protect the rights of human subjects and will be overseen and approved by the Georgia Tech Institutional Review Board (IRB). Informed consent of students, parental permission, and permission of other individuals involved in the project will be obtained prior to data collection, and researchers will follow all protocols approved by the IRB. Amendments will be added when necessary to obtain supplemental data. All the information collected will be confidential. ID numbers will be generated for each participant and used to tag all data. A master list linking ID numbers to names will be kept in a secure location, accessible only by research personnel, and housed within the Expressive Machinery Lab.Project-level evaluation data will include notes from project meeting observation, activity monitoring, and review of project documents.

Data and metadata will be generated following accepted scientific standards. Metadata, including the qualitative data coding schemes, variable names, and labels, will be captured. Qualitative and quantitative data will be coded using established software formats (e.g. Excel, SAS and SPSS) and standards. Data will be stored in password protected computers behind Georgia Tech's firewall as well as on a project directory structure on Box, and only research personnel approved by the IRB will have access to it. Box is a GT-approved platform for file storage, management, and sharing of sensitive data. Files will be viewable only to authenticated IRB-approved members of the team, will be encrypted using 256-bit AES (AES-256) and further protected by encryption key-wrapping, TLS 1.1 or TLS 1.2 encryption for files in transit, and daily automatic backups.

EarSketch student scripts are saved to the EarSketch database instance on Amazon Web Services (AWS), using the Amazon Aurora Relational Database Service (RDS), which is backed up hourly for 24 hours, weekly for 4 weeks, and monthly for 12 months. The database instance, including all logs, backups, and snapshots are encrypted at rest using AES-256. The AWS account is managed and secured by the Georgia Tech Office of Information Technology, with two-factor single sign-on authentication, centralized automated activity log shipping and monitoring, and secure campus VPC access to the AWS cloud infrastructure.

Analysis of aggregated data will be released via printed publications, conference presentations, and technical reports, following best practices in the field for data reporting and analysis. Publication of these analyzed data will occur during the lifespan of the project and after its completion. All of these systems will be in place for the 4-year minimum prescribed by the NSF, as well as for at least three years after the conclusion of the project or public release, whichever is later, with all data archived in the Box project directory structure, to be held by the PI. Copies of data on local hard drives will be backed up to the Georgia Tech CrashPlan Enterprise service.

Data Preservation: Scholarly Materials and Research @ Georgia Tech (SMARTech) will be used to archive the coded qualitative and quantitative data, as appropriate. SMARTech is an intellectual repository that is part of the MetaArchive Cooperative, a digital preservation network through the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP). Data will be archived in both Excel and comma-separated formats (.xlsx and .csv), and metadata will be archived in supplementary documents (.docx and .pdf). Data will be archived by 2 years after collection or at the conclusion of the project, whichever comes first. GTCMT will store hard copy and electronic files for the lifespan of the project.

Data Sharing: After the main research project findings have been accepted for publication the un-aggregated, qualitatively coded data and quantitative data (with appropriate metadata) will be made available upon request. These un-aggregated qualitative or quantitative data will be anonymous: they will only contain participant ID numbers. Access to raw (uncoded) video/audio, transcripts, interview, and questionnaire responses will be restricted to the project team due to IRB privacy restrictions.

We plan to share all teaching/PD materials as widely and freely as possible through our project website, except where restrictions due to intellectual property rights exist. The principal investigators for the project and their institutions will hold the intellectual property rights for the data. EarSketch, including its coding environment, digital audio workstation, and curriculum, will be available as a free online service. The service and related materials will be available via the EarSketch web site (https://earsketch.gatech.edu). By default, the music and code created by EarSketch users is private to them. It can only be shared if they choose to do so. EarSketch users can share their projects via a shareable URL or with a specific EarSketch user (such as a fellow student or a teacher). They can also download their projects as code or audio files to archive or to share through other platforms. EarSketch data collection practices are disclosed publicly to our users via the EarSketch Privacy Policy, located at https://earsketch.gatech.edu/landing/#/privacy.