The PennSTART Safety Standards Project

Data Collection

What data will you collect or create?

The primary data we will collect include: the safety plans submitted by users of the PennSTART facility, detailing their test scenarios at the site and the safety measures. We also collect data in terms of the actual tests and outcomes. In summary we collect information about both planned and observed uses of PennSTART facility in terms of test scenarios and safety measures.

How will the data be collected or created?

The data will be collected mostly through xlsx, docx, and pdf formats via communications between the CMU research team and the main stakeholder (RIDC). The data and information on the CMU side will be securely stored and versioned in the Box platform. For private information that can not be stored in the cloud, RIDC and CMU will communicate prior to storing the data and keep the information in local computing devices.

Documentation and Metadata

What documentation and metadata will accompany the data?

The data will be in text and numeric format in docx, xdlx, and pdf formats. The main tools used will be Microsoft Office and Adobe Reader.

Ethics and Legal Compliance

How will you manage any ethical issues?

All collected data will be anonymized to ensure participants cannot be directly identified. In cases where data needs to be matched with its source, pseudonyms may replace actual names, with a secure master list maintained separately. All sensitive data will be encrypted, both during storage and transfer, and strict access controls will be implemented to ensure only authorized personnel can access it.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

The data collected during this project will be owned by RIDC. Any information and data derived

from the original source will be owned by CMU. If necessary, Intellectual Property Rights (IPR) will be specified in a consortium agreement. Adhering to institutional and funder guidelines, the data will be licensed under the an open source license [e.g., "Creative Commons Attribution 4.0 International License (CC BY 4.0)"] for open reuse with attribution. Third-party data will be used respecting its original licensing terms, ensuring no unauthorized sharing or redistribution. Occasionally, data sharing might be delayed due to pending publications or patent applications, safeguarding the research team's intellectual contributions before broader data dissemination.

Storage and Backup

How will the data be stored and backed up during the research?

We will utilize the Box cloud service for our data storage needs, ensuring ample space and eliminating the necessity for additional charges. Box provides automatic backup features, ensuring data redundancy and minimizing risks associated with data loss. Our designated IT team member will oversee the backup process and recovery protocols. In the event of data loss or any other incident, the recovery process will be initiated through Box's built-in recovery features. This approach, while automated, aligns with our institution's guidelines, ensuring data is held securely and in compliance with all relevant policies.

How will you manage access and security?

Data security risks, including unauthorized access and data breaches, will be managed using stringent encryption protocols and multi-factor authentication provided by our primary storage solution. Access will be strictly controlled, granting permissions only to authorized personnel. Collaborators will access data through secure channels, ensuring data integrity and confidentiality. For data collected in the field, secure transfer methods, such as encrypted VPNs, will be employed to safely integrate it into our primary systems. When handling confidential data, we'll adhere to recognized standards, including compliance with ISO 27001, to ensure maximum security and data protection.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

For the PennSTART initiative, data essential for contractual, legal, or regulatory purposes will be mandatorily retained, while data deemed non-essential will be evaluated for its potential research value, especially in supporting the safety standards research for the Connected Deployment

Corridor. Foreseeably, the data will be instrumental in validating research findings, informing new studies, and potentially serving as a teaching resource. The data will be preserved for a minimum of two years to align with the initial launch of PennSTART uses. Based on its continued relevance and potential for further research applications, the retention period might be extended to support ongoing and future projects. Decisions on data retention will balance its potential reuse value, economic feasibility, and any preparatory efforts needed for data sharing and preservation.

What is the long-term preservation plan for the dataset?

The data generated from our project will be securely stored and managed using the Box cloud service, a reputable and established repository. Currently, the costs associated with our selected storage solution, Box, are covered under our institutional agreement, ensuring no additional charges for our project's data storage needs. We have allocated resources and budgeted time to prepare, document, and curate the data for both sharing and long-term preservation. Recognizing the importance of data longevity beyond the grant's duration, our choice of Box ensures that the data remains accessible and curated effectively, benefiting future research endeavors and potential collaborations.

Data Sharing

How will you share the data?

Our data will be made discoverable through references in publications, institutional repositories, and project-related communications. We intend to share the data with researchers, collaborators, and other stakeholders, ensuring that access is granted under specific conditions to maintain data integrity and confidentiality. The primary mechanism for data sharing will be through the Box cloud service, which provides secure and controlled access to authorized users. We anticipate making the data available post-analysis and upon the conclusion of the project's primary phase. Additionally, we will pursue obtaining a persistent identifier for our data, ensuring its traceability and consistent referencing. By adopting this approach, we aim to foster a culture of acknowledgment and citation for data reuse, building on our commitment to open and collaborative research.

Are any restrictions on data sharing required?

While we aim for open data sharing, certain restrictions may arise due to confidentiality, intellectual property rights, or the absence of consent agreements. To address these challenges, we will implement rigorous data anonymization and pseudonymization techniques, ensuring that sensitive information remains protected. We anticipate needing exclusive use of the data during the project's primary phase for analysis and validation, approximately two years. A data sharing agreement will

be established, outlining the terms of access, use, and redistribution. In cases where data holds significant confidentiality concerns, we may consider non-disclosure agreements to provide an added layer of protection, ensuring that our commitment to data security and participant privacy remains uncompromised.

Responsibilities and Resources

Who will be responsible for data management?

The responsibility for implementing the Data Management Plan (DMP) is jointly shared between RIDC and CMU. In our collaborative research framework, data management activities will be delineated based on expertise and infrastructure capabilities of each partner. For instance, CMU, with its research-centric infrastructure, will be pivotal in data analysis and quality checks. Data ownership and responsibilities concerning research data management will be explicitly detailed in a consortium agreement, ensuring clarity and accountability. Both RIDC and CMU are committed to upholding relevant policies, with designated individuals from each institution ensuring the seamless execution of the DMP.

What resources will you require to deliver your plan?

To effectively deliver this data management plan, we may require specialized software for data anonymization and robust encryption protocols. While our existing staff possesses foundational knowledge, additional training or potentially hiring specialists might be necessary to handle advanced data management tasks. Hardware enhancements, particularly secure servers with high storage capacities, might be essential to accommodate the data. Lastly, while our primary data repository, Box, is covered under our institutional agreement, any additional or specialized repositories might incur charges, which will be factored into the project's budget. All these resources are crucial to ensure the plan's successful execution and data's integrity.