Data Management Plan (DMP)

This data management plan (DMP) describes how the Safety through Disruption (Safe-D) National University Transportation Center (UTC) will comply with the U.S. Department of Transportation (USDOT) policy on the dissemination and sharing of research results. The DMP defines how Safe-D researchers will handle digital data both during and after a research project.

Data Description

Fueled by the inevitable changes in our transportation system, the Safe-D UTC endeavors to maximize the potential safety benefits of disruptive technologies through targeted research that addresses the most pressing transportation safety questions. Comprising the Virginia Tech Transportation Institute (VTTI), Texas Transportation Institute (TTI), and San Diego State University (SDSU), the Safe-D UTC will proactively promote safety through a data-driven collaboration among the nation's brightest researchers. Research projects will generate new data and will analyze new and/or existing data across four Safe-D theme areas: automation, connectivity, transportation as a service, and big data analytics. Data captured during the duration of the Safe-D UTC program are expected to contribute to the existing body of knowledge related to transportation safety and serve as the impetus for new research ideas and programs. As a result, the public will benefit from an improved transportation ecosystem stemming from the widespread use and application of these data and associated research.

A number of research projects will be conducted under the Safe-D UTC. While the members of the Safe-D consortium have a variety of existing data sets that may be mined, original data will also be collected as part of research activities. Data collection within each project will be completed in accordance with the timeline approved by the Safe-D directors prior to funding. Across Safe-D projects, all data will be collected during the UTC grant period, as approved by the USDOT. Original data will be collected in a manner that maximizes the utility of existing data, avoids unnecessary duplication, and leverages the knowledge gained from past efforts. Safe-D researchers will record new data using a variety of experimental methods, including laboratory (e.g., driving simulation, crash sled, and modeling), test-track (e.g., controlled track testing), and field (e.g., naturalistic driving) experiments. Some of the data will be observational in nature, particularly those data sets resulting from naturalistic driving data collection methods. Safe-D research projects are expected to produce a myriad of data types, ranging in scale from compact tabular data sets to large-scale complex naturalistic driving data. Data types will include numerical data, images, video, questionnaires, and audio. Computer code and models may also exist within the outputs of the Safe-D UTC research projects. Potential users of the Safe-D UTC data will include students, academic faculty, system designers and developers, road operators, and other individuals interested in conducting transportation safety research.

A subset of Safe-D research projects may receive partial funding from industry partners. These projects may generate data considered proprietary. If it is determined that this type of data collection will occur, the portion of the project under which proprietary data collection is completed will be funded entirely by the industry partner, such that the data will continue to be owned by the partner and will not be generated using any federal funding. In such cases, the data set will not be subject to public access requirements and will not be shared through the means described within this DMP. However, the Safe-D UTC expects that some of these industry partners may make some of these restrictions temporary and will allow such restrictions to expire once protection of the proprietary element(s) of data collection are no longer necessary. In addition, elements of data that contain personally identifying information may be removed prior to making the data set publicly available, in accordance with the Institutional Review Board requirements associated with the data collection efforts.

The principal investigator(s) of each research project will be responsible for managing all data throughout the life of the project, including all activities related to the preparation of the data for public access. After the data are archived in a repository and it has been verified that the data comply with pre-defined minimum requirements, responsibility will then be transferred to the curators of the repository, as described below.

The Safe-D Program Manager will guide research project principal investigators through compliance with the Safe-D DMP. Safe-D research teams will also have access to guides and checklists on the internal Safe-D Researcher Portal that will chaperone them through the compliance process. Compliance is ensured by the Safe-D administrators through review of quarterly reports detailing the technical and budgetary progress of the projects, including any issues encountered; bi-annual surveys, which will feed into larger grant-reporting requirements; and a project closure report confirming that all activities conducted under the project comply with Safe-D requirements, including this DMP.

Data Format and Metadata Standards

In general, data will be collected in a variety of formats determined by each Safe-D research team and will be converted into an open-source format prior to making them publicly accessible. When warranted by extenuating circumstances, data may occasionally reside in a custom format. However, such a format should be readily readable by a commercially available software package. In some instances, specialized research (e.g., vehicle simulations) requires the use of specialized software, which generates data in a proprietary format. Tools and software applicable to the data will vary based on the data format of the associated project but it is expected that data will generally be viewable through open-source software. Aforementioned exceptions to this, while uncommon, may be present for some specialized research.

The data generated by Safe-D research projects will be uploaded and archived in the VTTI Dataverse, a data repository maintained by VTTI based on the Dataverse platform. This infrastructure facilitates data archiving and sharing through the use of persistent identifiers, data set version control, and adherence to metadata standards. Access to the repository catalog is available through https://dataverse.vtti.vt.edu/. Safe-D UTC data sets will be versioned following introduction into Dataverse. To be published, any updates to the data sets will require a classification of "minor" or "major," and versioning will advance automatically based on this classification. To ensure that versioning is handled consistently, only data repository curators will be able to publish new versions of existing data sets in collaboration and coordination with the Safe-D research teams. After data sets are uploaded to the VTTI Dataverse, curators will verify that each data set complies with this Safe-D DMP. Prior to publishing, each principal investigator will be required to verify that the public data set produced matches his or her expectations and is an accurate representation of what was provided from the project.

Safe-D researchers will ensure that the data archived are understandable and usable by other researchers through the creation of descriptive products (including metadata), an explanation of the data collection method(s), and a data dictionary. These products will be managed and stored in the VTTI Dataverse, which uses standard-compliant metadata that can be mapped easily to standard schemas and exported into JSON format (XML for tabular file metadata) for preservation and interoperability.

Access and Sharing Policies

With the few exceptions noted above, all data collected by the Safe-D UTC will be made accessible via the VTTI Dataverse. Data made available to the public will not contain proprietary or confidential information. De-identification of data will be a necessary and required step prior to making data publicly available. De-identification includes removal of any potentially identifying information (PII) such as names, location, or other participant/data attributes which could possibly tie the data back to a participant. Because of this de-identification step, the resulting data made available to the public would not raise any concerns regarding privacy, ethics, or confidentiality.

VTTI has significant experience in data de-identification, which will be used in the review and censoring of potentially identifying data prior to publication. For example, the forward view of rear-end crashes captured during naturalistic driving studies often shows the license plate of the lead vehicle, which could subsequently be used to identify the participant through crash records. However, VTTI ensures the license plate is blurred; an example of this approach can be seen in the forward crash videos on the InSight website, which serves as a visualization tool for the SHRP2 data housed on Dataverse (https://insight.shrp2nds.us/data/category/events#/table/38/ [requires additional log in]). Another approach also used on InSight is to aggregate identifying data to a degree that makes re-identification difficult or impossible. This can be seen in the heat maps for the data collection locations in which thresholds for minimum number of trips (and minimum number of distinct drivers) were applied before data for a particular road segment could be shown on the map (https://insight.shrp2nds.us/data/category/trips#/map/1).

Re-use, Redistribution, and Derivative Products Policies

After data are uploaded to the VTTI Dataverse repository, data management rights will be transferred to the curators of the VTTI Dataverse. Uploaded data will not have any associated intellectual property rights transferred to the data archive. Following upload to the VTTI Dataverse, Safe-D data will be made available for open sharing under the Creative Commons Zero (CC0) universal public domain dedication. Under CC0, data and derivative products will be available for reuse and redistribution without restriction.

Archiving and Preservation Plans

As described above, data will be archived on the VTTI Dataverse repository at https://dataverse.vtti.vt.edu/. The VTTI Dataverse meets the criteria outlined in the Guidelines for Evaluating Repositories for Conformance with the DOT Public Access Plan. The VTTI Dataverse promotes an explicit mission of digital data archiving, which is described on the Dataverse website (http://www.dataverse.org), and is listed by the USDOT as a Data Repository Conformant with the DOT Public Access Plan at https://ntl.bts.gov/publicaccess/repositories.html.

Data set preparation and submission to the VTTI Dataverse repository for archiving will occur prior to Safe-D research project end dates defined in each research project timeline. Upon publication, each data set will be assigned a Digital Object Identifier (DOI). The VTTI Dataverse uses the EZID DOI minting service. The use of other identifiers (e.g., ARK) is also supported, if deemed necessary for a research team. At a minimum, the VTTI Dataverse will retain the data through the duration of the Safe-D UTC program. At the end of the program, if UTC funding does not continue, it is VTTI's intention to continue to support the SAFE-D UTC Dataverse as part of its own institutional Dataverse. If this were not possible due to funding constraints, the Dataverse would be migrated to Virginia Tech's Library Services, who hosts its own data repositories. In any of these eventualities, the expectation is that the data will be available for a minimum of 5 years from program inception.

Prior to archiving on the VTTI Dataverse, Safe-D research project data will be stored on the data management systems of the consortium institution that is collecting the data. Back-up, disaster recovery, off-site data storage, and other redundant strategies are used by each consortium member institution, which protects data from accidental or malicious modification or deletion. Safe-D research teams will follow the usual processes employed by each consortium institution for these purposes. Once data are archived, the VTTI Dataverse will handle these processes. The VTTI Dataverse is hosted by Amazon Web Services in a scalable and redundant platform. In addition, snapshots of the VTTI Dataverse environment are taken periodically to ensure resiliency in the unlikely event that restoration is required at any point.