C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION

Connected Cities for Smart Mobility towards Accessible Resilient Transportation (C2SMART)

Master Data Management Plan

Version 2.0

(Revised April 23, 2018)

Lead Institution

New York University

Partner Institutions

City College of New York
Rutgers University
University of Texas, El Paso
University of Washington, Seattle

C2SMART Data Management Plan

C2SMART is a Tier-1 University Transportation Center composed of a consortium of universities: New York University (NYU) (lead), City College of New York (MSI), Rutgers University, University of Texas El Paso (MSI), and University of Washington Seattle. The main research priority of the C2SMART is "Improving Mobility of People and Goods" with a focus on the topic area of "Smart Cities." C2SMART's mission is to build a solution-oriented research center that uses resources from a range of cities among its consortium members as a decentralized but comprehensive living laboratory. The Center studies a number of challenging transportation problems and field tests novel solutions in close collaboration with end-users, city agencies, policy makers, private companies, and entrepreneurs.

In coordinating research, education, and workforce training, the Center will ensure that data of all types will be managed and organized for security, consistency, and public dissemination when appropriate. This document details the general requirements for all activities funded by the Center, whether it's conducting research, educating students, training professionals, or providing outreach to connect industry, government and academia.

C2SMART researchers will follow the guidelines and policies in this Master Data Management Plan (DMP). For each individual research project, researchers will also create a narrative Project DMP, based on the USDOT guidelines[1]. Researchers may reference or quote from the C2SMART Master DMP as appropriate but should be sure to call out in their project DMP the unique qualities of their research, and where their DMP differs from this Master DMP. Researchers wanting more guidance on the U.S. DOT Public Access Plan can find guidance at https://ntl.bts.gov/public-access

[1] https://ntl.bts.gov/public-access/creating-data-management-plans-extramural-research

1. Expected Data

The Center expects a variety of data to be gathered or generated depending on the types of funded activities. This data falls into a number of different areas including infrastructure, users, and the environment, and can involve public as well as private companies. These are divided into different activities:

Research projects – these are funded projects led by one of the PIs within the Center

- Publications: working papers, project reports, and open access articles
- Code and test instances prepared for interoperability between different city settings
- Data generated from experiments or surveys
- Private data shared by agency or private company

Educational activities – these are long and short courses, projects, and internships designed to disseminate knowledge from the Center's research activities to students attending the consortium institutions.

- Course notes and syllabi with consideration of involvement from different consortium institutions
- Summary of attendees and outcomes
- Project/internship deliverables
- Video/webcasts of presentations, lectures

Outreach activities – these are workshops, symposiums, hackathons, invited speaker seminars, and student-run conferences held between consortium institutions.

- Video/webcasts of presentations/seminars/invited talks
- Conference proceedings
- Hackathon codes and experimental data
- Contact/mailing list for all involved active faculty, students, staff

All data created from the Center's research will be stored and maintained to ensure long-term accessibility. All data that will be produced needs to be identified by project PIs, both at the start and end of a project. Data-specific restrictions for release, if any, should be clearly documented and submitted in the project's Scope of Work (SOW).

2. Data Formats and Standards

The data will primarily be stored either as code files, images/videos, txt format for outputs, and pdfs for documentation. It is expected a variety of file formats will be generated from Center research, which we will seek to migrate to a single stable format in accordance with Library of Congress standards[2]. Except for private data shared by agencies and private companies, and certain videos/webcasts, all data produced from the Center's research will be made publicly accessible.

Study data will be hosted on a central server at NYU Tandon School of Engineering for the duration of the study, while a permanent version will be added to the Zenodo data repository following completion of the study (see "5. Archiving of Data"). A machine-readable .json metadata file should be produced for each primary source data. PIs should refer to the Project Open Data Metadata Schema, chosen by the U.S. DOT as the preferred schema in their Public Access Plan, to develop the metadata.

A working paper series will be used for hosting unpublished working papers. PIs can publish them in other outlets. A project report is expected for each funded project. The report should clearly highlight the data in the repository generated from the project. Reports may be withheld from sharing with the public until after publication of content.

Code can be of any language. Test instances should be able to run specific input data to get expected output data. Test instances should allow researchers at a different consortium institution or partner to apply at their scale.

[2] https://www.loc.gov/librarians/standards

3. Access to Data and Data Sharing Practices and Policies

Data from the research projects funded by the Center will generally be made publicly accessible. Exceptions to this policy are data that contain personally identifiable information, confidential business information, or classified information. In these cases, notes are needed to explain why the entire or part of the data cannot be made publicly accessible. Typically, the level of sharing depends on the nature of the data.

For example, software tools that implement the model/algorithms of research should be shared after the intellectual property issue is properly addressed. Simulation data and other types of model-generated data can be shared without any restrictions. Infrastructure and control data can be shared upon obtaining approval from transportation management agencies who manage the infrastructure (such as city or state Departments of Transportation). Road user and vehicle data can be shared after removing personal identifiable information. Environment data can be shared upon the approval of the data owner. Other types of data may also be shared at appropriate levels depending on the way of data collection, content of the data, and their actual formats.

At the same time, protecting research participants and guarding against the disclosure of identities and/or confidential business information is an essential norm in scientific research. When working with human subjects, researchers will follow Institutional Review Board (IRB) policies of their affiliate institutions and should seek and identify IRB approval before beginning the study or collecting any data. If needed, proper documents will be prepared to address these issues and outline the efforts that will be taken to provide informed consent statements to participants, the steps that will be taken to protect privacy and confidentiality prior to archiving the data, and any additional concerns (e.g., embargo periods for the data).

In case it is impossible to anonymize the data in a manner that protects privacy and confidentiality while maintaining the utility of the dataset, the necessary restrictions on access and use should be clearly stated. In matters of human subject research, the informed consent forms should describe how the collected data will be shared with the research community and whether additional steps, such as an Institutional Review Board (IRB), may be used to protect privacy and confidentiality.

4. Policies for Re-Use, Re-Distribution

Data will include a "source" tag, typically associated with the project report or its own unique digital object identifier (DOI).

The Creative Commons Attribution 4.0 International (CC BY 4.0) license[3] will be utilized for all re-use and re-distribution of data, in accordance with federal guidelines[4]. This license allows for users to copy and redistribute data, modify and build upon the material even for commercial purposes, as long as attribution is provided to the authors/creators of the data.

Production of the derivatives based on software packages (i.e., new development based on the source codes) will be handled on a case-by-case when a written request to the Center is received in order to protect the intellectual property (IP) of the project team. In such cases, the GNU General Public License v3.0[5] will be utilized for all re-use and re-distribution.

Open-source codes and developmental efforts will also be detailed on and shared via Github as appropriate to the nature of each project or study.

[3] https://creativecommons.org/licenses/by/4.0/
[4] USDOT "Plan to Increase Public Access to the Results of Federally-Funded Scientific Research Results Version 1.1" Published December 16, 2015
https://www.transportation.gov/sites/dot.gov/files/docs/Official%20DOT%20Public%20Access%20Plan%20ver%201.1.pdf
[5] https://www.gnu.org/licenses/gpl-3.0.en.html

5. Archiving of Data

All of the publicly accessible data will be physically stored on a central data repository hosted in the Center headquarters at NYU Tandon School of Engineering, with a concurrent online version located on the Zenodo repository, which is conformant with the U.S. DOT Public Access Plan, located at https://zenodo.org/. In cases where code is being stored in the repository, the Zenodo repository will also be linked to the project or study's Github via Zenodo's DOI-enhanced URI.

The Project Open Data Metadata Schema will be used to locate the data. The repository will allow users to upload, archive, and manage project data. The project data needs to be uploaded to the repository for a project to be considered completed.

Data that is to be made publicly accessible shall follow the U.S. DOT Public Access Plan, as noted in: https://www.transportation.gov/mission/open/official-dot-public-access-plan-v11

The repository shall be indexed in the following site:
http://www.re3data.org/search?query=transportation

It shall also be added to the National Transportation Library (https://ntl.bts.gov/). A data package, including the final report, public datasets, the project DMP, the machine-readable metadata files, and other documentation, should be sent to NTL at ntldatacurator@dot.gov. If data files are large, an email requesting a secure large file transfer interaction should be sent first.

If the Center is ended, the public data in the repository will be uploaded to an existing publicly available repository that conforms with the US DOT Public Access Plan:
https://ntl.bts.gov/publicaccess/repositories.html.