Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D1BH4R

Title: SMART Detroit MODES

Creator: Michael Hale - **ORCID:** 0009-0004-4055-6078

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

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

Funding opportunity number: 87 FR 58187

Grant: SMARTFY22N1P1G31

Template: Digital Curation Centre

Project abstract:

The City of Detroit's Office of Mobility Innovation has been selected to receive a $2 million award from the USDOT Strengthening Mobility and Revolutionizing Transportation (SMART) grant program for the Detroit Mobility Optimization through Data for Equity and Safety (Detroit MODES) project.

Detroit MODES will deploy state-of-the-art intelligent infrastructure to create a series of smart intersections, collect key roadway safety data, derive real-time impacts of roadway safety interventions, and ultimately enable the City of Detroit to deliver equitable road safety outcomes where they are needed most, quicker and more effectively. The Detroit MODES project team is comprised highly specialized local and international experts, partners, and AI-driven technology. This program prioritizes working with local talent and partners, and will introduce more equitable access to workforce training opportunities for Detroiters in smart infrastructure.

Detroit MODES has the potential to revolutionize the way Detroit approaches road safety and the adoption of new transportation modalities. We believe that this project can become a model for other cities in the future and will further enhance our position as the global leader in mobility.

Start date: 09-15-2023

End date: 03-15-2025

Last modified: 01-22-2024

Copyright information:

SMART Detroit MODES

https://cidetroitmius-my.sharepoint.com/:f:/r/personal/gearat_detroitmi_gov/Documents/2-%20Detroit%20MODES%20-%20SMART%20Grant/DMPEP/Data%20Management%20Plan%20(DMP)/TYME%20DMPEP/DMP%20Attachement?csf=1&web=1&e=EOvFDg

Stakeholder Engagement Data

Qualitative data will be collected to identify the needs and ideal experiences of the users of traffic/mobility data (stakeholders). This data will be collected through in-person interviews conducted with stakeholders who have volunteered to take part in the initiative. Interviews will cover topics including: current workflows for accessing and utilizing traffic/mobility data, pain points in current processes, identification of ideal future workflows, and how data should be presented, consumed, and utilized to achieve desirable outcomes. Participants will be identified only through first name, and as an "internal stakeholder" (City of Detroit), "external stakeholder," or "community member"; Personal Identifiable Information (PII) will not be collected. Qualitative data collected and the resulting final reports will be staged in secure folders provided by the City of Detroit.

System Data

Primary Data Source Overview

UrbanLogiq will collect and ingest two primary sources of data:

Miovision intersection vehicle, pedestrian, and other traffic count and classification data

Derq vehicle and pedestrian counts, near-miss, traffic accident, and pedestrian intent data

See the section on data collection for details on data type, volume, and format. Also see other relevant sections for more details on the subjects of data sharing, retention, and access.

It is anticipated that approximately 2 years' worth of historical data will be ingested for this study, where applicable.

Ingested data will take the form of events in time associated with real-world assets. Using the primary Miovision and Derq sources as concrete examples, vehicle and pedestrian count and classification events will be associated with smart intersections, roadways, or crosswalks where applicable.

Third Party Data Sources

The team is currently assessing a variety of third-party data providers. All of which consist of mobile phone movement data and/or connected vehicle data. The data that is being assessed will consist of speeds and volumes. The timeframe of this data will be for a 2-year period.

Existing Data

The team would be reliant on historic data sources already available to the city. UrbanLogiq has however worked with dozens of other cities on similar projects, enabling accelerated ingestion and time to insight.

Some existing city asset data (e.g. road segment geometry) may also be ingested as needed.

Please refer to **Appendix A** - Data Source List for more information.

Some data will have a Digital Object Identifier (DOI), that will simplify discoverability and Video Data will be collected through direct RTSP streams. SPaT data will be collected through a UDP stream on the traffic network. Count and Event Data will be shared through the Derq Dashboard export, REST API.

The file name will include the Intersection Identifier, Data type and Timestamp in epoch (example: SiteA-Counts-1704771000.csv). File versioning will be made unique using timestamps.

After the system's configuration, Derq will perform an initial QA review involving humans in the loop verifying performance manually as well as periodic automated QA reviews to ensure that performance is maintained.

More Data Collection Methods

The preferred method for data acquisition is via automated API service integration. For REST API endpoints, UrbanLogiq plans to use a batch data polling approach. Alternative options include file-based ingestion, via the preferred data formats listed below and in **Appendix A**.

Miovision leverages a combination of REST API endpoints and file export functionality, depending on the permanence of the intersection infrastructure. This is a pilot project, however Miovision will be making an API data stream available to UrbanLogiq; it will also be making improvements to its API systems.

The team is able to ingest historical count data sources already available to the city (e.g. consultant reports). Structured data is preferred to accelerate ingestion, e.g. sources such as CSV, Excel, JSON, or XML.

Data is assumed to be in the rough order of magnitude range of ~10-100MB/ day per data source. Data will be compressed upon ingestion into the systems.

Please refer to sections on data storage, versioning, and restoration for additional information.

More Data Consistency and Quality Assurance

Sample data will be collected for initial calibration of data quality dimensions with both internal and external stakeholders. The intent is to ensure that the data used to make decisions is as reliable and accurate as possible.

Most data capture will be programmatically controlled and automated. Changes to the data and data capture code will be reviewed by the team staff, including data scientists, engineers, and an internal data quality team. Where applicable, we will leverage tools such as quality checklists, spot checks, and reconciliations against available cross-reference data sources. Code changes and data handling will be documented and reviewed by multiple parties.

A User Guide will be included outlining information available within the Dashboard, and a user guide on how to access such data, export it, and interpret it. An API document will also be made available as needed. Consent for data preservation and sharing will be explicitly obtained from data contributors, ensuring transparency about the purposes, duration, and conditions of data preservation and sharing. The Derq System will use common data formats and communication protocols to upload the data on the Derq Dashboard or stream it through the API.

The team aims to generate data that is as intuitive, navigable, and easy to understand as possible. This includes features such as date selectors, categorical filters, performance metrics, and charts. Data sources with intuitive names are categorically listed and filterable. These elements and more will be explained during the scheduled training sessions.

In addition, team's staff are able to access (via API) and relay relevant underlying metadata associated with raw data sources. This includes raw informational details like creation date, units of measurement, data formats, assumptions made, variables, and vocabularies. All of the original data is retained in our secure Azure containers. Much of this metadata is stored inline next to the transformational logic as well as in explanatory documents and systems such as: Gitlab Merge Request descriptions, code comments, and Confluence documentation. Reaching out to our Customer

Success team is the first step should any party wish to better understand certain transformational logic or data better.

The team's systems inherently store activity logs associated with data access and updates. This includes internal metadata structure, as well as the data catalog object structure for details such as date of creation, last edited by, access control logs, etc. In-house data engineers and developers have access to this information to verify data integrity and to audit changes made by users. The team uses an internally developed metadata format, the description of which is available to licensees.

UrbanLogiq's workspaces are additionally accompanied by a boilerplate "About this Workspace" modal window that provides a brief explanation of that data type to help with interpretability.

Fig 1 - Intersection Count "About this Workspace" Modal

The team only collects username and contact information in its Dashboard but doesn't collect any sort of critical Personal Identifiable Information (PII). Count data and event data do not collect any PII by design. All video data processed by the system is anonymized in such a way that no PII can be inferred from it. For any other data types stored within the system, if any, we will obtain approval from the project owners and stakeholders. All the data reported by the system is only made available to registered and logged in users of the system who have given prior consent for us to store, report this data and create visualizations for these users. We used industry best practices to ensure that the data is stored and transferred securely, including, but not limited to, using password protection on the servers where the data is stored as well as using secure connections to transfer the data.

All data provided through the team's platform has been obtained legally, either through a contract with the data provider, or directly from clients with consent. Where data is requested for sharing or re-using with additional parties, formal written consent must be obtained, and the appropriate parties will be notified.

Data planned for ingestion will not be PII. Intersection traffic counts and vehicle categories are not typically accompanied by identifiable information. While crash or near miss incidents may be linked to personal information via publicly available sources, we will strip out any PII encountered in these datasets. Nevertheless, data ingested into our platform is stored in a manner that protects identity and obfuscates identifying factors.

A team member has its own internal data governance board to manage ethical concerns, should they arise. They have developed a series of enforceable policies and procedures regarding ethical issues, PII, data sharing, security and privacy, networked systems, information exchange, as well as information sensitivity and classification. They are certified and currently in good standing with ISO 27001 and 27701 standards with attestations for 27017 and 27018.

The team adheres to all Canadian and United States intellectual property regulations, including both data and the platform supportive systems. We have obtained consent to use all data in our systems from the appropriate data controllers and will remove it when the consent is revoked, or the validity period expires.

The team only collects username and contact information in its Dashboard but doesn't collect any sort of critical Personal Identifiable Information (PII). Count data and event data do not collect any PII by design. All video data processed by the system is anonymized in such a way that no PII can be inferred from it. For any other data types stored within the system, if any, we will obtain approval from the project owners and stakeholders. All the data reported by the system is only made available to registered and logged in users of the system who have given prior consent for us to store, report this data and create visualizations for these users. We used industry best practices to ensure that the data is stored and transferred securely, including, but not limited to, using password protection on the servers where the data is stored as well as using secure connections to transfer the data.

Most of the team's systems are made in-house, so licensing of 3rd party software becomes less relevant. UrbanLogiq's code and technology is itself copyrighted (CommunityLogiq Software ©). UrbanLogiq complies with relevant open-source licensing requirements where applicable for external systems.

See (How will you manage any ethical issues?) for more details surrounding identity protection. Also see the section on data sharing, access, and security for details on relevant security and privacy concerns.

Data is stored in the team's cloud buckets in the US, leveraging industry-standard secure and durable data storage. Additional storage server replicas may be used for redundancy and availability. Automatic Backups are also enabled with long term storage options. Team's DevOps/Infra Manager will be responsible for ensuring proper data backup and recovery. No data will be stored on endpoints as per our industry best practices. Also note that the team is pursuing the SOC 2 certification on data cybersecurity.

The team already has access to sufficient commodity data storage capacity via Azure cloud; as such it will not charge for additional data storage products or services.

Data will be backed up nightly. Data file revisions are retained in Azure and can be recovered as needed after updates. Data will be retained indefinitely unless consent expires for the data. The team's CTO and/ or trained senior developers will be responsible for these nightly backups as well as any restoration operations, should they become necessary.

The team can roll back data streams to earlier states as needed. Azure data backup and restoration processes are utilized and tested nightly.

Data is automatically replicated within the United States using multi-region replication for blob data. The data in Azure is always replicated to ensure durability and high availability. Data storage on local machines is minimized and discouraged as much as possible.

Data access is provided through our team's Dashboard. Authentication required an email/password Authentication method with mandatory Multi Factor Authentication, providing a time-based session with expiry. Authorization is implemented through RBAC (Role Based Access Control) to prevent unauthorized access to data, implementing a Zero Trust Architecture. Both Authentication and Authorization are monitored and logged using Audit Trail methods.

The team also has a SOC2 Compliance framework implementation under-progress. Data storage can also be secured using state-of-the-art At-Rest encryption, as required.

Also access to data / services used by the team is tightly controlled via a number of methods, including but not limited to:

- User access provisioning and deregistration
- User access management
- User policies and responsibility training
- Secure log-on
- Network access control
- Administrator access control
- Source code access control

Data access will be limited to those who require it, adhering to privacy by design principles. The team's data scientists and data engineers will work with this data on a regular basis. Data access provisioning for 3rd parties will be controlled by the team's CTO and trained data engineering staff.

The team may collect, hold, process, or share personal data, a valuable asset that needs to be suitably protected. Every care is taken to protect personal data from incidents (either accidentally or deliberately) to avoid a data protection breach that could compromise security. Our infrastructure was designed to be secure and GDPR compliant from the ground up.

The team employs the most rigorous of data protection handling procedures to all of its data, ensuring that storage and transference of data is compliant with GDPR and ISO 27001 and 27701 standards. The team elects to follow General Data Protection Regulation (EU) 2016/679 ("GDPR") guidance to have in place an institutional framework designed to ensure the security of all personal data during its lifecycle, including clear lines of responsibility. This applies to all staff of the team, including temporary, casual or agency staff and contractors, consultants, suppliers and data processors working for, or on behalf of the team.

Data generated and stored by the system includes event and count data, in the form of video snippets and JSON files resulting from the Derq processing pipeline. All data is retained within the team's system only when there is a continued and valid reason to store or process the data. The data is retained as desired to facilitate benchmarking and comparisons over extended periods of time. In some instances, the data could even be used to enhance any forensic or root-cause analysis for a given event. Since no PII is recorded and reported, no data needs to be destroyed. Additional research uses of the data will include performance improvements and further developments to the team's platform. This data is typically stored for a year after the license expiry or for the retention period specified in the contractual agreement or in the City's data retention policies.

Also, the team has a responsibility to protect the integrity and confidentiality of personal data held with regards to our clients, employees, and partners. Employees also have an obligation to protect the integrity and confidentiality of personal data and to prevent unauthorized disclosure of data whether it is oral, printed, handwritten or computer based. All information, in any format held by the team, must be destroyed in a way which does not breach the data protection rights of our employees, contractors and customers.

Raw and processed data (including personally identifiable data) will be retained or destroyed in accordance with legal, ethical, contractual, compliance, and conformity requirements. We will only retain data or personal information for as long as necessary to fulfill the purposes we collected it for. To determine the appropriate retention period for personal data, we consider legal retention requirements, the amount, nature, and sensitivity of the personal data, the potential risk of harm from unauthorized use or disclosure of personal data, the purposes for which we process personal data and whether we can achieve those purposes through other means, and the applicable legal requirements.

In addition, the stored team's data can be used for predictive model training, university partnership research projects, etc. however only with the consent of the data controller.

Requests for data retention or destruction will be fielded via our customer success department or our leadership team. We will retain or dispose of data as required by its contractual, legal, or regulatory obligations. If necessary, data will be converted to file formats that are more readily accessible to the relevant parties.

Derq data is stored in cloud servers (such as an AWS S3 service), providing long term and durable data storage. Using this service to store the data collected at any given intersection would be at a minimal charge per intersection per year. Additionally, using services such as AWS intelligent tiering, low frequency accessed data will automatically be transferred to low-cost long-term storage tiers, causing little to no impact to time/efforts to preserve data long term.

In addition, long term data storage and preservation will utilize similar Azure storage systems and methods as used during active operations. Azure commodity systems allow for multiple data storage tiers depending on activity level. As UrbanLogiq is not charging for data storage costs, no additional budget has been allocated for long term storage, sharing, or preservation. UrbanLogiq staff will ensure every effort is made to effectively curate and manage the data beyond the active engagement, subject to requirements.

Data will be shared through the team’s Dashboard. Users that have access to the data will be approved and managed by the city. The team’s system can provide or enable a user with a manager role to add, remove and modify user profiles with Standard roles and control who has access to the data. Those users can be either internal or external to the City’s organization. Data will be made available as soon as the processing and upload of the data is complete (i.e., in near real-time).

Also, users will have access to the team’s platform, which has intuitive interfaces for data discovery via use cases. Clients will be able to contact the team’s customer success staff if they have any questions about data availability and platform navigation. Data will become available to clients as soon as it is ingested into the team’s platform.

Data sharing will be enabled for clients where there is an established contractual agreement for intended purposes. Data sources, where provided, are directly exportable from the team’s platform by clients. Consultant reports and raw data, with accompanying metadata, is exportable by the platform’s users. Available dates of each data source are provided directly in the platform. Platform users may additionally share generated reports with other authorized users (i.e. once a use case report has been generated for a selected location).

Fig 2 - Share and Export Functionality

Data can be shared via various secured methods. Our chosen data storage formats enable straightforward export and sharing of data as needed. The team has several methods available for securely sharing data such as SFTP and 3rd party Azure containers.

Some of the preferred methods of transferring data are as follows:

- SFTP via onramp.UrbanLogiq.com
- SFTP via customer-provided server

Information will only be sent that is necessary for the stated purpose. Unnecessary data, PII, and any data not required will be redacted or removed completely (as appropriate) before transfer. Relevant parties should be notified in writing prior to any usage of data beyond intended purposes.

Data will be persisted via intuitive naming. For example, using data source and data type, such as Miovision Turning Movement Counts. This will enable clients to easily identify and use their data. As already stated, data is persisted in team’s Azure blob storage.

The team shares access to the data it generates exclusively with its customer organizations that have bought a license to operate the system. The data is made available only to registered and authorized users and cannot be readily shared outside of the platform without access to the Dashboard or the system’s API. A team member doesn’t share city specific data externally, but the city can decide, if it wishes, to share its data to third-party entities with the use on NDA or not, which might in turn require a data sharing agreement with the team member or the City directly. Meanwhile, IPR developed from the generation, reporting and use of the data will remain the property of the team.

Potential restrictions in data sharing focus around contractual and legal obligations with the various involved parties. Should continue access to data ingested by the team be required by an additional party, a data sharing agreement should be discussed. The team will flag any potential sources of data restriction during the ingestion process.

Data conversion may be required to enable some parties to utilize the data, which should be considered in any drafted data sharing agreement.

The responsibility of implementing the DMP goes back to the Development Operations (DevOps) and Infrastructure Manager who will oversee the entire process, under the supervision of the Engineering Manager and ultimately the CTO. The QA manager will also be in charge of the performance evaluation and quality assurance step of the process.

Apparently, the data engineering team is responsible for implementing this data management plan on a day-to-day level. Data ingestion (capture), storage, monitoring, and sharing are also the responsibility of the data engineering team. Developers are also responsible for documentation, accessibility, and quality of the data and metadata.

It is the responsibility of senior management, including the CTO and Data Privacy Officer, to ensure that this plan is reviewed and revised as necessary. Managing backup and restoration activities, as well as areas such as contractual and legal compliance, platform security, privacy, and access provisioning are primarily the responsibility of the senior developers and technical leaders. The leadership is ultimately responsible to ensure that relevant policies will be respected.

Privacy and security are everyone’s responsibility. All parties involved must do their part to ensure that intellectual property, proprietary data, and private information do not fall into the wrong hands.

The team will deploy its on-premises server as well as its software engine (including cloud hosting, storage and applications). The team doesn't require any additional resources or expertise to deliver on the plan. Charges can be applied by data repositories at the time of renewal after the initial operation period.

Our staff and contractors are engaged and ready to deliver on this data management plan. Data science specialists will be employed or contracted to deliver insights and to support decision making. Senior developers and data engineers will ingest and provision different data sources, as well as develop and implement specific use cases for our platform users. Our customer success team will be leading training and documentation around platform usage, as well as deliver progress reports and updates to the relevant stakeholders.

The team already has access to relevant commodity hardware and software needed to fulfill its obligations. No new charges to the client will be applied by its data repositories or cloud infrastructure.

Planned Research Outputs

Collection - "More information necessary, could you please provide us with guidance ."

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
More information necessary, could you please provi ...	Collection	Unspecified	Open	None specified		None specified	None specified	No	No