Data Management and Sharing Plan (DMSP)

This Data Management and Sharing Plan (DMSP) defines the strategy for managing and sharing the digital final deliverables produced by the Center for Transportation, Environment, and Community Health (CTECH), a Tier 1 University Transportation Center (UTC) funded by the U.S. Department of Transportation (Grant No. 69A3551747119). Headquartered at Cornell University, CTECH includes four consortium members: Cornell University, University of California Davis, University of South Florida and The University of Texas El Paso. The aim and purpose of this DMSP is to present details regarding our efforts to preserve and make available all final deliverables created by CTECH projects.

The document will cover the following:

1. Data description
2. Archiving and Preservation Plans
3. Standards used
4. Access Policies
5. Re-Use, Redistribution and Derivative Products Policies

1. Data description

For the purposes of the DMSP, CTECH will use the following definitions:

- Final Project Datasets: Recorded factual material commonly accepted in the scientific community as necessary to validate research findings at the individual project level, i.e., the data needed to reproduce the final results.
- Project Metadata: The set of data that describes and gives information about and context for the dataset as recommended by the Data Documentation Initiative (DDI).
- Dataset description document: Describes all the variables in a dataset equivalently to a database schema, also known as data dictionary. For example, the measurement units used, variable labels, label values etc. This document should specify the data position of each variable, describe the contents of each variable, and identify the range of possible codes and the meanings of the codes.
- Code: Any routines, scripts, queries, or other code needed or desired to reproduce final results. The code should contain comments or other documentation that guides its execution by others.
- Final Documentation: Final reports, papers, handbooks, guides, manuals, or presentations derived from research and produced as a final deliverable.
- Final Deliverables: Final documentation, final datasets with their associated metadata, data description documents and, if applicable, code.

CTECH will only deposit final deliverables in the CTECH document archive. Individual CTECH research projects may create and use a wide variety of information in digital form including:

experimental, observational, and simulation data; codes, software and algorithms; text; numeric information; images; video; audio; and associated metadata. CTECH will not archive these raw datasets or interim products.

2. Archiving and preservation plans

Final deliverables as defined above will be archived in the CTECH Collection hosted by the Cornell University Library's institutional repository, eCommons (https://ecommons.cornell.edu/), for preservation and access. The library is committed to long-term preservation of the binary form of the digital object. Datasets will be made openly available online upon approval/permission from relevant sources. Data files in eCommons are written to state-of-the-art, high-availability servers in the Cornell Data Center. Servers have full emergency power and are backed up daily to a secondary location on campus. A replica copy of the backup data is kept at Weill Cornell Medical Center's IT facility in New York City for disaster recovery purposes. eCommons provides each item with a persistent URL (handle and/or DOI) to facilitate citation, allows assignment of Creative Commons and other usage.

In addition, these final deliverables, including datasets and accompanying material, will also be made publicly available through links on the CTECH website http://ctech.cee.cornell.edu/data-management-and-sharing-plan/ and be deposited in the National Transportation Library. The CTECH eCommons Collection will store CTECH files in perpetuity.

3. Standards used

Datasets will be archived in platform-independent, non-proprietary digital formats. Data in proprietary formats such as Microsoft Word, Excel and PowerPoint may be stored temporarily, but CTECH will archive final documents as portable document format (.pdf) files, and final data as comma-separated values (.csv), JavaScript Object Notation (JSON) and/or text (.txt) files. Each PI will submit final deliverables to CTECH in both native format (Microsoft Word, Excel or PowerPoint commonly) and in .pdf, .csv, JSON and/or .txt. These application-independent file types are less likely to lose their backward compatibility with new software upgrades. In the future, if other file types (e.g., audio, video, image, geospatial) need to be addressed, the optimal preservation format as defined by the Library of Congress will be accepted. For example, geospatial data currently require .shp, .shx, and .dbf files for preservation and future access, all of which would be archived. Whenever possible, each individual project will attempt to generate code using freely accessible programming languages such as Python and R, and only use publicly accessible and achievable libraries.

When submitting the final datasets to the CTECH repository, the metadata, data descriptions and code for each project will be stored with their associated final data files. Archived data files will be described with the DDI compliant metadata fields that have been widely adopted by the international data archives community. The fields populated during the creation of descriptive metadata – filename, data type, author, abstract, keywords, publisher, geographic coverage, temporal period of collection, response rate, funding, rights, etc. – will enhance search and discovery of project data.

4. Access policies

To facilitate and encourage data sharing, public use versions of the final datasets (de-identified and filtered as necessary to preserve privacy and/or intellectual property) will be made accessible online via the CTECH website and eCommons. Along with these data, CTECH will also preserve and provide access to supporting materials necessary to interpret and use the data appropriately. The CTECH eCommons collection will automatically assign a unique and persistent digital object identifier (DOI) for each dataset to ensure that the data will be accessible online even if their location should change. The DOI will also link the material to the CTECH website and the National Transportation Library. The eCommons system also generates standardized data citations to encourage attribution to the project team.

CTECH is currently not aware of any reasons that might prohibit the sharing of public-use versions of final datasets. In certain cases, as determined by the lead PI on each project, an embargo period of one year may be placed on the data to allow the project team to retain first use rights of the data. It is important that investigators have sufficient time to analyze the data to produce and publish peer-reviewed manuscripts of their research (without time pressure) before the data are available to external users. After the one-year embargo period, the data will be made publicly available via the CTECH website and eCommons. If survey data are included in the final datasets, the publicly available responses will be voluntary, anonymous, confidential, and unidentifiable. Any other data with personally identifiable information (PII) will be anonymized prior to deposit. Results will be released only as aggregate statistics. PIs will be responsible for removing all identifiers from datasets before release. In the event that a final dataset includes embargoed data, restricted access will be imposed until open access is granted by the PI or the one year embargo ends, or an exception to this timeframe is agreed to by CTECH, whichever comes first.

As mentioned earlier, the storage and backup of project data and codes during the active phase of the projects will be coordinated at an individual project level by each PI. During this phase, CTECH recommends that raw and intermediate data files and codes be stored and backed up on an access-restricted, password-protected server as determined by the PI and their affiliated

university requirements. It is recommended that code be managed via a version control system. Each PI will be responsible for ensuring that their chosen computing support follows optimal policies and procedures for safeguarding sensitive data. This includes provisions for secure data storage, transmission, access restrictions, and incident management. Only the PI and authorized project personnel will be allowed access to project files. In addition, the PI is responsible for addressing their respective Institutional Review Board (IRB) policies and ensuring compliance with any conditions placed on approval.

5. Re-Use, Redistribution and Derivative Products Policies

The materials generated through CTECH projects are intended to be reusable for validation and future research. All final datasets in .csv or .txt format will carry a Creative Commons, CC – "Public Domain Dedication." The user may share by copying and redistributing the material in any medium or format and may adapt by remixing, transforming, and building upon the material. CTECH assumes that consumers of this data will conform to good scientific practices and that proper credit will be given via citation. Final code generated by CTECH projects will similarly be made available through a GNU General Public License (GPL).

All content archived in .pdf format will be readable but will be secured against copying and editing.