Data Management Plan

1 Data and Artifacts Generated

We have identified the following types of artifacts and data that will result from this project:

1. Physical samples and prototypes: Over the course of the project several physical samples and proto-types of taxi sensing modules may be produced.

2. Software/algorithms: Extensive software development forms the core of this research effort.

3. Numerical simulations and calculations: Over the course of the project numerical simulations and calculations will be performed using MATLAB and other software available at Carnegie Mellon University and digital data will be generated mainly in the form of tables and images.

4. Observational measurements: Our deployments will result in extensive data gathering, including vehicle location, trajectories, route selection, ride requests, air quality from our deployment.

5. Measurements and monitoring data: In addition to the data collected in support of the deployment applications themselves, extensive data will be collected regarding how the software systems (FSM) operate.

6. Metadata: The prototypes and the measurement and monitoring data will be accompanied with appropriate metadata (i.e., reports and publications) that will contain detailed descriptions, design and manufacturing procedures, test-setup explanations, and lists of the equipment used. These will help in the re-creation of the samples and prototypes, as well as the understandability and reproducibility of the tests.

2 Data and Metadata Standards

The prototype systems will be made as described in the project description. These prototypes will include the instrumented taxis. The design is open sourced as well as any additional physical artifacts generated as part of the project will also be open sourced.

Unprocessed data resulting from numerical simulations and calculations will have its original format as provided by software used. Unprocessed measurement and monitoring data will be stored in text files (extension .txt) containing basic data to identify the measurement (i.e., date, time, parameterization, etc.), as well as additional columns specific to the measurement.

Processed and validated data obtained from measurements will be stored in Microsoft Excel files (extension .xlsx). Processed data for measurements will show the values of the measurands (e.g. position and temperature) in standard engineering units. Diagrams may be included with the files as necessary. Plots will be stored as images or documents as appropriate (extensions .jpg, .pdf, .eps). Any research notes taken during experiments and deployments will be stored in Acrobat Reader format (extension .pdf). The meta-data will be delivered in Acrobat Reader format (.pdf) and Microsoft Powerpoint (extension .pptx). Digital results data is expected not to exceed 2 GB. Data from sensor deployments could be larger, but less than 1TB.

3 Policies for Access and Sharing

3.1 Raw Data:

Raw data collected as part of this project falls into two categories. 1) generalized observation ground truth data that can be collected through human observation and publicly obtainable. Such data will be shared similar to anonymized data. However, it is the nature of this research to explore more fine-grained infor-mation that can potentially track and reasonably identify the drivers. These data will be treated as human

subject data and be stored in accordance to CMU human subject research policy. The data stored will be encrypted and only accessible by members of the research group until they are properly anonymized. Furthermore, we will comply with various regulations for proper storage and disposal of identifiable information for any human subjects associated with our research.

3.2 Anonymized Data:

Processed, anonymized and validated data will be made publicly available on the PIs CMU websites. The websites are backed-up daily. The data will be freely accessible and download statistics will be maintained. Processed and validated data sets, and accompanying metadata, will be made available to the public no later than 12 months after the initial conference and/or journal publication. The data will be available through the websites for an initial period of five years. Several five-year extensions may be granted depending on the interest of the users. If no extension is granted, the data and the metadata will be removed from the websites, but they will still be available upon request. Software tools and hardware design documents will also be distributed via Google Code.

4 Policies and Provisions for Re-use and Re-distribution

In general, no restrictions on the use of data will be imposed for research and educational purposes. Although commercial uses will not explicitly be supported, R&D efforts in industry will be supported. However, the experimental nature of the data will be highlighted; the PIs will thus decline any and all responsibility for the use of the data through a disclaimer. A requirement of acknowledgement will be imposed on the use of the data. The re-use policies for publications will be governed by the policies enacted by the publishing entities (e.g., ACM, IEEE, etc.).

5 Plan for Archiving and Preservation of Access

Unprocessed data will be stored on the servers within the departments of the PIs immediately after the experiment, and back-up copies will be uploaded to the online backup services. The servers are backed-up daily and managed by the individual university and departments' computing services. Processed and validated data will also be stored on the servers and on DVDs no later than six months after the tests.

The data will be kept archived in the presented manner for an initial period of ten years following the completion of this research. Several five-year extensions may be granted depending on the interest of the users and technological progress in the research area.