DATA MANAGEMENT PLAN

UTC Title: Center for Connected Multimodal Mobility (C^2M^2)
UTC Director: Dr. Ronnie Chowdhury (Clemson University)
UTC Associate Directors: Dr. Nathan Hyunh (University of South Carolina), Dr. Gurcan Comert (Benedict College), Dr. Judith Mwakalonge (South Carolina State University), and Dr. Dimitra Michalaka (The Citadel)

I. Types of Data

Data produced as a result of the proposed project is expected to include, but not limited to, web sites, publications, research posters, student application codes, and archival-quality data that is acquired and curated from a variety of text and other sources. Text documents including technical papers, theses, posters, and software documentation are expected to include figures in the form of tables and graphs; codes are expected to include source, test routines and data, building and installation instructions as well as ancillary scripts as needed.

In the proposed plan, data sets generated from experiments or obtained from existing sources may fall in the following categories:

1. ***Connected Vehicle / Sensor Data***: This category includes data generated from sensors and devices embedded in connect vehicles. Connected vehicles may include, but are not limited to, passenger vehicles, transit vehicles and heavy duty freight trucks.
2. ***Transportation Infrastructure and Traffic Data***: The team will conduct real world as well as simulation experiments for various roadway corridors in South Carolina. This will require collecting roadway geometric, traffic and traffic control related data.
3. ***Human Factor related Data***: Human performance/behavior data collected from human subject studies may be used. Examples include human factor related data collected from field experiments and driving simulator studies.
4. ***Personal Device (Social Media and Cell Phone) Data:*** Data from social media applications, such as twitter, may be used by the center. Additionally, data from personal devices from users such as pedestrians may be used.
5. ***News and Weather Data:*** News and weather data may be used by the center.

II. Data Formats and Standards

Since various tools and techniques will be involved in the process of generating and collecting large amounts of data, metadata (e.g. data headers from simulation and real world testing data) will be made available in their original format for reusability and originality. Detailed data collection methods and procedures will be documented corresponding to each metadata set available in at least one of the following formats: plain text, PDF, HTML, MS Word, or LaTeX. The aggregated data will be in plain text format and may be compressed for space-saving purposes. Data dictionaries will be provided for all collected data sets, and standards used for data and metadata format and content (absent existing standards or standards deemed inadequate) will be documented with proposed solutions and remedies.

III. Roles and Responsibilities

The lead UTC partner from each university will be responsible for making their data publically available for access. The Center for Connected Multimodal Mobility will also have a UTC Coordinator, who will oversee the implementation of the data management plan.

IV. Policies for Access and Sharing and Provisions for Appropriate Protection/Privacy

All actions of data access and sharing in this research will conform to Institutional Review Board (IRB) requirements at each participating institution. All institutions in this research will obtain IRB approval before utilizing any human subject data. All private identifiable information regarding human subjects will be strictly pulled from the data for publication and data sharing purposes to protect personal privacy; the participant will be de-identified such that the data cannot be traced to the participant. All use of the data will conform to the purposes documented in the original informed consent.

All data necessary for reproduction of research activities will be made available to the community. If the data sizes are small, they can be hosted directly via the UTC's website and other public resources such as the USDOT Research Data Exchange (RDE). If the data are large, samples of the data will be hosted and the full dataset will be provided upon request. The availability of raw data will be determined on a case-by-case basis. All raw data generated from public sources (e.g., public traffic infrastructure, weather stations, news) and federally funded resources (e.g., experimental testbeds) will be made available. Raw data provided to UTC via research agreements with private entities (e.g., telematics data from LexisNexis) will not be made available. However, data collection and data query procedures will be provided so that other institutions having access to these resources can reproduce the research results.

Access to, sharing and distributing of code, data, and documentation will be via web sites and publishing in conferences, journals, and monographs. The Lead Center will maintain a web server with support from Clemson Computing and Information Technology (CCIT). CCIT maintains a web site linking all Clemson University programs at http://www.clemson.edu/ces/computing/.

The Lead at each university intends to publish and present the results of the research at relevant workshops, conferences, etc. Students will be expected to participate in research and present results. Theses and dissertations that result from this project will be maintained and made available by the library at each participating university.

Clemson University will host the primary site for storing and disseminating results of the proposed project. The original author of each document or code, and his or her corresponding university will retain ownership of the intellectual property.

V. Policies and Provisions for Re-Use, Re-Distribution

Data from this research will be significant for use in research involving field experiments, driving behavior modeling, fundamental studies in car-following theory, and microscopic and macroscopic traffic simulations. Scholars are welcome to use the available data for their independent research in relevant areas with citation to originator of the data sets.

VI. Plans for Archiving and Preservation of Access and Data Integrity (Long-term digital data

The generated data, project reports and other research products (e.g., published papers, software, original proposals, quarterly project reports, and manuals) will be stored on data servers located on the Clemson University campus.

The university has a professional Information Technology (IT) department that can help with data backup on a regular basis. This will ensure the long-term safety and integrity of the data. The information on Clemson University data servers will be shared through the USDOT Research Data Exchange (RDE). Metadata such as data dictionaries will be stored on the two data servers and RDE. Interested researchers, subject to approval, can download the data. Since all the data will be preprocessed and stored in standard format, there is no need for transformation in order to make the data reusable.

Clemson University has two data centers available to the university's research and education community consisting of 2 petabytes of redundant tape library storage used for backup and archive purposes. In addition, tens of terabytes of spinning disk storage are also available for hot standby and tape staging areas at a remote data center. Clemson's primary data center is securely located in the Information Technology Center (ITC) at Clemson's Innovation Campus and Technology Park. The data center is the home of both enterprise and HPC systems, and is staffed by CCIT staff on a 24-hour basis from a state of the art network operations center within the ITC. Clemson's secondary data center, located on the main Clemson campus, provides for business continuity.

Clemson's data storage systems are monitored in Clemson's Network Operations Center (NOC). The NOC utilizes cutting edge technology, equipment, and monitoring tools to proactively monitor the network infrastructure and critical computing systems and services provided by the university community, overseeing functionality, capacity and performance. The NOC operates 24-7 to provide top-level support and customer service to the university's research and academic communities and our affiliated partners.

The data centers are physically protected by 24-7 surveillance cameras that are actively monitored and recorded. Monitoring of security anomalies is done via but not limited to Security Incident and Event Management based solutions with auto-reporting to the NOC and Clemson's security team should an event occur. Access to the facility is protected by card access with a role-based scheme as to authorization. Additional layers of access are afforded to the machine room in which personnel have to demonstrate need, and participate in proper training that is renewed annually on applicable state and federal regulations (i.e., HIPAA, FERPA, FPPA, etc).