**476 Co-designing Safety-Enhancing ADAS with Transit Operators: DATA MANAGEMENT PLAN**

**Types of Data Collected —**

**Survey Response data:** We will generate survey response data during our survey of transit operators. This data will include numerical answers to Likert-style questions as well as text answers to free-response questions. We will also collect basic demographic data about participants. We will not collect participant names as part of the survey but will collect their local union number and geographic location at the city and state level. Surveys will be administered on mobile or desktop computers using an online survey tool (Qualtrics). Participants will consent before completing the survey. Survey data will be exported and stored on CMU-managed Google Drive servers as .csv files.

**Audio, video, and photo data:** Audio data from the workshops and interviews will be collected. All participants will be asked to consent before the collection of this data. We will also ask participants for permission to show video clips from the diary study in publications, demonstrations, and talks. We will only display footage of participants who have explicitly consented to this use.

We will use a third-party transcription service to generate transcripts of the words spoken during the audio and video recording. These will be de-identified, using a participant number instead of a participant’s name. Video and audio files will use standard formats, including: Quicktime (.mov) and MPEG4 (.mp4). Transcript data will be stored as spreadsheets using .csv, .xls, or .txt files.

**Field notes**: We will generate field notes from our observations when conducting interviews and facilitating workshops with transit workers and related stakeholders. Field notes will only contain de-identified data, where participants will be referred to by a participant number as well as their social role (manager, operator) and demographic data. Field notes will be stored as text files using .doc or .txt files.

**Synthesis data:** We will use a variety of approaches to synthesize our qualitative and quantitative data into actionable insights. These include grounded theory, thematic analysis, and affinity diagrams. The data that comes out of these efforts will include codes generated as a part of the analysis, emergent themes, insights about the meaning of the data, and opportunities for new research and development of new technologies or policy recommendations. The synthesis data will consist of tables, lists, descriptions, and rationales. This will not contain any identifiable information. Synthesis data will be stored as documents and spreadsheets using .csv, .xls, .doc, or .txt files.

**Metadata Standards —**

The team will utilize common data structures and file collections for data we gather. All data gathered directly from participants will be stored in traditional CSV and plaintext file formats alongside data specification documents for easy interpretation of data files. All video collected will be stored in popular digital formats (e.g., Quicktime [.mov] and MPEG4 [.mp4]).

**Policies for Access and Sharing —**

All human subject research and sharing of data created via human subjects research will follow up-to-date IRB protocols, as established by Carnegie Mellon University. All research team members will complete the CITI training on human subjects research, and the team will regularly review their human subjects procedures and handling of participant data during team meetings. IRB protocols will include policies on data sharing.

**Policies and Provisions for Data Sharing and Re-Use —**

We will make all of the de-identified data available to other researchers and to the general public. We will make the data available via a project web site hosted at Carnegie Mellon University. In some potential cases, we will work to increase the reach of our data by making it available through commercial services, such as GitHub.

**Data Retention, Access, and Storage Facilities —**

The project team will use only CMU-authorized servers for long-term storage of all data, and they will rely on archiving procedures developed by the College (School of Computer Science). All servers have transparent and password-protected access to the Google drive System and the Andrew File System, a shared file space, and to one another through the Network File System protocol. The HCII department maintains terabytes of secondary storage with regular backup. Another mechanism we will use for archiving data will be analyzing and publishing our results in open scientific literature. The team typically retains all data for at least five years beyond the end of each study to support longer publication cycles and students still working on their PhD theses.