Data Management and Sharing Plan

The PI and co-PIs of this proposal are committed to maintaining and disseminating data developed during the conduct of the proposed research in consistent with the requirements as articulated in the requirements.

Overview

This project will produce a variety of data including publicly and privately available data, statistical data sets, qualitative data sets, travel demand modeling modifications and related program codes and data sets, transportation network simulation data from case studies, and computer codes. Data will also be obtained from a variety of sources (including outputs of the project) and will be stored during the funded project period in the PI's laboratory.

The strategic objectives of this Data Management Plan are to: (i) store, archive, and curate all data; (ii) utilize resources from Consortium Universities such as hardware, software, and staff expertise; (iii) collaborate with international researchers and practitioners to ensure proper representation of data and data formats; and (iv) share curated data in a timely manner. We will ensure that any qualified user can 1) reproduce the analyses, and 2) use the data for other relevant studies, if appropriate.

The data types will include i) research progress, ii) synthesized data as well as simulation data, and iii) methods and programming codes for data analyses. Backups will be set automatically using the Co-PIs' backup system such as Apple's Time Capsule.

General Data Management Plan.

The data generated in this project consist of the following types: 1. Research data. Such raw data will be stored on password-protected computers of the PI, Co-PIs and key research staff; be shared through secure cloud-based "Box" system, which encrypts data both in transmission and while stored in the cloud, provides detailed auditing of who has acted on and viewed files, watermarks of files, and allows view-only files that can prevent others from downloading. Where possible, personally identifying information will be kept in separate files from response data to ensure anonymity of human participants. 2. Research results data. Such anonymized or aggregate data will be published through conferences, journals, and technical reports; 3. Source code and Software. Such data will be maintained in an SVN server for version control and is only accessible by project participants of three collaborative universities. It may be released as open source software to stimulate the research interest of the research community upon project completes; 4. Testbed and simulator experimental data. Such data are collected for system diagnosis and evaluation purpose during project experiments and software simulations. These data are preserved and may be released through our project website.

Data files formats

In general, all non-proprietary data and codes will be stored using unencrypted and uncompressed ASCII or Unicode files. Travel demand model modifications will be made in the

TransCAD environment. Many of the input and output files of the model are in the database format (*.dbd), the matrix format (*.mtx), or the binary format (*.bin). For some files, such as curriculum material, PDF files will also be stored and backed up. Formats and standards of some other files are TBA (to be developed within the project).

Computer codes

Statistical analysis programs will be written in R or other languages, which is a high-level programming language and software environment for statistical computing, data analysis and graphics. R is part of the GNU project and its source code is freely available under the terms of the GNU General Public License. Copies of the code will be shared through eCommons.

Data Availability and Sharing

The publicly available data described above will be made available after the research team has had the opportunity to produce publications and conference presentations to meet its funding obligations. Distribution and availability of the data will be subject to restrictions by the original source such as those regarding copyright.

During the project, data will be shared only among collaborators. At the conclusion of the project, however, the expectation is that these data will be of general interest to the research community and the public. After an embargo period of 1 year to allow for primary scholarly dissemination, the project data sets will be submitted or made available through an alternative appropriate repository.

Non-proprietary data, codes, modeling results, curriculum material and outreach activities will be shared through eCommons, the official data repository of Cornell University. Estimates and other parameters of the statistical analysis, database and matrix data from the network analysis, and materials from the social welfare analysis will be shared through tables in scientific articles. All materials are indexed in search engines, including Google Scholar. eCommons at Cornell ensures long-term digital archiving and preservation, acting as an additional source for long-term storage of the data.