Plan Overview

A Data Management Plan created using DMPTool

Title: Tomasic_Universal Transit Assistance

Creator: Anthony Tomasic

Affiliation: Carnegie Mellon University (CMU)

Principal Investigator: Anthony Tomasic

Data Manager: Anthony Tomasic

Funder: United States Department of Transportation (DOT) (transportation.gov)

Funding opportunity number: 12/5/2017 - 12/4/2018

Template: U.S. Department of Transportation Public Access Guidance v1

Last modified: 12-05-2017

Copyright information:

Tomasic_Universal Transit Assistance

Data description

Describe the data that will be gathered in the course of the research project, including whether the data should be preserved for long-term access.

Data Description

The purpose of this research is to build computational models that reliably predict the origin/destination of user trips for a transit service. This prediction takes the form of a machine learning algorithm designed to incorporate relevant information to predict origin/destination.

Policies for Access and Sharing

We will apply a request to the IRB board to make the data publically available after a period of time. If approved, the dataset will be added to the Center's data repository.

Data format and metadata standards

Describe the standards and machine-readable formats that will be used in the course of the research project.

The data is available as SQL database dumps in ASCII text, a common, open, non-proprietary data format. The process to generate these dumps is simple - a standard database dump tool is used.

In order to understand the data, we will include a data dictionary that describes the data as text comments on the SQL schema. The data is managed and stored as a single large tar.gz file. To read or view the data, a researcher unzips and pipes the data to an SQL database system. The file is self contained with respect to schema and data.

The data is stored in a relational database and consists of approximately a year of Tiramisu user transit ride traces (several thousands of records). The data was gathered through the use of the Tiramisu mobile transit information system application and the

location services available on mobile devices. Potential users of the data include (i) researchers interested in constructing models of transit rider behavior and (ii) transit system planners interested in constructing models of transit rider behavior. Thus the potential value is only in constructing and validation of models. The data has no value outside of these constraints. The constructed models have value to traffic planners.

Policies for access and sharing

Discuss the access policies that will apply to the data, so as to protect against the disclosure of identities, confidential business information, national security information, etc. and whether public use files may be generated from the data.

With respect to public access, we will submit an IRB requesting this change to the data, since it was originally collected without permission to generally distribute. We may have to anonymize the data (although at the moment the data is relatively anonymous). The PI will be responsible for managing the data. He will confer with colleagues on a regular basis to insure adherence to this data management plan.

Policies for re-use, redistribution, derivatives

Discuss the policies for re-use, re-distribution and derivative projects.

Carnegie Mellon University or the home institution of the faculty holds the IP for data created by the project.

Currently the data rights are retained by CMU and there are no rights to re-use, redistribution or derivative works.

If the data is approved for distribution by CMU IRB and CMU IP Office, the data will be distributed into the public domain.

Plans for archiving and preservation

Outline the plans for archiving and preservation, specifying where research data will be

deposited, and specify that data will be deposited at the time of initial publication of any related peer-reviewed journal article.

1. The Mobility21 UTC will archive all data on CERN, https://cds.cern.ch/, which is an approved site of the USDOT.

2. When a project submits a final report, the faculty will have 60 days to archive their data on CERN.

3. Faculty will maintain the data until it is uploaded to CERN.

CERN is a approved data repository by USDOT, we assume the following is pre-approved by DOT

4. Describe how back-up, disaster recovery, off-site data storage, and other redundant storage strategies will be used to ensure the data's security and integrity.

5. Describe how data will be protected from accidental or malicious modification or deletion prior to receipt by the archive.

6. Discuss your chosen data archive's policies and practices for back-up, disaster recovery, off-site data storage, and other redundant storage strategies to ensure the data's security and integrity for the long-term.

7. Indicate how long the chosen archive will retain the data.

8. Indicate if the chosen archive employs, or allows for the recording of, persistent identifiers linked to the data.

9. Discuss how your chosen data repository meets the criteria outlined on the **Guidelines for Evaluating Repositories for Conformance with the DOT Public Access Plan page.**