Connected Communities for Smart Mobility towards Accessible and Resilient Transportation for Equitably Reducing Congestion (C2SMARTER Center)

Grant Period: June 1, 2023 – May 31, 2029
U.S. DOT Grant Number: 69A3552348326

Center Data Management Plan

Version 1.1

Effective Date: October 5, 2023

Lead Institution

New York University

Partner Institutions

New York City College of Technology
North Carolina A&T University
Rutgers University
Texas Southern University
University of Texas-El Paso
University of Washington

For any questions or further information please contact:

Shri Iyer, Managing Director
shri.iyer@nyu.edu

Kaan Ozbay, Center Director
kaan.ozbay@nyu.edu

C2SMARTER Data Management Plan

C2SMARTER is a Tier-1 University Transportation Center composed of a consortium of universities: New York University (NYU) (lead), New York City College of Technology, North Carolina A&T University, Rutgers University, Texas Southern University, University of Texas-El Paso, and University of Washington. The main research priority of the C2SMARTER is *Reducing Congestion* with a 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.

The purpose of this Data Management Plan (DMP) is to assist Center-funded faculty, staff, students, and partners in producing research outputs in accordance with sponsor (U.S. Department of Transportation) and University Transportation Center (UTC) program requirements. In coordinating research, education, and workforce training, the Center aims to ensure that data of all types will be managed and organized for security, consistency, and public dissemination as required. 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.

C2SMARTER researchers will follow the guidelines and policies in this Center Data Management Plan. For each individual research project or effort, researchers will also create a narrative Project DMP, based on the U.S. DOT guidelines[1]. Researchers may reference or quote from the C2SMARTER Center DMP as appropriate but should be sure to tailor their project DMPs to the unique qualities of their research, and call out where their DMP differs from the guidelines set forth in this Center DMP. Individual project DMPs are meant to serve as living knowledge management tools, and should be reviewed and updated regularly throughout the project's life cycle: at minimum, each time there is a significant change in the research project, the data collected, or in project personnel. 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. Data Description

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
- Data generated from experiments or surveys, or output of Center-developed simulations, models, or other tools
- Secured, sensitive, or confidential data shared by agency or private company
- Code and test instances of software and tools prepared for interoperability between different city settings

Educational activities – these are long and short courses, projects, and internship programs 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, including surveys
- 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
- Summary of attendees and outcomes, including surveys
- Hackathon codes and experimental data
- Contact/mailing list for all involved active faculty, students, staff

In developing Project DMPs, the Data Description should describe what the type of output data is, and address the scope and scale of the data. Where applicable, characteristics of the data in as much detail as available should be included, including relationship to other datasets. If the data contains any risks such as a security concern or includes personally identifiable information (PII), it should be specified.

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

To sufficiently answer these questions, the following should be addressed as applicable:

1. Name of the data/dataset, project, or data producing program.
2. The purpose of the data in relation to the research or project activity.
3. A description of the data that will be generated in terms of nature and scale (e.g., numerical data, image data, text sequences, video, audio, database, modeling data, source code, etc.).
4. Methods for creating the data (e.g., simulated; observed; experimental; software; physical collections; sensors; satellite; enforcement activities; researcher-generated databases, tables, and/or spreadsheets; instrument generated digital data output such as images and video; etc).
5. The period of time data will be collected and frequency of update.
6. The relationship between the data and other existing data.
7. The potential value of the data over the long-term for the project and the general public.
8. Rationale for any restrictions on the data, such as lack of public access.

2. Data Formats and Standards

Format of data storage will depend on the nature of data. 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]. Numerical data shall be stored as csv or machine readable text format for outputs, whereas code files will primarily be stored as scripts.

Media may take the form of images/videos, txt format for other outputs, and pdfs for documentation. These file formats are generally platform-independent and non-proprietary. When possible, open-source and replicable formats shall always be selected as the standard. It is understood that certain research requires the use of proprietary softwares with formats that may not be interoperable. These shall be handled on a case to case basis.

Code can be of any language. Test instances should be able to run specific input data to get expected output data, and should allow researchers at a different consortium institution or partner to apply at their scale. Codebooks or README should also be included to facilitate this replication.

It is important that metadata are included in a standardized format. Where applicable, 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.

In developing project DMPs, teams shall seek to answer the following:

As general guidance you may consider addressing the following:

1. List the format(s) of the data indicating if they are open or proprietary.
2. If proprietary, provide rationale for using those standards and formats.
3. Describe how versions of data be signified and/or controlled.
4. Describe metadata schema and how metadata be managed and stored.
5. Indicate what tools or software is required to read or view the data – if not a standard format, describe the rationale and how to work with the data

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

6. List what documentation exists or will be created in order to make the data understandable by other researchers.
7. Describe quality control measures.

3. Access to Data and Data Protection

Data from the research projects funded by the Center will generally be made publicly accessible after the conclusion of the research study. During the course of the study, Center-wide data shall be stored on a central server at the NYU Tandon School of Engineering. This server employs authentication methods tied to users' university-provided login credentials to manage access. Only those that are listed as principal investigator, co-principal investigator, or project personnel, shall have access to the data over the course of the study, except in the case of where a non-disclosure agreement or other confidentiality clause governs further restriction to data access. For individual projects, data storage during the course of the study will be handled on a case by case basis based on the requirements of the project and principal investigator.

In general, data from research projects funded wholly or in part by U.S. DOT must 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. Software tools that implement the model/algorithms of research should be shared after any intellectual property issues are properly addressed in consultation with researchers' universities' legal offices. 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) to prevent any security issue. 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. Further, when working with, or conducting research that includes Indigenous populations or Tribal communities, C2SMARTER researchers will adhere to the CARE Principles for Indigenous Data Governance https://www.gida-global.org/care.

In developing project DMPs, the following questions should be answered:

1. Describe how and where data will be stored and how others will access them.
2. State who will have access to the data including any permission levels or division of responsibilities.
3. Indicate whether the data contain private or confidential information and how disclosure of identities and/or confidential business information will be prevented.
4. Describe how any intellectual property issues have been or will be handled to not impede sponsor required public access.

4. Policies for Re-Use, Re-Distribution

Following the conclusion of research or project activity, all collected data will be added to the Center's Zenodo data repository. 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 in coordination with the intellectual property (IP) requirements of the researchers' home institutions. By default, 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.

The U.S. DOT also reserves a royalty-free, nonexclusive and irrevocable license to reproduce, publish, or otherwise use and to authorize others to use the work for government purposes.

In developing DMPs for projects, discuss if any modifications to this protocol will be employed. Specifically, if there are intellectual property issues that affect re-use or re-distribution, and how they will be handled. In general, U.S. DOT's rights are not modifiable. Any questions shall be directed to the National Transportation Library Data Services (public.access@dot.gov).

[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 shall be archived and stored on the Zenodo repository which is conformant with the U.S. DOT Public Access Plan, located at https://zenodo.org/. Datasets will be tagged to the Center's community. 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's Repository & Open Science Access Portal (ROSA-P). 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 U.S. DOT Public Access Plan: https://ntl.bts.gov/publicaccess/repositories.html.

Changelog

2023-10-05: Version 1.1 – updates:

- Standardized Center & Master DMP nomenclature to Center
- Clarify references and differences between National Transportation Library, UTC program and repositories
- Insertion of Codebooks/Readme in Section 2 on producing code
- Insertion of language in Section 3 on work with indigenous or tribal communities

2023-09-30: Version 1.0
2023-07-31: Version 0.9