

SHRP 2 Safety Data, Update 2**April 10, 2015**

This is the second in a series of updates about Phase 1 of the implementation of the SHRP 2 Safety Data by TRB in collaboration with its partners, Federal Highway Administration (FHWA), the American Association of State Highway and Transportation Officials (AASHTO), and the National Highway Traffic Safety Administration (NHTSA). As was noted in the first of these updates, “Phase 1” is governed by a Safety Data Oversight Committee (SDOC). The SDOC recently held a meeting at the Beckman Center in Irvine, California to map out the next steps of the five year effort.

“Phase 1” is an experimental program designed to make the safety data as widely available as possible to qualified researchers and to gather information needed to lay the foundation for use of the data beyond the first five years. Several efforts have recently been concluded to help inform the SDOC for its governance activities. Three of these are:

- An Environmental Scan conducted by TRB staff, who interviewed operators of large databases around the United States.
- A draft and not yet published Environmental Scan conducted by US Department of John A. Volpe National Transportation Systems Center for the FHWA Turner Fairbank Highway Research Center (TFHRC) which focused on the development and operation of remote, secure data enclaves.
- A TRB Request for Information (RFI) which was issued in January 2015 to request input from potential data users and stakeholders about the future of SHRP 2 Safety Data and Phase 1 operations.

Selected highlights of the results of each are summarized below. These “takeaways” were presented to the SDOC and are now being used to move the program ahead during Phase 1.

TRB Staff Environmental Scan Takeaways

- Downloads of original data are generally not permitted. Cleansed, de-identified public use datasets are available for download. Public use datasets need user community input to most effectively meet their needs.
- Various statistical safeguards such as randomization and GPS coordinate clipping are used to protect participant privacy (e.g. personally identifying information, or “PII”).
- Synthetic dataset techniques can be used create data for public use, but require extensive validation.
- Database operators develop and maintain a clear process for data requests and approval, including an expert review committee or staff review. Several database owners require that data use requests be consistent with the purposes of the dataset.

- Access to any sensitive data occurs via a secure data enclave (SDE) or not at all (i.e. only database operator staff can touch PII). Some operators place the onus on researchers to create a secure environment for sensitive data that is made available to them.
- Data base owners often conduct a disclosure risk review of any derived data released from secure data enclaves to ensure that the derived data are adequately aggregated and de-identified.
- Researchers return or destroy the original data once they are finished and have published research results.
- User fees are charged by some database operators to help sustain the data assets; some are flat application or use fees and others are more directly tied to recovery of labor costs to construct custom datasets. There is an inherent tension in charging fees between encouraging use and promoting data resource sustainability.
- A process exists for citing use of the database and tracking database use and publications generated through use of the data.

Turner Fairbank Highway Research Center Environmental Scan Takeaways

- Providing different levels of data access based on user needs is a good practice. These levels could include everything from web sites containing summary data to secure data enclaves.
- A key to success is user-driven investment in tools (e.g. websites, guides, documentation, and software) and training for users. Technical assistance is ubiquitous but levels vary based upon resources.
- Public use data sets and summary datasets support more widespread use of data. Databases with confidential data had access models similar to those in TRB scan: a data request review, data analysis only in a secure environment, and an output disclosure review.
- Some federal data centers charge user fees.

TRB Request for Information (RFI) Stakeholder Recommendations and Takeaways

- Explore collaborative consortium-based approaches to curating the SHRP 2 data and combining it with data from other naturalistic driving studies and related datasets to leverage expenses and broaden utilization.
- Promote an active user community that openly shares the approaches to and results of their work including software tools, analysis techniques, and reduces datasets. Consider peer review of shared datasets to ensure quality control.
- Secure external source(s) of funding to maintain the database infrastructure following Phase 1. User fees cannot be expected to fully cover operating costs without making data access and use prohibitively expensive.
- General user access can be satisfied with reduced datasets, for example de-identified public use datasets.



TRANSPORTATION RESEARCH BOARD

OF THE NATIONAL ACADEMIES

- Expert users will want direct data access, but these represent a minority of users. Any remote data access must maintain data security and PII protection equivalent that maintained for the central database
- Stakeholder community engagement is crucial, especially in terms of generating and sharing data extracts (public use datasets) and developing and sharing common tools and lessons learned from using the data.