Safety Research Using Simulation (SAFER-SIM)
University Transportation Center
Data Management Plan

Creator(s): Dawn Marshall, Jacob Heiden
Affiliation: University of Iowa
Last modified: February 15, 2017

Data description

The SAFER-SIM UTC will require each individual researcher to submit detailed data descriptions for their individual research projects per this plan as outlined in the guidance. Individual data management plans will:

1. Name the data, data collection project, or data producing program.
2. Describe of the purpose of the research.
3. Describe the data that will be generated in terms of nature and scale (e.g., numerical data, image data, text sequences, video, audio, database, modeling data, source code, etc.).
4. Describe methods for creating the data (e.g., simulated; observed; experimental; software; physical collections; sensors; satellite; enforcement activities; researcher-generated databases, tables, and/or spreadsheets; instrument generated digital data output such as images and video; etc.)
5. Discuss of the period of time data will be collected and frequency of update.
6. If using existing data, describe of the relationship between the data you are collecting and existing data.
7. List potential users of the data.
8. Discuss the potential value of the data have over the long-term for not only SAFER-SIM, but also for the public.
9. If you request permission not to make data publicly accessible, explain rationale for lack of public access.
10. Indicate the party responsible for managing the data.
11. Describe how you will check for adherence to this data management plan.

Data format and metadata standards

Data gathered from transportation-related research varies and includes, but is not limited to the following: response times, gap choices, speed, time-to-collision, travel times, vehicle miles traveled, crashes, signal timings, video logs, land-use, infrastructure sensors, traveler behavior, driver behavior, and trip generation. The data is typically found the formats listed below:

- MS Excel (.xls)
- Video files (.xml, .csv, .mpg, .avi, .mov, .wmv)
- MS Excel Macro (.xml)
- Comma Separated Values (.csv)

- Portable Document Format (.pdf)
- Joint Photographic Experts Group (.jpg)

Researchers will be required to report the formats used when gathering data and they must list if they are open or proprietary. If using proprietary data, the SAFER-SIM Center will require the lead researcher to provide a rationale. It is expected that researchers will include in their reports how that data has varied, if any, from its original format.

Data gathered from SAFER-SIM funded projects must adhere to the standards used in the transportation industry. Metadata will be included with each dataset that describes the context, content, and structure of the data. One metadata schema that may be used to describe the data includes, but is not limited to, the National Transportation Library's Dublin Core Metadata Guidelines (https://ntl.bts.gov/tools/metadata). Researchers may use a nonstandard schema when necessary but must document rationale in final report. Metadata will be managed and stored similarly to collected data, either in a separate file or included in the dataset. Principal investigators will manage metadata before, during, and after data collection. Metadata will be stored in the SAFER-SIM Harvard Dataverse repository with all other data.

Tools or software required to read or view the data will be described in the project technical report. The SAFER-SIM Center expects that the data generated by funded projects is of good quality, and because of the interdisciplinary nature of transportation engineering, these standards will vary. The principal investigator of each individual project is responsible for ensuring the data is accurate and complete.

All principal investigators for individual projects will be required to:

1. Have final datasets that are not proprietary in a standard data format of the field, such as csv.
2. If principal investigators are using proprietary data formats, they will be required to discuss their rationale.
3. Include metadata describing the context, content, and structure of the final version of data shared to the public.
4. Describe how they will document the alternative formats they are using and why.
5. List what documentation they will be creating in order to make the data understandable by other researchers.
6. Indicate what metadata schema they are using to describe the data. If the metadata schema is not one standard for their field, and discuss their rationale for using that scheme.
7. Describe how the metadata be managed and stored.
8. Indicate what tools or software is required to read or view the data.
9. Describe their quality control measures.

Policies for access and sharing

The principal investigator is responsible for how the data is managed and secured during the experimental process. Once the project is completed, the data will be publically available via the SAFER-SIM repository in the Harvard Dataverse (https://dataverse.harvard.edu/dataverse/safersim). SAFER-SIM researchers are required to upload their data within 60 days of their project end date. Because some transportation-related research requires the use of human subjects, permission from the

Institutional Review Board (IRB) where the research originated will be obtained prior to publishing onto the public sites for data sharing.

Principal investigators will be required to address any access restrictions in the project data management plan they submit to the SAFER-SIM Center. For individual project data management plans, principal investigators will address issues and outline the efforts they will take to provide informed consent statements to participants, the steps they will take the protect privacy and confidentiality prior to archiving their data, and any additional concerns (e.g., embargo periods for their data). If necessary, they will describe any division of responsibilities for stewarding and protecting the data among other project staff. If principal investigators will not be able to de-identify the data in a manner that protects privacy and confidentiality while maintaining the utility of the dataset, faculty will describe the necessary restrictions on access and use. If an individual research project includes human subject research, researchers will be required to go through University of Iowa IRB or their home institutions IRB, if they have one.

Principal investigators will be required to address the following:

1. Describe what data will be shared, how data files will be shared, and how others will access them.
2. Indicate whether the data contain private or confidential information. If so
    a. Discuss how you will guard against disclosure of identities and/or confidential business information
    b. List what processes you will follow to provide informed consent to participants.
    c. State the party responsible for protecting the data.
3. Describe what, if any, privacy, ethical, or confidentiality concerns are raised due to data sharing.
4. If applicable, describe how you will de-identify your data before sharing. If not:
    a. Identify what restrictions on access and use you will place on the data.
    b. Discuss additional steps, if any, you will use to protect privacy and confidentiality.

Policies for re-use, redistribution, derivatives

University of Iowa or the home institution of the principal investigators holds the intellectual property rights for all data generated by SAFER-SIM funded projects. The data is also subject to the General Provisions of Grants for 2016 University Transportation Centers, Item #16 (Patents and Copyrights), pages 11-15. All data gathered and collected, which are provided publically, are the property of the PI. The PI will indicate in the final report, which, if any rights are to be transferred to the data archive. The PI must also indicate in the final report how the data will be licensed for reuse and redistribution.

If using proprietary data, principal investigators will be required to cite the data source and license under which they used the data in their project data management plans.

In general, principal investigators will address the following in their project data management plans:

1. Name who has the right to manage the data.
2. Indicate who holds the intellectual property rights to the data.
3. List any copyrights to the data. If so, indicate who owns them.
4. Discuss any rights be transferred to a data archive.
5. Describe how your data will be licensed for reuse, redistribution, and derivative products.

Plans for archiving and preservation

Principal investigators or their delegate will responsibly manage data before, during, and after data collection. Principal investigators will ensure data management meets their Institutional Review Board's standards. The SAFER-SIM UTC will archive all data on Harvard's Dataverse, https://dataverse.harvard.edu/dataverse/safersim, which is an approved site of the USDOT. Principal investigators will have 60 days following the end date of their project to archive their data on Dataverse. Principal investigators will maintain the data until it is uploaded to Dataverse. Principal investigators will describe how data will be protected from accidental or malicious modification or deletion prior to receipt by the archive.

Dataverse is an approved data repository by USDOT (https://ntl.bts.gov/publicaccess/repositories.html). Harvard University Information Technology (HUIT) in collaboration with Harvard Library, and the Institute for Quantitative Social Science (IQSS) hosts Harvard's Dataverse repository and maintains a full backup of all data and directories. This means that there is always a full, recent off-site copy of the Dataverse repository. Dataverse's policy for digital archiving is part of the institution's general mission to preserve all of its archival collections and to ensure their availability for current and future use. More specifically, this policy for preserving our digital data collections is meant to ensure continued access to born digital and digitized data, to ensure their authenticity, and to maintain data quality using the best digital archival practices. The repository backs up all of the application/system files and databases nightly. It is stored off-site for 45 days. All research data files in the repository are replicated every 4 hours to a second off-site storage array. Data content of the Dataverse into the DRS Storage Infrastructure, which makes use of storage management software to create a tape copy of data to be stored for the long term at the Harvard Depository. Dataverse preservation policy can be found here: http://best-practices.dataverse.org/harvard-policies/harvard-preservation-policy.html

Once a dataset is published, the repository guarantees archival and long term access to that dataset with a DOI persistent identifier provided by the California Digital Library's (CDL) EZID service (DataCite member). In order to ensure long term accessibility of the dataset in the Harvard Dataverse, once a dataset is published it cannot be unpublished and can only be deaccessioned under extreme circumstances, such as a legal requirement to destroy that dataset.