Data Management Plan
Version 1
October 30, 2020

Introduction

This document constitutes a Data Management Plan (DMP) for the Transit – Serving Communities Optimally, Responsibly, and Efficiently (T-SCORE) Center, a US Department of Transportation Tier 1 University Transportation Center (UTC) led by the Georgia Institute of Technology. Center partners include the University of Tennessee – Knoxville, the University of Kentucky, and Brigham Young University (BYU). The center aims to define a set strategic visions that will guide public transportation into a sustainable and resilient future, and to equip local planners with the tools needed to translate their chosen vision into their own community. In order to achieve this goal, the center is undertaking two distinct research tracks: a Community Analysis (CA) track focused on understanding the trends and challenges transit agencies and riders face, and a Multi-Modal Optimization and Simulation (MMOS) track focused on developing computational planning tools that agencies can use in these efforts. The two tracks will generate, obtain, and use data in different ways, though both tracks are committed to sharing data, products, publications, and presentations with the transportation community. Through the sharing of information, the center hopes to build a repository of resources that researchers and scholars will use for a period of time that extends well beyond the lifetime of the center. The following sections describe the protocols and procedures that will be followed at T-SCORE in relation to data protection, data sharing, and data dissemination.

Data and Products

The data collected and created/assembled as part of T-SCORE research activities will have long-lasting value. Researchers and scholars around the world may be interested in analyzing the data and extending the simulation work in other contexts. T-SCORE believes that it would be of value to preserve the data for long-term access well beyond the life of the T-SCORE UTC.

In order to better understand the challenges faced by transit agencies, the CA track will conduct surveys and interviews with transit agencies and experts who can expand on the strategic challenges targeted by the center. Through these efforts, T-SCORE researchers will gather open-ended qualitative data on the perspectives of these agencies. These interviews will be conducted only upon receiving any necessary approvals from the Institutional Review Boards (IRB) at the respective institutions. Transit agency and expert interview data will include, but not necessarily be limited to:

- Input on two primary transit visions, i.e., transit as a social good and optimizing transit ridership
- Statements on goals or objectives a transit agency may (or should) pursue
- Insights on transit as a social good
- Insights on transit systems for maximizing efficiency or ridership
- Comments on transit agency strategies for effective balancing these goals

Researchers on this track may also conduct surveys of transit riders in regions that have implemented mobile fare payment technologies. These surveys are likely to involve online and /or mail-back paper surveys and will be conducted only upon receiving any necessary approvals from the Institutional Review Boards (IRB) at the respective institutions. Transit rider survey data could include, but not necessarily be limited to:

- Socio-economic and demographic data of household/persons
- Transit trip frequency and purpose
- Transit access mode
- Transit fare value and payment method
- Traveler attitudes and values data, which includes information about how individuals feel about various mobility options, built environments, travel experiences, and the environment.

In addition to collecting data through surveys, T-SCORE researchers on the CA track will also use various existing survey and census data sets. These include, for example, the National Transit Database (NTD) data set and the American Community Survey (ACS) of the Census Bureau.

The CA track may also request some ridership-related datasets directly from transit agencies or other organizations that provide estimates of transit ridership. This could include, but not be limited to, route-level transit ridership counts, automated passenger counting (APC) data about boardings and alightings, and estimates of transit ridership based on usage of transit-related smartphone apps. Storage and usage of these datasets will be in accordance with the provider of the respective dataset.

The MMOS track involves computational models of transit and modern mobility service offerings. Researchers on this track will generate synthetic populations and travel diaries for two metropolitan areas and use a network representation of the road and transit systems. These synthetic populations and diaries will then inform the optimization and simulation programs, which will generate results dataset. In the course of testing the performance and fidelity of these simulations, the MMOS track researchers will make use of existing household travel surveys and onboard transit surveys conducted by other public agencies, as well as GPS trace data of Transportation Network Company (TNC) vehicles obtained through these public agencies.

From a broader perspective, T-SCORE will generate a number of products that may also be treated as data for purposes of this data management plan. These products include the following:

- Publications including papers, articles, reports, brochures, policy briefs, survey forms, and newsletters
- Analytical, statistical, and computational models and algorithms
- Program codes, software systems, and geo-spatial visualizations of small and big data
- Educational materials including lesson plans, course notes, lecture videos, and course projects

All of these entities will be treated as “data” by the T-SCORE team and appropriate protocolsand procedures will be followed to store and disseminate the products.

Standards and Formats

T-SCORE researchers and scholars will use appropriate industry standards and machine-readable formats for all products generated through T-SCORE activities. All data sets will be stored and archived in accordance with best practice. Data will be made available and archived in widely used formats such as ASCII (text) and CSV formats. Other data formats will be explored to ensure wide and open access across a number of platforms and software environments. At a minimum, data documentation will be provided in PDF or text format, with researchers exploring the use of XML and YAML for creating metadata. T-SCORE researchers will create metadata (i.e., data about the data) in a manner that is easily understood by end users and the wider community. The Data Documentation Initiative (DDI) Alliance (www.ddialliance.org) is an industry standard that will be consulted to ensure that metadata is prepared in a manner that facilitates documentation, discovery, and inter-operability.

The T-SCORE team believes in the principles of open access and will generate products in formats that can be readily and easily accessed. All reports, documents, articles, and papers will be made available in PDF format. In addition, where appropriate, selected documents will be made available in HTML format so that individuals can access content even without a PDF reader. Models and algorithms will likely be coded in software specific scripts, however, to the greatest extent possible, the team will open-source their code and algorithms and will host the source code in public repositories on GitHub. Documentation of these software will be provided within and alongside the repository, either in the repository README file or in the repository wiki as appropriate. The T-SCORE MMOS track researchers will develop their simulation models in BEAM, an open-source project led by staff at Lawrence Berkeley National Laboratory. By doing so, T-SCORE will facilitate open access to models and algorithms to the extent possible. Educational materials may include presentations in both PowerPoint and PDF formats, videos on a YouTube channel, and other reference documents and lesson plans in MS Word, MS Excel, PDF, and / or Canvas open course modules.

Data Access Policies

T-SCORE researchers will exercise the necessary precautions and follow standard safeguards and IRB protocols to ensure that confidential data is protected and data privacy standards are met. Every survey data collection and data analysis effort will be subjected to IRB scrutiny at the respective institutions to ensure protection of human subjects. All survey data sets will be stripped of all personally identifiable information to create public use data sets that will not compromise the safety or privacy of subjects. These public use data sets will be posted and made available in ASCII and CSV formats (with appropriate metadata) through the T-SCORE UTC website.

The full data sets with personally identifiable information will be stored on secure storage systems at each institution with appropriate access control procedures. The project will work with the compliance teams at each institution to ensure that the proper security protocols are followed. These data sets will be used by T-SCORE researchers and scholars for performing in-depth data analysis and modeling, but results will be published and disseminated so that no personally identifiable information is released. Only aggregate statistical summaries and models will be documented in reports and papers. The confidential data sets will be protected in this manner in perpetuity.

MMOS track code will be made available through the T-SCORE GitHub organization, and linked through the T-SCORE website. Release binaries of the model software will be indicated using GitHub's "Releases" features. Input files for the model software for both demonstration cities will be made available through web storage drives, and documentation on obtaining these files will be included with the releases. The input files will be distributed as compressed CSV and XML text files, as appropriate to the specific data structures represented in the files.
Sensitive data products obtained from agency partners – for example the TNC and household travel survey data used to validate the MMOS track models – will be obtained by T-SCORE researchers under non-disclosure agreements that govern the further dissemination of these data. In general, these external datasets will not be considered "T-SCORE data" and will therefore not be subject to this data management plan.

All other products of T-SCORE will generally be made available to the broader research community through the center website. Attempts will be made to ensure that all papers, articles, presentations, policy briefs, and reports will be made freely available to the community through the center website, unless academic journal polices prohibit this and open-access versions are cost prohibitive. Similarly, educational materials will be made freely available and no restrictions will be placed on access. Codes, models, and algorithms will also be posted at the center website; these resources will be made available to the research community in the spirit of open access. All codes, models, and algorithms will be open source and distributed under the Apache 2.0 (https://opensource.org/licenses/Apache-2.0) or GNU General Public License Version 3 of the Open Source Initiative (https://opensource.org/licenses/GPL-3.0), except for pre-existing code that the center uses or extends; in that case, the license on the extended project will remain.

T-SCORE will participate in ensuring that the products of the center are made available through widely used transportation libraries and repositories. The two repositories where T-SCORE will make products accessible are the National Transportation Library (NTL) (http://ntl.bts.gov) and the Transportation Research Information Services (TRIS) (http://www.trb.org/InformationServices/InformationServices.aspx and https://trid.trb.org/). T-SCORE leadership will work with UTC program managers to ensure that products of T-SCORE that can be accommodated in NTL and TRID are copied from the center website and made available through these repositories.

Reuse and Redistribution of Research Data

The T-SCORE team will make research data available to the broader community to foster research in transit strategy development, and build a community of scholars dedicated to advancing the state-of-the- art in transit simulation research. As mentioned earlier, T-SCORE will produce public use versions of data sets in which all personally identifiable information has been stripped away. These data sets will be made available to the research community through the T-SCORE website. Scholars will have to register and complete a simple online form to access the public use data sets. In this way, T-SCORE will have a tracking system to track level of usage and interest in the products of the center, and to reduce the likelihood of inappropriate use of the data. In addition, the public use data sets (together with metadata) will be made available through Dryad (https://datadryad.org/stash). All T-SCORE research

reports, papers, articles, policy briefs, newsletters, and documents will be made available via the T-SCORE center website. In the event that the T-SCORE website is sunset, another Georgia Tech digital repository will be identified so that the materials continue to be accessible to the worldwide community beyond the life of the center. Educational materials will also be made available on the T-SCORE website, and videos (where applicable) will be made available via a YouTube channel. All webinars will be recorded and links to the webinars will be posted to the T-SCORE center website.

The confidential data sets that include personally identifiable information will not be made available through open access to the worldwide community. These datasets will remain on secure storage systems housed at the center's partner institutions under the proper IRB/SSP protocols. The project will work with the relevant compliance teams at their institutions. However, it may be of value to make such data also available to the broader research community so that studies that require detailed location data can be undertaken in the future by scholars around the world. T-SCORE will work to find appropriate ways to share such data if possible in the future. Through the mechanisms described above, T-SCORE will ensure that data and products are available for reuse and redistribution, subject to data privacy laws and human subject protection. Codes, algorithms, and models will be made available to the community through GitHub, which includes a tracking system to assess level of interest and usage in the products of the center.