Phase 2 Data Management Plan (DMP)

Georgia Department of Transportation: Safe Trips in a Connected Transportation Network ITS4US Deployment Project

www.its.dot.gov/index.htm

Final Report — August 21, 2023
FHWA-JPO-22-974

ITS4US

IT'S TRANSPORTATION FOR ALL OF US

U.S. Department of Transportation

Produced by Georgia Department of Transportation (GDOT)
U.S. Department of Transportation
Intelligent Transportation Systems Joint Program Office
Federal Highway Administration
Office of the Assistant Secretary for Research and Technology
Federal Transit Administration

Notice

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof.

The U.S. Government is not endorsing any manufacturers, products, or services cited herein and any trade name that may appear in the work has been included only because it is essential to the contents of the work.

Technical Report Documentation Page

1. Report No.

FHWA-JPO-22-974

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle

Phase 2 Data Management Plan (DMP)

Safe Trips in a Connected Transportation Network ITS4US Deployment Project

5. Report Date

August 21, 2023

6. Performing Organization Code

(Delete and insert information here or leave blank)

7. Author(s)

Alan Davis (GDOT), Kofi Wakhisi (ARC), Bennet Foster (ARC), Randall L. Guensler (Georgia Institute of Technology), Angshuman Guin (Georgia Institute of Technology), Polly Okunieff (GO Systems and Solutions), Natalie Smusz-Mengelkoch (ICF)

8. Performing Organization Report No.

(Delete and insert information here or leave blank)

9. Performing Organization Name and Address

Georgia Department of Transportation – One Georgia Center
600 West Peachtree NW,
Atlanta, GA 30308

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

693JJ32250011

12. Sponsoring Agency Name and Address

U.S. Department of Transportation
ITS Joint Program Office
1200 New Jersey Avenue, SE
Washington, DC 20590

13. Type of Report and Period Covered

Final

14. Sponsoring Agency Code

HOIT-1

15. Supplementary Notes

The following USDOT partners are supporting this document development: Elina Zlotchenko (Program Manager), Sarah Tarpgaard (Agreement Officer), and Norah Ocel (Agreement Officer Representative)

16. Abstract

The Georgia Department of Transportation (GDOT) ITS4US Deployment project, Safe Trips in a Connected Transportation Network (ST-CTN), is leveraging innovative solutions, existing deployments, and collaboration to make a positive impact using transportation technology to support safety, mobility, sustainability, and accessibility. The ST-CTN concept is comprised of an integrated set of advanced transportation technology solutions (connected vehicle, transit signal priority, machine learning, predictive analytics) to support safe and complete trips, with a focus on accessibility for those with disabilities, older adults, and those with limited English proficiency.

The Data Management Plan (DMP) provides an inventory of the datasets and their characteristics related to the GDOT ITS4US project –ST-CTN. The inventory includes datasets that are ingested, generated, processed and exported by the ST-CTN system including static, realtime, and archived datasets. The plan includes information on data governance, management, security and privacy policies, storage and access.

17. Keywords

ITS4US; Deployment; ITS; Intelligent Transportation Systems; Safe Trips in a Connected Transportation Network; ITS4US; Data Management Plan

18. Distribution Statement

19. Security Classif. (of this report)

Unclassified

20. Security Classif. (of this page)

Unclassified

21. No. of Pages

(72)

22. Price

N/A

Revision History

Table i. Revision History

Name	Date	Version	Summary of Changes	Approver
ARC Team	23 August 2021	1.0	Final Phase 1	USDOT
GDOT Team	26 June 2023	2.0	Draft Phase 2	Bennett Foster
GDOT Team	31 July 2023	2.1	Final Phase 2	Bennett Foster
GDOT Team	21 August 2023	2.2	Final Phase 2	Bennett Foster

PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Document Purpose

This Data Management Plan (DMP) provides an inventory of the datasets and their characteristics related to the Georgia Department of Transportation (GDOT) ITS4US project – Safe Trips in a Connected Transportation Network (ST-CTN). The inventory includes datasets that are ingested, generated, processed, and exported by the ST-CTN system including static, realtime, and archived datasets. The plan includes information on data governance, management, security and privacy policies, storage, and access, as well as the relationship of the data to performance measures.

During the Phase 1 effort, the DMP covered the ST-CTN project's preliminary approach to managing the datasets including identifying security, privacy and governance policies related to existing datasets and describing proposed system datasets and their dependent standards. As the system design and testing phases progress, additional details on data needed for performance measures were added.

This document is intended for technical reviewers, independent evaluators (IE), designers, researchers, and other persons interested in the datasets and their derivatives that drive or are produced by the ST-CTN system.

1.1.1 Organization of this Document

This document includes the following sections:

Section 1: Introduction – description of the document purpose, system concept and an overview of the data schedule and needs.

Section 2: Data Stewardship – description of the owners and stewards of the data curation (quality, storage, retention, and access), privacy provisions, and relationship to performance data including baseline datasets.

Section 3: Data Standards – description of adopted data standards, versioning control and metadata approach.

Appendix A: Acronym and Glossary – tables that list the acronyms and terms used throughout this document.

Appendix B: Comprehensive List of Standards – description of the standards used for ingesting and managing the datasets.

Appendix C: Data Impact Log – the data impact log serves to capture changes to the DMP since the Phase 1 version.

1.1.2 Change Control

This DMP is a living document that is updated during each phase of the three-phase project development – concept development (Phase 1), design and test (Phase 2), operate and evaluate (Phase 3).

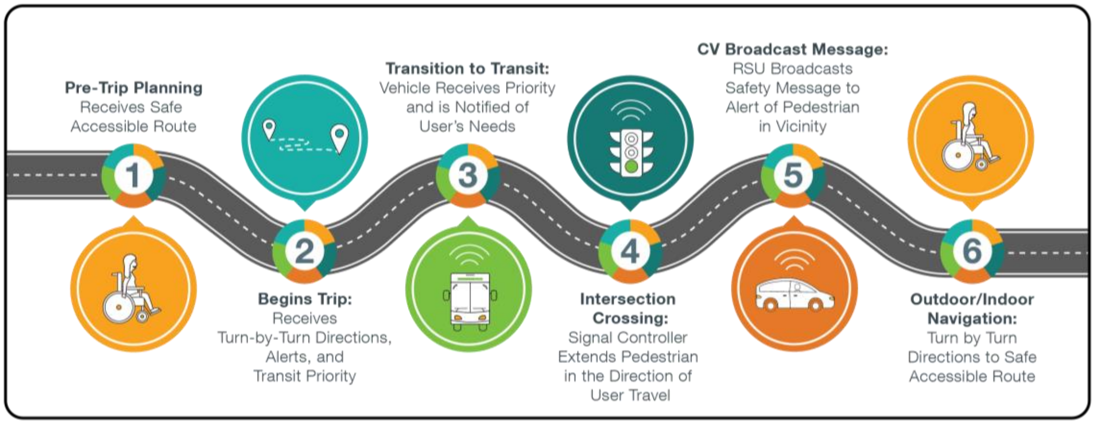
During this deployment phase, a more refined set of datasets, related access levels, storage and rules of curation and standards are described. As the project operation phase is rolled out, this plan will be finalized to reflect the most up-to-date, as-built information. Each version number will be incremented numerically in the Revision History table (see page i). The whole number, starting with version 1.0, will reflect the phase, and the decimal will reflect incremental updates during a phase, as necessary.

Specifically, the change to the document will be reflected not only in the **Revision History** table, but will be captured in the **Dataset Impact Log** included in **Appendix C.** The list contained in **Table 13** includes the state of the dataset by DMP version, and the reason for the change or removal (if applicable). Several rules are implemented to ensure the integrity of the comprehensive list:

- No identifier will be reused or replaced for a different dataset. Currently, each dataset is associated with a unique identifier.
- References to the specific DMP table and row will reference where the change was made. For example, if the Data Custodian was replaced or standard compliance changed due to an update of the standard specification.
- DMP version changes will include the previous and current versions from which the change was made. This ensures that changes may be made each time the document is reissued.
- When datasets undergo multiple changes between versions, only one entry will be inserted into the table, and all the reasons for the change(s) shall be listed in the reason column.

1.2 System Concept

The ST-CTN project aims to upgrade and integrate existing technologies and services to assist underserved populations with completing their complete trip successfully, safely, and reliably. The vision of the project is to provide users complete trip functionality with directions, conditions, and status on the links between trip legs that are personalized based on the user’s profile, while connecting the user to Connect Vehicle (CV) infrastructure to provide safer trips and more transportation network awareness. As an illustration of how the ST-CTN system will be used, transit-based trips were delineated into six segments (as depicted in **Figure 1**) to allow for easier understanding and a greater breakdown of priorities and goals.


Pre-Trip Planning
Receives Safe Accessible Route
1
2
Begins Trip:
Receives Turn-by-Turn Directions, Alerts, and Transit Priority
Transition to Transit:
Vehicle Receives Priority and is Notified of User's Needs
3
4
Intersection Crossing:
Signal Controller Extends Pedestrian in the Direction of User Travel
CV Broadcast Message:
RSU Broadcasts Safety Message to Alert of Pedestrian in Vicinity
5
6
Outdoor/Indoor Navigation:
Turn by Turn Directions to Safe Accessible Route

Source: ARC, 2020

Figure 1. Traveler's Complete Trip

The delineated trip segments include the following steps and project components:

- **Step 1 Pre-Trip Planning.** The traveler plans for and receives a safe accessible route.
  - The ability to customize trip preferences based on the user's abilities.
- **Step 2 Begins Trip.** The traveler begins their trip and receives turn by turn directions, alerts, remote pedestrian activation, and can trigger transit signal priority (TSP) if the user requires additional time boarding or alighting a transit vehicle, is unable to stand for long periods, or is sensitive to weather conditions.
  - Turn by turn, shortest path, directions along pathways that meet user-defined preferences.
  - Provides support services for users if they become disoriented or have issues accessing defined paths.
  - Activates TSP for buses if the user requires additional time boarding or alighting a transit vehicle, is unable to stand for long periods, or is sensitive to weather conditions.
- **Step 3 Transition to Transit.** The traveler transitions to transit and the transit vehicle receives priority and is notified of a user's needs. TSP can be triggered if the bus is running behind schedule due to a longer boarding time needed by a user.
  - Provides users with transit trips that have accommodations that meet user-defined preferences.
  - Sends alerts to transit vehicles when users need additional time to board, navigate internally, or alight the transit vehicle.
  - Remotely requests service from transit vehicles while waiting to board or alight.
  - Triggers TSP if the bus is running behind schedule due to a user needing additional time to board or alight.

- **Step 4 Intersection Crossing.** When crossing a signalized intersection, the traveler indirectly interacts with the signal controller in that an authorized system call is made to the controller which extends the pedestrian phase in the direction of user travel.
    - Allows the user to communicate with connected intersections if they are unable to reach or press the crosswalk button.
    - Provides the user with information about the intersection crossing and adds time to the crossing if needed.
- **Step 5 CV Broadcast Message.** Roadside units (RSUs) broadcast safety message to alert CVs of pedestrians in the intersection.
    - Provides the ability for users to remotely request service from transit vehicles while waiting to board or alight.
    - Provides communications to CVs from pedestrian crossing signal system (via RSUs) to make them aware of pedestrians crossing a roadway.
    - Provides communications between transit vehicles and travelers waiting at a transit stop to make them aware of each other.
- **Step 6 Outdoor/Indoor Navigation.** The traveler is provided with turn-by-turn directions to a safe accessible route.
    - Hands-free navigation via mobile apps and/or wearables and accessible channels (haptic, voice, text).
    - Alerts and dynamic rerouting in response to changes in path conditions.
    - Provides the user with accessible routes into and through transit hubs within the project area.
    - Provides users with updates on the operating status of indoor infrastructure such as elevators and escalators.

System development and system integrations completed within the scope of this project will enable travelers – specifically those in the underserved community – to program and safely complete single mode or multimodal trips that are based on their abilities; improve the transition between modes by providing additional details to users and transit service operators; suggest dynamic routing changes based on infrastructure condition and calculated delay.

The existing initiatives that are being leveraged to support the proposed ST-CTN system are defined in more detail below as well as those components that will be developed specifically to support ST-CTN project evaluation. The icons and colors depicted below are used throughout the DMP to clearly identify the critical components of ST-CTN. In some cases, partner agencies are upgrading the services within their current systems to create a more robust dataset or toolset for the ST-CTN program.

OTP for G-MAP. Atlanta Rider Information and Data Evaluation System (ATL RIDES) includes an Open Source Software (OSS) multimodal trip planning and mobile application, integrated mobile fare payment options, and a Connected Data Platform (CDP) using regional General Transit Feed Specification (GTFS) transit service data. The tool supports multi-agency context, multilingual support, and mobile app turn-by-turn directions. The OpenTripPlanner (OTP) architecture facilitates integration with additional OSS tools including a data analytics engine, call center module with application programming interfaces (APIs), and account management system. The existing ATL RIDES application will be modified and enhanced based on ST-CTN project needs and leveraged to create a new, independent application which will be differentiated as the *OTP for G-MAP* subsystem (hereafter referred to as G-MAP).

SIDEWALKSIM. SidewalkSim is an asset management system and shortest path (lowest impedance) routing tool for pedestrian pathways. Site inspections provide more detailed ADA and inclusive design and condition data for use in pathway accessibility analysis. SidewalkSim identifies the best path between any two points in the pedestrian network, given the set of pathway characteristics and any user-specified needs and route penalties.

CV1K. The Atlanta region is home to one of the largest CV deployments in the United States – Regional Connected Vehicle Infrastructure Deployment Program (CV1K). CV1K is deploying interoperable CV technologies at signalized intersections throughout the Atlanta region using both Dedicated Short-Range Communications (DSRC) and Cellular Vehicle to Everything (C-V2X) technologies to deliver safety and mobility-based applications. The program provides support to configure, operate, and maintain CV infrastructure and applications, including TSP. Gwinnett County is one of the largest recipients of the first phase of this deployment.

CVTMP. Gwinnett County's Connected Vehicle Technology Master Plan (CVTMP) sets out to develop and improve economic viability and quality of life, address the needs and challenges to motorized and non-motorized modes, establish guidelines for deploying technology, and have broad applicability to Gwinnett, other local jurisdictions, and across the state—to set the standard for implementing CVs. Among the high priorities is establishing a mobile accessible safety program and alternative strategies for TSP in Gwinnett County.

STM. The Space Time Memory (STM) platform processes traffic volume and speed data from multiple monitoring and modeling sources, tracks network performance measures, and predicts evolving route conditions using traditional and machine learning techniques. The STM projects trip trajectories through the transportation network, as network conditions change in space and time. This tool will be applied to analyze and predict performance through the multimodal transportation network. The shortest path analysis will be applied to the combined roadway, transit, sidewalk, and shared-use path networks, allowing routing decisions to incorporate travel time, safety, and other impedances into path selection.

PMD. The PMD is a data storage and distribution tool that archives operational, survey and performance metrics of the system. The PMD is envisioned to ingest, quality-check, and curate all types of data – static and dynamic, structured and unstructured, open and private datasets. The PMD will store the datasets so that they can be viewed and accessed based on user roles. The PMD will include a public PMD for the public to access open data presented on an interactive dashboard to the public.

1.3 Data Schedule

As the project progresses, the Data Schedule will extend through all three phases of the project and into the operational and maintenance period. The DMP will be updated to reflect the most current inventory of datasets (and their characteristics) needed to operate and continually evaluate system operations and performance.

For Phase 1, only limited information was available for determining the timing of data milestones. During Phase 2, as the system design is rolled out, additional schedule information is detailed about data milestones. **Table 1** provides a schedule of data milestones. The times and dates are intended to provide a general expectation of when data milestones are expected to occur.

Table 1. Data Schedule

ID	Event Title	Description	Date
1	Draft Phase 1 DMP is delivered to USDOT	• Draft Phase 1 DMP with basic information known at the time of writing.	July 2021 (Phase 1)
2	Final Phase 1 DMP	• Phase 1 DMP is updated with USDOT comments addressed.	August 2021 (Phase 1)
3	Select data specification for data flows	• Identify and profile data standards for use in project or develop draft specifications for each subsystem (with no applicable standard); • Acquire and reconcile data dictionaries from each subsystem.	Ongoing through Agile process May 2023 – June 2024 (Phase 2)
4	Data specification profiles and curation processes	• Describe specific schemas and profiles (i.e., Interface Control Document (ICD)) and common dictionary / semantics used for current systems (G-MAP, CV, STM). • Describe curation and metadata processes for shared datasets. • Describe method of access. • Submit to Institutional Review Board (IRB) for data with personally identifiable information (PII).	Ongoing through Agile process May 2023 – June 2024 (Phase 2)
5	Draft Phase 2 DMP is delivered to USDOT	• Draft Phase 2 DMP.	June 26, 2023 (Phase 2)
6	Final Phase 2 DMP	• Phase 2 DMP is updated with USDOT comments addressed.	August 21, 2023 (Phase 2)

ID	Event Title	Description	Date
7	Draft Phase 2 Data Privacy Plan (DPP)	• Draft Phase 2 DPP.	July 17, 2023 (Phase 2)
8	Final Phase 2 Data Privacy Plan (DPP)	• Phase 2 DPP is updated with USDOT comments addressed.	September 5, 2023 (Phase 2)
9	Notice of Privacy Management Consistency	• Memorandum confirming consistency with the Privacy Management Plan.	September 11, 2023 (Phase 2)
10	Initial meeting with USDOT data team to set expectations and review forthcoming data	• Meeting to review data with USDOT and walkthrough the DMP.	September-October 2023 (Phase 2)
11	Pedestrian data collection	• Initial collection of data on current conditions starts. • **Collect Gwinnett Sidewalk Data:** Organize and train staff to collect Gwinnett County sidewalk datasets. Collect, review, and generate into STM format. • These collection periods will include public right of way (PROW) and other datasets. No PII will be shared. • Data storage is ready for initial input.	June - October 2023 (dependent on weather) (Phase 2)
12	Performance Measure Requirements	• Describe datasets needed to generate performance measurement (and methods for collecting data).	December 2023 (Phase 2)
13	Enterprise Data Governance (EDG) Data Committee	• Establish EDG data committee to govern data for integrated datasets. • The EDG data committee will include USDOT and IE representatives as observers.	April 2024 (Phase 2)

ID	Event Title	Description	Date
14	Baseline metric data provided to USDOT	• Baseline datasets and metadata files are made available for the USDOT and the IE to access.	Starting in April 2024 when beta testing begins (Phase 2)
15	Month of testing of applications begins	• Initial upload of "after datasets" are collected and stored on project research data storage systems through testing.	Starting in April 2024 when beta testing begins (Phase 2)
16	System Test Results Summary submitted (test report will include a section on data fidelity)	• System Test Results Summary submitted to USDOT.	April 2024 (Phase 2)
17	System Test Results Summary Documentation submitted	• System Test Results Summary Documentation submitted to USDOT.	April 2024 (Phase 2)
18	Data accessed by USDOT	• Daily updates of *after case data* are available to USDOT and IE to access.	May 2024 (Phase 2)
19	Initial data samples provided to USDOT	• Initial Data samples (e.g., SidewalkSIM, General Transit Feed Specification (GTFS)) are created, validated, and submitted to USDOT for review. • Note: The datasets will evolve due to the Agile approach to developing G-MAP and the ongoing deployment of CV applications.	May 2024 (Phase 2)
20	Data Review	• Data Review conducted with USDOT and IE to ensure datasets and metadata files are complete.	May-June 2024 (Phase 2)
21	Phase 3 data provided to USDOT	• Performance and other data supporting a comprehensive assessment of deployment impacts to be shared with IE.	Pending discussions with IE (Phase 3)

1.4 Data Needs Summary

The following sections provide an overview of the data that will be required to support the functionality of the ST-CTN system.

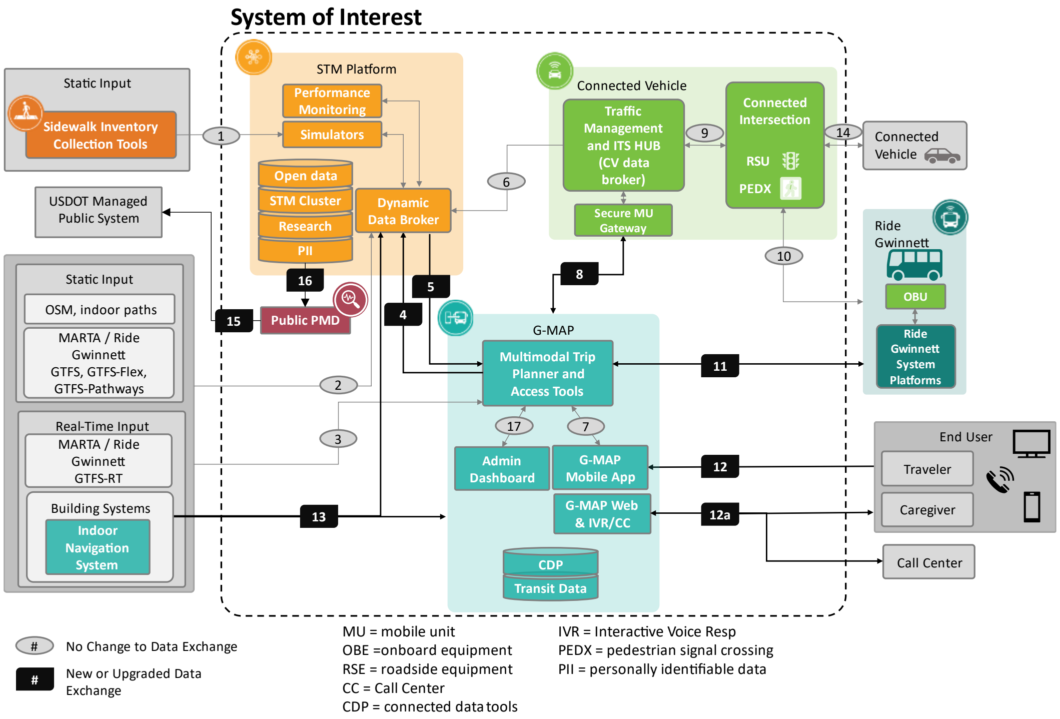
1.4.1 Data Needs Summary

The Data Needs are described by the Functional Architecture Viewpoint which describes the functions and information flows between the major subsystems and with external systems, including the data needed to support the performance measures.

Source: GDOT, 2023

Figure 2 provides the ST-CTN high-level functional and information flow diagram (also referred to as the context diagram) in more detail and corresponds with the data flows described in **Table 2**.

Critical ST-CTN data exchanges are identified by number and color in the context diagram and correspond to the exchanges described in **Table 2**. These descriptions are focused only on the scope of work required to implement the new system and information that is necessary to provide context to that work. As described in **Section 1.2**, there are a number of existing and on-going efforts that are being performed outside the scope of this project that are critical to the success of ST-CTN.



Source: GDOT, 2023

Figure 2. Data Flow / Functional Architecture for ST-CTN System

The system information flows are identified by number in the functional diagram above and described in **Table 2**. The grey oval labels indicate existing data exchanges that will be utilized with no change to the current data exchange. Black rectangular labels indicate data exchanges that will be new or upgraded to support the ST-CTN system.

Table 2. Critical ST-CTN Information Flow Descriptions

Data Exchange ID	Description
1	Sidewalk inventory data, including accessibility features to the STM Platform simulators
2	Static data from various existing sources to the STM Platform dynamic data broker and G-Map OTP Network and other tools including GTFS, OSM, Indoor pathways
3	Dynamic or realtime data from various existing sources to the STM Platform dynamic data broker and G-Map Middleware
4	G-MAP Mobile App logs and trip feedback
5	STM Network Impedance API
6	CV and Traffic Operations Messages: signal phasing and timing road characteristics, traffic data
7	Open Trip Planner (OTP) APIs and G-MAP APIs
8	Mobile Accessible Pedestrian Signal System (PED-SIG)
9	CV messages includes PSM, BSM and other CV messages and Signal Controller monitoring and control information
10	Transit signal priority (TSP) and other CV application messages
11	Ride Request messages
12	G-MAP Mobile App APIs and Traveler exchange – profile, trip plan, settings, notifications, feedback, etc.
12 A	G-MAP Call Center and Trusted Companion – profile, trip plan, settings, feedback (subset of information exchanges tailored for website, call takers and caregivers)
13	Dynamic information from building facilities, including beacon signals to the G-MAP Mobile App
14	CV data (i.e., PSMs)

Data Exchange ID	Description
15	Project data for USDOT-managed Public System
16	Public facing data and visualizations for reporting on system operations and performance
17	G-MAP Administration Tool configuration and control message exchanges

1.4.2 Data Overview

The datasets that meet the needs of the ST-CTN are listed in **Table 3**. The table lists each dataset, referenced with a unique identifier for each dataset and subset of dataset. Each dataset type, scope, and reference to the context diagram are listed in the table. The datasets are described by the following columns:

- **Dataset Identifier (ID)**: Unique identifier related to every dataset or dataset subset.
- **Data Exchange ID (EX ID)**: References interface flow(s) in the Context Diagram (see **Source:** GDOT, 2023
- Figure 2).
- **Dataset Type**: The category associated with the dataset content. Values include (Asset, Asset Condition, Crowdsource, CV, Land Use, Network, Network Operating Conditions, System-Customer Performance, Transit, Vulnerable Road User (VRU) Modes, Weather, Mobility Service API).
- **Dataset Name:** A name or title for the dataset.
- **Dataset Description**: Describes a short description of the dataset, its purpose, origin (particularly which subsystem ingests, uses, or generates it), general content and current status.
- **Dataset Subset Description**: If a portion of the dataset is extracted for special use, the subset is described with its purpose and use.
- **Data Collection Methods**: Describes the method by which data are collected, including input sources. Collection methods include:
  - External – acquired as input from external sources,
  - Derived – summarized, fused or integrated data generated from multiple datasets, or
  - Collect / Forward -- created, collected, forwarded and stored data from system. The *collect / forward* method includes user-input transactions (e.g., between APIs), web forms, or user-tracking methods (e.g., trace data from mobile phones).

Information on data file format will need to comply with Data Standards and profile provisions, so data file formats are included in **Table 12**. The collection method may change over time and impact the quality of the data. Hence data collection methods will be described in the technical metadata description (see **Section 3.3**).

Table 3. Data Overview Table

Data ID	EX ID	Dataset Type	Dataset Name	Dataset Description	Dataset Subset Description	Collection Method
3	2	Network	Whole Road Network	Comprehensive roadway network for Metro Atlanta, including all facility type roadways (referred to as links) and intersections (referred to as nodes). The network is mapped to, and reconciled with, all other network data sources (serving as the master network). The full dataset serves as the underlying disaggregate link-node structure for all roadway networks in the region. It includes nodes needed for future Activity Based Model (ABM) and simulation model application (e.g., large parking lots that input/absorb demand). It is the basis for link-to-link mapping between multi-provider roadway networks. Working datasets are generated for pathway and impedance analyses. Link-and-nodes are collapsed to improve algorithm processing time. Research analysis subsets are created for case studies.	Data subsets are employed in: 1. STM Network structures 2. Connections between sidewalk, transit, and road networks 3. Updating the OSM network 4. Connecting data across travel demand and simulation models	Derived

Data ID	EX ID	Dataset Type	Dataset Name	Dataset Description	Dataset Subset Description	Collection Method
5	5	Network	STM Network	Road and pathway network employed by the STM for impedance calculations and shortest path analyses. Includes all ABM links and as many links from the whole road network as deemed necessary to support mode and pathway analyses. Network is processed and stored with full network and impedance data (historic STM contains all links over time for research and machine learning), working data for current conditions (previous two hours), and forecast conditions from machine learning projections (future one hour).	Data subsets are employed in: 1. Connections between sidewalk, transit, and road networks 2. Updating the OSM network 3. Connecting data across travel demand and simulation models	Derived
8	2, 3	Network	OpenStreet Map (OSM) Network	OpenStreetMap network is needed to support G-MAP OTP engine and STM simulator component. OSM serves as the basis for all routing processes in G-MAP app. The extracted OSM network will be updated to reflect the whole road, STM, and sidewalk networks to ensure data compatibility. APIs will provide updated linkages and transfer-of-path impedance costs from the STM to the OSM format for full compatibility with the routing app. Note: the updated OSM data will not be published during operations; rather, the project team will work with OSM consortium to update permanent changes to the network when necessary.	Data subsets are employed in: 1. G-MAP Wayfinding 2. Connectivity between STM and OSM reference network 3. Processes designed to update OSM spatial accuracy	External Input and Derived during operations

Data ID	EX ID	Dataset Type	Dataset Name	Dataset Description	Dataset Subset Description	Collection Method
10	1	Network	Sidewalk Network	Link-and-node structure for all sidewalks and potential sidewalks developed from parcel-level land use and roadway-link data. Full network includes sidewalks that do not yet exist (coded as width=0 and high link impedance). This dataset is collected through the Sidewalk Collection Tools.	Data subsets are employed in: 1. Referencing between STM and OSM wayfinding network 2. Impedance calculations 3. Shortest path planning analyses	Derived
11	1, 13	Network	Indoor Pathways	The description of indoor pathways including the location and description of vertical conveyances and planned or current obstructions. Includes connectivity to the sidewalk network. Data will be formatted in OSM structure for use in G-MAP app wayfinding.	Data subsets are employed in: 1. Referencing between STM and OSM wayfinding network 2. Impedance calculations 3. Shortest path planning analyses (with impedance)	External Input
15	6	Network Operating Conditions	NaviGAtor Data	Roadway facility volume and speed data mapped to the whole road network and STM links. GDOT roadway operating condition datasets are integrated into the STM for machine learning predictions, performance metrics, and research. Working data for lane-by-lane and corridor speed, volume, and vehicle class splits.	Data subsets are employed in: 1. STM speed and volume data for machine learning 2. Performance measurement	External Input
20	2	Network Operating Conditions	Modeled Future Operating Conditions	Model-predicted (regional ABM and simulation) future on-road spatial and temporal operating conditions (e.g., volume and speed data) mapped to STM links for conditions that may not have been encountered in the observational data used in machine learning analyses. Each modeling run generates subsets of data employed in the STM.	Data subsets are employed in: 1. STM speed and volume data from model predictions for machine learning	Derived

Data ID	EX ID	Dataset Type	Dataset Name	Dataset Description	Dataset Subset Description	Collection Method
25	5	Network Operating Conditions	Network Impedance API	New data exchange to communicate changes in network impedance values for complete paths to the G-MAP app using OSM/OTP data structures. The API will be developed during the agile development cycles in collaboration with STM and G-MAP platform developers.	Data subsets are employed in: 1. Machine learning 2. Performance measurement	Derived
26	1	Assets	Roadway Design and Condition Data	Roadway characteristics typically carried with planning models and operating-characteristic tracking (number of lanes, lane width, speed limit, design capacity, etc.). STM carries all available design elements from each vendor data source for use in machine learning analyses.	Data subsets are employed in: 1. Machine learning 2. Performance measurement	Derived
27	1	Assets	Roadway Intersection Design and Condition Data	Intersection design and operations data for vehicle operations (intersection lane design, bay length, lane-by-lane signal technology and configuration, sensors, timing plans, etc.).	Data subsets are employed in: 1. Machine learning 2. Performance measurement	Derived
29	1	Assets	Pedestrian Intersection Asset Design and Condition Data	Sidewalk ramps, curb cuts, crossings, pedestrian signals, and signage at signalized intersections. Referenced to sidewalk crossing network links and used in impedance calculations. Subsets are generated by asset type for performance reporting and scenario analysis for accessibility improvement scenarios (ramps, curb cuts, crossings, etc.).	Data subsets are employed in: 1. Pedestrian impedance calculations 2. Wayfinding via shortest path	Derived

Data ID	EX ID	Dataset Type	Dataset Name	Dataset Description	Dataset Subset Description	Collection Method
30	13	Assets	Building Pathway Asset Design and Condition Data	Building-interior pathway assets such as door access, thresholds, ramps, push-button activations, signage, etc. Assets (such as exterior doors) reference sidewalk approach links and interior pathway links. These are used in impedance calculations.	Data subsets are employed in: 1. Pedestrian impedance calculations 2. Wayfinding via shortest path	External Input
31	13	Assets	Building Wayfinding Asset Design and Condition Data	The location of wayfinding signs and announcements in facilities including transit hubs and stations. Includes status of current obstructions and vertical-conveyances status (e.g., operating, out of order, under maintenance). Used in impedance calculations and shortest path generation.	Data subsets are employed in: 1. Pedestrian impedance calculations 2. Wayfinding via shortest path	External Input
32	2, 3	Transit, Assets	Transit Stop Asset Design and Condition Data	Bus stop shelters, landing pads, benches, approaches, door access points, ramps, signage, etc. Assets (such as benches) reference sidewalk network links and transit location. These are used in impedance calculations. These features are not currently within GTFS or OSM features but can be used in server-side impedance calculations.	Data subsets are employed in: 1. Pedestrian impedance calculations 2. Wayfinding via shortest path	External Input
33	2, 3	Transit	Transit Vehicle Asset Design and Condition Data	Information about the accessibility of specific transit vehicles (lift presence/ configuration/ design, lift operational status, etc.) for which realtime automatic vehicle location (AVL) data are employed. These features are not currently a GTFS or OSM feature, but will be used in server-side impedance calculations.	Data subsets are employed in: 1. Pedestrian impedance calculations 2. Wayfinding via shortest path	External Input

Data ID	EX ID	Dataset Type	Dataset Name	Dataset Description	Dataset Subset Description	Collection Method
34	2, 3	Transit	GTFS (Ride Gwinnett)	General Transit Feed Specification data files, including accessibility attributes for Ride Gwinnett.	Data subsets are employed in: 1. Wayfinding for G-MAP 2. Development of the TransitSim network for multi-modal impedance calculations	External Input
35	2, 3	Transit	GTFS (MARTA)	General Transit Feed Specification data files, including accessibility attributes for MARTA.	Data subsets are employed in: 1. Wayfinding for G-MAP 2. Development of the TransitSim network for multi-modal impedance calculations	External Input
36	2, 3	Transit	GTFS Realtime (Ride Gwinnett)	GTFS-RT API for Ride Gwinnett. Data indicating transit vehicle location in realtime, event data, and transit vehicle arrival and departure times from a stop.	Data subsets are employed in: 1. Wayfinding for G-MAP 2. Development of the TransitSim network for multi-modal impedance calculations 3. Performance measurement	External Input
37	2, 3	Transit	GTFS Realtime (MARTA)	GTFS-RT API for MARTA transit service. Data indicating transit vehicle location in realtime, event data, and transit vehicle arrival and departure times from a stop.	Data subsets are employed in: 1. Wayfinding for G-MAP 2. Development of the TransitSim network for multi-modal impedance calculations	External Input

Data ID	EX ID	Dataset Type	Dataset Name	Dataset Description	Dataset Subset Description	Collection Method
39	6	CV	BSM	The basic safety message (BSM) is used in a variety of applications to exchange safety data regarding vehicle state and location. The BSM data will be used in this application to enhance the network operations state information in the STM.	Data subsets are employed in: 1. STM traffic operations state updates 2. Machine learning 3. Mobility performance measure computations	External Input
44	8	CV	Ped-X	A series of messages associated with pedestrian signal control including change interval, clearance time, phase, and walk interval. Archived Ped-X messages will be used in the STM for performance monitoring.	Data subsets are employed in: 1. Mobility performance measure computations 2. Performance measurement	Collect/ Forward
45	12, 17	Mobility Service API	Trip options	Trip options calculated from OTP Routing Engine. Calculated itinerary results when inputting an origin and destination in the OTP engine based on the personalized OSM network.	Data subsets are employed in: 1. Benchmark performance assessment 2. Trip destination and purpose research 3. Route-adherence research to improve impedance factors	Collect/ Forward

Data ID	EX ID	Dataset Type	Dataset Name	Dataset Description	Dataset Subset Description	Collection Method
46	2	VRU Modes	VRU categories	List of categories and their default edge impedance values. The enumerated list will correspond to the list of disabilities and assistive devices offered in the G-MAP preference menu.	Data subsets are employed in: 1. Benchmark performance assessment by VRU category 2. Trip destination and purpose research by VRU category 3. Route adherence research to improve impedance factors by VRU category	Derived
51	4	System-Customer Performance	Mobile App Logs	G-MAP mobile app log files which include all the trips, trip preferences, and travel results - as well as users' app usage logs - will be forwarded to the STM dynamic data broker for analysis and aggregation into performance measures.	Data subsets are employed in: 1. Benchmark performance assessment 2. Socioeconomic impact assessment research	Collect/ Forward
52	4	System-Customer Performance	Traverse Data	Customer traverse data through the system (in space and time at highest practical resolution) for use in performance assessment (response times, wait times, travel times, etc.) and that can be compared to recommended routes to refine impedance calculations and route recommendations.	Data subsets are employed in: 1. Benchmark performance assessment 2. Socioeconomic impact assessment research	Collect/ Forward
53	4	System-Customer Performance	Trip Feedback Reports	G-MAP trip reports and feedback from app users including survey data from app users.	Data subsets are employed in: 1. Benchmark performance assessment 2. Socioeconomic impact assessment research	Collect/ Forward

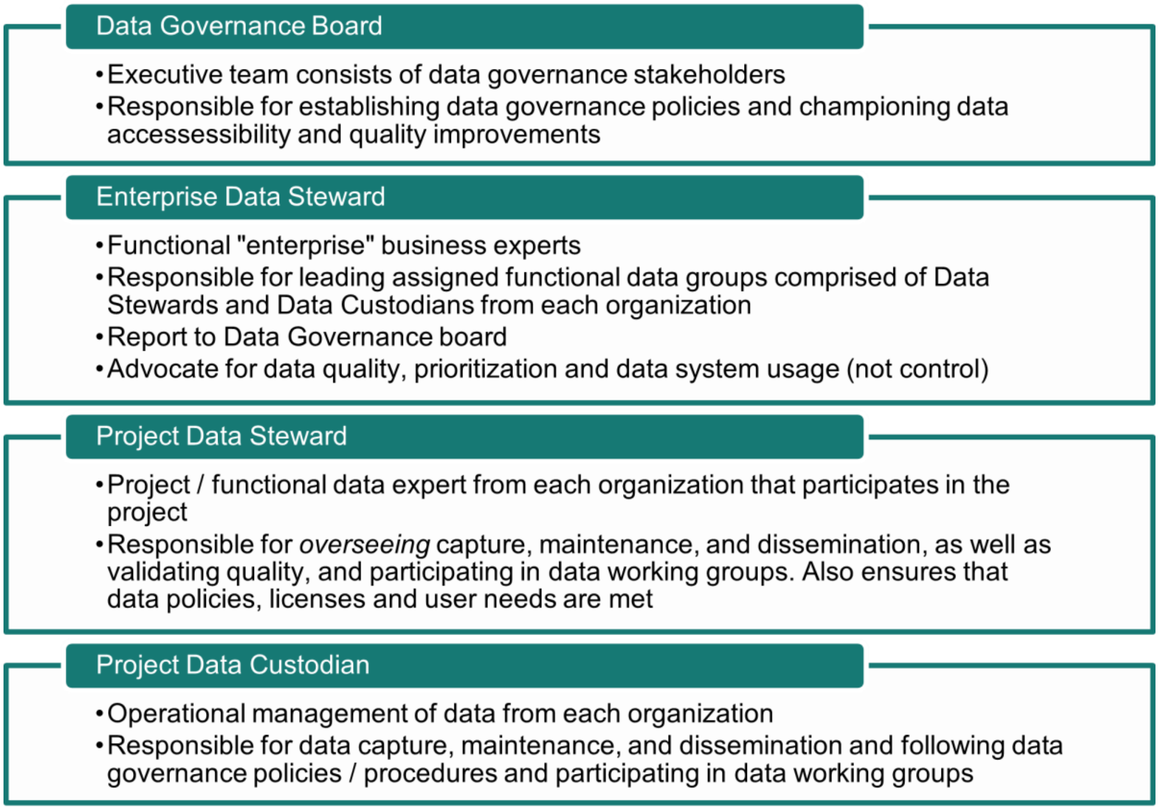
Data ID	EX ID	Dataset Type	Dataset Name	Dataset Description	Dataset Subset Description	Collection Method
58	16	System-Customer Performance	ST-CTN Performance Measures Data	Ongoing random sample data collection conducted through the G-MAP app will gather customer opinion data on system performance. Standardized questions on a Likert scale and open comment fields will be used to collect data.	Data subsets are employed for: 1. Performance assessment. Likert scale values will be collected to gauge changes in individual satisfaction with specific system features and outcomes.	Derived
59	16	STM Performance Logs	STM Communication Logs	The STM will continuously track inbound and outbound communications with timestamps for use in assessing latency.	Data subsets are employed for: 1. Operational performance assessment. Logged timestamps will be used to continuously track and quantify latency along each communications leg.	Derived
60	16	STM Performance Logs	STM Impedance Calculation Logs	The STM will continuously track the time at which impedance calculations begin and are completed to assess computational speed.	Data subsets are employed for: 1. Operational performance assessment. Logged time stamps will be used to continuously track and quantify algorithm speeds.	Derived
64	2	Transit	Ridership: Fixed Route	Transit vehicle ingress and egress counts collected by Ride Gwinnett's automated passenger count (APC) equipment.	Data subsets are employed for: 1. Performance assessment. Pedestrian count data will be used to assess changes in vehicle occupancy and passenger throughput for transit metrics.	External Input

Data ID	EX ID	Dataset Type	Dataset Name	Dataset Description	Dataset Subset Description	Collection Method
65	2	Transit	Ride Gwinnett Complaint Log	Ride Gwinnett maintains an electronic incident log that contains the records of individual passenger complaints that reach the call center.	Data subsets are employed for: 1. Sidewalk asset and impedance calculations. 2. Subset of complaints associated with transit service, routes, stop locations, navigation, and other factors employed in user-related performance metrics. 3. Performance measurement	External Input

2 Data Stewardship

A data governance framework establishes a formal organizational structure on the capture, curation, maintenance, and dissemination of data. Because the ST-CTN project is driven by exchanging and analyzing data for traveler trip generation and execution, data curation, quality, access, privacy, and security, the project will establish a formal data governance organization to manage and oversee integration and access to the data. Governance organizational hierarchy, roles and responsibilities are described in **(adapted from source: ARC Data Governance, 2019)**

Figure 3. This governance organization is described and the schedule for convening and facilitating the Board is described in the [IPFP].



Data Governance Board

- Executive team consists of data governance stakeholders
- Responsible for establishing data governance policies and championing data accessessibility and quality improvements

Enterprise Data Steward

- Functional "enterprise" business experts
- Responsible for leading assigned functional data groups comprised of Data Stewards and Data Custodians from each organization
- Report to Data Governance board
- Advocate for data quality, prioritization and data system usage (not control)

Project Data Steward

- Project / functional data expert from each organization that participates in the project
- Responsible for *overseeing* capture, maintenance, and dissemination, as well as validating quality, and participating in data working groups. Also ensures that data policies, licenses and user needs are met

Project Data Custodian

- Operational management of data from each organization
- Responsible for data capture, maintenance, and dissemination and following data governance policies / procedures and participating in data working groups

(adapted from source: ARC Data Governance, 2019)

Figure 3. Data Governance Roles for the ST-CTN ITS4US Project

In general, the *project data custodians* are subconsultants of public agencies that fund and operate subsystems of the ST-CTN project. The *project data custodians* process, manage, maintain, and validate a dataset that is captured, used or generated by one or more of the subsystems. The *project data steward,* typically staff of a public agency, operates the project and oversees the activities of the data custodian(s). The *enterprise data steward* (EDS) is a committee that addresses crosscutting issues, policies and quality standards needed to ensure an interoperable environment to exchange data. Finally, the *data governance board* is an executive team of public sector stakeholders that champion the data governance policies.

A *data owner* may be a data steward if the data are generated from a system owned by the stakeholder, however, it may be the person about whom the data were captured. For example, a transit agency is both the data owner and steward of GTFS data; this is different from an app user who is the data owner and the app developer is the data steward of a trip trace or account information stored in an account management system. It is then the responsibility of the data steward to enforce policies to protect PII and delete data for which the owner does not wish to be saved. In addition, data may be acquired from a third-party organization, in which case their license agreement governs usage and distribution rights.

The Project Data Steward role assigned to specific datasets is listed in **Table 4.** The Project Data Custodian role is currently under contract to a data steward organization, as listed in **Table 4**.

The EDS and the EDG committee will be held by ARC for the duration of this project (the role will not be listed in **Table 4**). Details of the roles and responsibilities of the EDS and EDG are described in the [IPFP].

The ARC data governance framework, described above, will be implemented in parallel with the BAA/NOFO governance roles described in Error! Reference source not found. As the Federal Sponsor, USDOT will become the owner of the data where the capture, generation or derivation is within the scope of this project. Sponsorship will shift to ARC after the completion of Phase 3.

2.1 Data Owner and Stewardship

Roles are assigned to each dataset that is captured, ingested, generated, used, and disseminated in the ARC ST-CTN project that is consistent with the USDOT-managed Data Store governance model. **Table 4** lists roles associated with each dataset with the following columns:

- **Dataset Identifier (ID)** – unique identifier assigned to each dataset (only full datasets are included) from **Table 3**.
- **Dataset Title** – name of the dataset from **Table 3**.
- **Derived / External –** indicates datasets that are derived (or generated) by the project and not proprietary or subject to data sharing restrictions or external input to the system.
- **Data Owner** – owner of the dataset, organization that licenses or grants access to the dataset. The data owner is the person or organization with the authority, ability, and responsibility to access, create, modify, store, use, share, and protect the data. Data owners have the right to delegate these privileges and responsibilities to other parties. During project testing and initial operations, the dataset owner is the organization designated in Dataset Owner column.

- **Data Steward** – responsible for overseeing the capture, curation, maintenance, and dissemination of the dataset, as well as enforcing the data policies in compliance with license agreements. Specifically, the data steward, at the direction of the data owner, is the person or organization that is delegated the privileges and responsibilities to manage, control, and maintain the quality of a data asset throughout the data lifecycle, including delegating the curation and operations to the data custodian (as described in the governance process). The data steward may also apply appropriate protections, restrictions, and other safeguards depending on the nature of the data, subject to the direction of the data owner.

- **Federal Sponsor** – the project sponsor of the datasets where the capture, generation or ownership is within the scope of this project. The federal sponsor will assume the role of Data Owner once the dataset(s) are provided to them per BAA/NOFO requirements later in the project during Phase 3. The federal sponsor is only responsible for the dataset when it is generated by the project and not subject to data sharing restrictions.

Table 4. Dataset Roles in Data Governance Organization

ID	Dataset Title	Derived / External	Data Owner	Data Steward	Federal Sponsor
3	Whole Road Network	Derived	ARC	ARC/GA Tech	USDOT ITS JPO
5	STM Network	Derived	STM	STM (GA Tech)	USDOT ITS JPO
6	NaviGAtor Network	External	GDOT	GDOT	Out of scope
8	OSM Network	External	OSM	G-MAP and ARC (GA Tech)	Out of scope
10	Sidewalk Network	Derived	ARC	GA Tech	USDOT ITS JPO
11	Indoor Pathways	External	INS Vendor	Gwinnett County	Out of scope
15	NaviGAtor Data	External	GDOT	GDOT	Out of scope
20	Modeled Future Operating Conditions	Derived	STM	ARC/GA Tech	USDOT ITS JPO
25	Network Impedance API	Derived	STM	STM (GA Tech)	USDOT ITS JPO
26	Roadway Design and Condition Data	Derived	STM	STM (GA Tech)	USDOT ITS JPO
27	Roadway Intersection Design and Condition Data	Derived	STM	STM (GA Tech)	USDOT ITS JPO

ID	Dataset Title	Derived / External	Data Owner	Data Steward	Federal Sponsor
29	Pedestrian Intersection Asset Design and Condition Data	Derived	Gwinnett County	Gwinnett County	USDOT ITS JPO
30	Building Pathway Asset Design and Condition Data	External	Gwinnett County	Gwinnett County	Out of scope
31	Building Wayfinding Asset Design and Condition Data	External	Gwinnett County	Gwinnett County	Out of scope
32	Transit Stop Asset Design and Condition Data	Derived	Gwinnett County	Gwinnett County	USDOT ITS JPO
33	Transit Vehicle Asset Design and Condition Data	Derived	Ride Gwinnett	G-MAP	USDOT ITS JPO
34	GTFS (Ride Gwinnett)	External	Ride Gwinnett	G-MAP	Out of scope
35	GTFS (MARTA)	External	MARTA	G-MAP	Out of scope
36	GTFS Realtime (Ride Gwinnett)	External	Ride Gwinnett	G-MAP	Out of scope
37	GTFS Realtime (MARTA)	External	MARTA	G-MAP	Out of scope
39	BSM	External	GDOT/GCDOT	GDOT/ GCDOT	Out of scope
44	Ped-X	External	G-MAP /GCDOT	G-MAP /GCDOT	Out of scope
45	Trip Options	External	G-MAP	G-MAP	Out of scope
46	VRU categories	Derived	ARC	ARC (GA Tech)	USDOT ITS JPO
51	Mobile App Logs	Derived	G-MAP	G-MAP	USDOT ITS JPO
52	Traverse Data	Derived	GA Tech	GA Tech	USDOT ITS JPO
53	Trip Feedback Reports	Derived	ARC	ARC	USDOT ITS JPO

ID	Dataset Title	Derived / External	Data Owner	Data Steward	Federal Sponsor
58	ST-CTN Performance Measures Data	Derived	G-MAP / GA Tech	GA Tech	USDOT ITS JPO
59	STM Communication Logs	Derived	G-MAP / GA Tech	GA Tech	USDOT ITS JPO
60	STM Impedance Calculation Logs	Derived	G-MAP / GA Tech	GA Tech	USDOT ITS JPO
64	Ridership: Fixed Route	External	Ride Gwinnett	Ride Gwinnett	Out of scope
65	Ride Gwinnett Complaint Log	External	Ride Gwinnett	Ride Gwinnett	Out of scope

2.2 Data Storage and Retention

This section of the DMP identifies the servers and the systems that will be used to manage, store, and maintain data for the project. The sections that follow provide details for each dataset employed in the project about where the data will be hosted (system type), how data will be managed, how long data will be retained, what data storage and retention policies will be implemented, and how security provisions will be managed.

When these external datasets arrive, the system reviews and validates the quality of the data prior to its usage. Many of the datasets are used and discarded unless the source system retains the dataset; when relevant, the project team will archive the dataset after its use and retain the data through the project period. Although external data storage and retention provisions are out of scope of this project, the project team will review the retention provisions to ensure that key datasets that are needed for performance measurements are retained in project data storage systems and made available to USDOT and the IE.

2.2.1 Storage Systems

The storage systems that are used in the project among the three subsystems are identified in **Table 5**. The datasets are described by the following columns:

- **Data Storage System Name**: the name of the data storage system the dataset will be stored in. Only one data system in the project system will be identified as the single source of primary data.

- **Data Storage System Subsystem and Status**: denotes the subsystem steward organization of the data storage system and status of the dataset. Status values include: under development, implemented, operational. Note: operational implies that the data store may be implemented and operational in another instance of the system. In most cases, the data store will be augmented to include new data elements generated from this project.

- **Dataset IDs and Title(s)**: list of Dataset IDs and Titles contained in the data store.
- **Initial Storage Date**: The initial date that data will be available in each data storage system.
- **Frequency of Update**: Describes how frequently the data will be updated in the data storage system once ingestion begins (i.e., "Continually," "Daily," "Weekly," "Monthly," "Annually," "Unknown," "As needed," "Irregular," or "None planned").
- **Archiving and Preservation Period**: The duration for which the dataset will be maintained in each data storage system.

Table 5. Storage Systems

Data Storage System Name	Data Storage System Subsystem	Dataset ID and Title	Initial Storage Date	Frequency of Update	Archiving and Preservation Period
G-MAP Connected Data Platform Module	G-MAP; operational	44. Summary of PED-X transactions 45. Trip Options (raw) 51. Mobile App Log (raw) 53. Trip Feedback Reports (raw)	Phase 2, April 2024	Continuous	Through project period (see [G-MAP DMP])
Open Data Server	Public PMD; operational expected in Phase 2, January 2024	10. Sidewalk Network	Phase 2, April 2024	Monthly	Through project period
Open Data Server	STM; operational expected in Phase 2, January 2024 Note: although most of the datasets originate in the GDOT/GCDOT ITS Data Hub, they are not archived.	3. Whole Road Network 8. OSM Network 46. VRU categories	Phase 2, April 2024	As Needed	Through project period
PII Server Processed Data on Research Server	STM; operational expected in Phase 2, January 2024	51. Mobile App Logs 52. Traverse Data 53. Trip Feedback Reports 58. ST-CTN Performance Measures Data 59. STM Communication Logs 60. STM Impedance Calculation Logs 65. Ride Gwinnett Complaint Log	Phase 2, April 2024	Daily	Through project period

Data Storage System Name	Data Storage System Subsystem	Dataset ID and Title	Initial Storage Date	Frequency of Update	Archiving and Preservation Period
Research Server	STM (subset on STM Cluster); operational expected in Phase 2, January 2024	11. Indoor Pathways 15. NaviGAtor Data 20. Modeled Future Operating Conditions 26. Roadway Design and Condition Data 27. Roadway Intersection Design and Condition Data 29. Pedestrian Intersection and Pathway Asset Design and Condition Data 30. Building Pathway Asset Design and Condition Data 31. Building Wayfinding Asset Design and Condition Data 32. Transit Stop Asset Design and Condition Data 33. Transit Vehicle Asset Design and Condition Data 64. Ride Gwinnett Ridership Data 39. BSM 51. Mobile App Logs (processed) 53. Trip Feedback Reports (processed)	Phase 2, April 2024	Daily	through project period and/or compliant with data provider archiving policy
STM Server Cluster	STM; operational expected in Phase 2, October 2023	5. STM Network	Phase 2, January 2024	As Needed	through project period
STM Server Cluster	STM; operational expected in Phase 2, October 2023	25. Network Impedance API	Phase 2, October 2023	Continuous	through project period

Data Storage System Name	Data Storage System Subsystem	Dataset ID and Title	Initial Storage Date	Frequency of Update	Archiving and Preservation Period
TRANSIT-data-tools	G-MAP; operational	34. GTFS (Ride Gwinnett) 35. GTFS (MARTA) 36. GTFS Realtime (Ride Gwinnett) 37. GTFS Realtime (MARTA)	Phase 2, April 2024	As Needed	Compliant with ATL and ARC archiving and retention policy; configurable for each dataset (see [G-MAP DMP])

2.2.2 Data Storage System Description

For this project, data storage and retention procedures are required to host and manage static network and demographic data (traditional file storage), user interface systems and the space-time memory (database and cloud database systems), and to handle PII and research activities (dedicated secure research servers). To implement these systems, the ST-CTN project team will manage a mix of servers for this project:

- Private cloud servers will be employed in the provision of user interface systems.
- A server farm will host the space-time memory systems.
- Open-source servers will handle open-source data for operational systems, public access and to push data to the U.S. DOT-managed Public System (https://its.dot.gov/data/).
- Third-party servers will feed data from outside sources into the project systems.
- Research servers accessible only to the project team will be used in system development, development of APIs and key performance indicators (KPIs), and management of PII.

The data storage systems will be designed, tested, and deployed in Phase 2. The G-MAP and STM operational data storage systems are currently deployed or will be deployed prior to collecting the datasets.

TRANSIT-data-tools is a G-MAP data store, with managed and controlled access by the IBI Group (data custodians). The data storage is configurable for each GTFS dataset. Realtime transit information is not archived or retained (it is available from Ride Gwinnett upon request).

G-MAP Connected Data Platform Module is a G-MAP data store, with managed and controlled access by the IBI Group (data custodians). The Data Platform is currently deployed since it was included in the first G-MAP sprint. Additional fields will be added as new features are incorporated into G-MAP through the development period.

Open Data Server will be an open data portal where the public may access the data (available data is dependent on data license and use requirements). The design, structure, and user interfaces for the open data portal is estimated to be deployed in Phase 2, April 2024. The Open Data Server will be hosted in the cloud through ARC's Microsoft Power BI subscription, and will provide public access to open data, and provide project data to the U.S. DOT-managed Public System. The data will be submitted to U.S. DOT via the DataHub which is a web-based system designed for storing and publicly sharing open data generated by ITS JPO funded projects. Associated Metadata will be provided per the requirements of the site (https://its.dot.gov/data/datahub-submission/). Design provisions for the method of sharing the data will be provided in the SDD.

STM Hosted Data Stores: The realtime data and historic data archives implemented as part of the STM will be hosted on a Georgia Tech server farm dedicated to this project. These servers follow security and privacy protocols outlined by the Institute (Information Security Procedures, Standards, and Forms – Georgia Tech Cyber Security (gatech.edu)). These servers manage operational, private, and personally identifiable datasets. Specific data stores include the following:

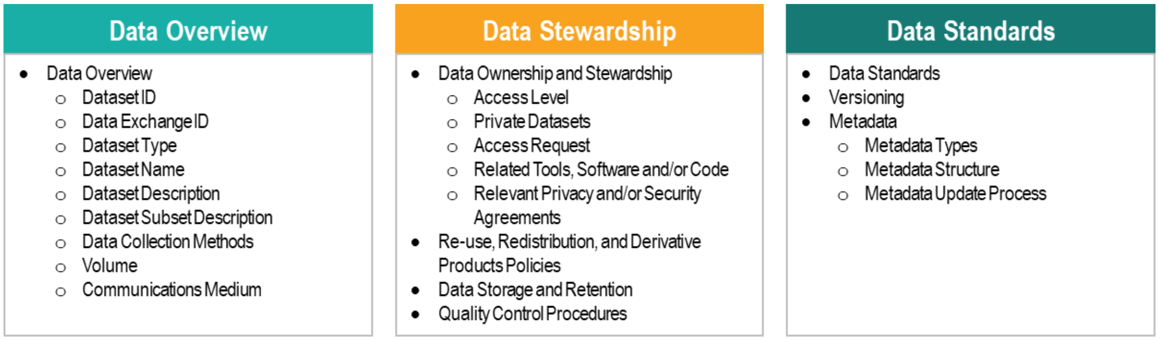
- **STM Server Cluster:** A cluster of compute and storage servers that store and preprocess historic operation and mobility information and translates them into information. The STM ties in with the Partnership for an Advanced Computing Environment (PACE) cluster at Georgia Tech to leverage high-performance computing for processing the archival data while it uses the STM cluster resources for integrating realtime information with the historical conditions.

- **Research Server:** Multiple high-resource computers (e.g., 96 cores, 512 GB RAM) are used for the development, training and execution of the Machine Learning processes.

- **PII Server:** Separate resources are maintained for handling PII data at Georgia Tech. The PII servers reside on a separate secure network that restricts access to other networks even within the GA Tech Civil Engineering network. The access to the server requires physical presence in the secure access lab (badge-access controlled) in the GA Tech Civil Engineering building.

2.3 Data Sharing Framework

The Data Sharing Framework applies specialized rules for sharing information across the data components -- overview, stewardship, and standards depicted in **Figure 4**. The framework consists of appropriately prepared system control, performance, and evaluation data, stripped of PII, that will be made available to the USDOT and posted in a timely fashion on public-facing resources, freely available to the public and research community.



Data Overview

- Data Overview
  - Dataset ID
  - Data Exchange ID
  - Dataset Type
  - Dataset Name
  - Dataset Description
  - Dataset Subset Description
  - Data Collection Methods
  - Volume
  - Communications Medium

Data Stewardship

- Data Ownership and Stewardship
  - Access Level
  - Private Datasets
  - Access Request
  - Related Tools, Software and/or Code
  - Relevant Privacy and/or Security Agreements
- Re-use, Redistribution, and Derivative Products Policies
- Data Storage and Retention
- Quality Control Procedures

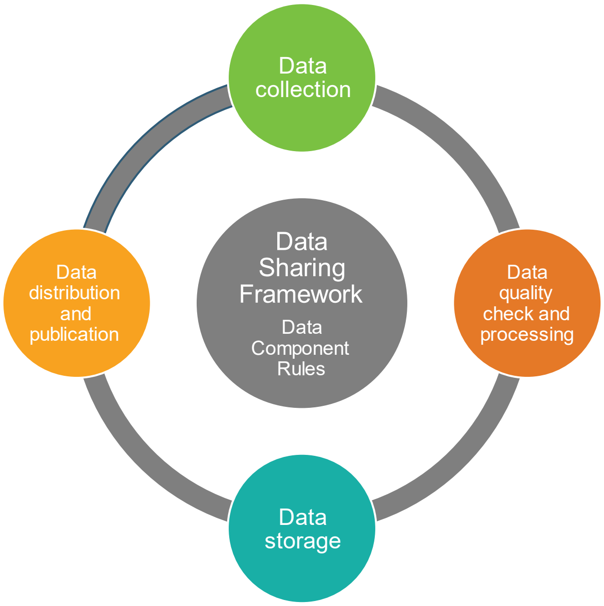
Data Standards

- Data Standards
- Versioning
- Metadata
  - Metadata Types
  - Metadata Structure
  - Metadata Update Process

Source: ARC, 2022

Figure 4. Data Sharing Framework

Appropriate curation processes will be applied to specific datasets to move data from collection to distribution including data quality checking and processing (e.g., anonymization), storage and metadata management, distribution, and publication functions. Figure 5 provides a visual representation of the data sharing framework where data collection, quality control, storage, and distribution/publication are governed by data component rules.



Source: ARC, 2021

Figure 5. Curation Processes in Data Sharing Processes

The data component rules implemented in the data curation process (**Source:** ARC, 2021

Figure 5) are discussed throughout this plan including data overview (**Section 1.4.2**); Data Stewardship including quality control (**Section 2**); and Standards, Versioning and Metadata Management (**Section 3**).

2.4 Data Quality Control

As each derived public dataset is collected and published, the data steward will generate a data curation plan as specified in the SyRS (see **Req ID 6.1.0-148**) and documentation of associated data quality processes (see **Req ID 6.1.0-149**). The data curation plan will document "data quality processes and control including data collection (ingestion, acquisition), verification and quality checking, storage (extraction, transformation, loading), and distribution procedures." The data curation plans will document data quality processes for each derived dataset that includes "each step from acquisition through storing, providing access and deprecating." As noted in **Section 3.3**, the curation processes - including data quality and control - will be documented in the metadata when made available through the Public PMD.

Several acquired datasets will also include data curation plans. These include:

- Whole Road Network Composition (1.3.0-035)
- Traveler Feedback Data (6.1.0-151)
- Transit Data (GTFS static / realtime) (6.1.0-154)
- Signal Control Data (GC DOT TCC) (6.1.0-157-158)

The initial data curation approach and extraction-transformation-load (ETL) design will be included in the System Design Document (SDD) and the operational data curation procedures will be included in the Comprehensive Maintenance and Operations Plan (COMP).

When data curation plans are developed, the following data quality elements will be considered:

- Accuracy – the data are required to be accurate.
- Relevancy – the data are required to serve the intended use.
- Completeness – the data are required to be complete.
- Timeliness – each dataset will need to be current as specified within system requirements.
- Consistency – the data are required to have consistent formatting as described within the metadata.

2.5 Privacy

The ST-CTN project includes many datasets used in a wide variety of applications. Maintaining privacy and security of the data is paramount. Each dataset is assigned a data owner and data steward, as described in **Section** Data Owner and Stewardship**2.1**, who is tasked with maintaining the accessibility and integrity of the data. Data access is categorized as either open (i.e., publicly available data) or private. Private datasets contain data that may not be shared with external users due to operational, proprietary, research, or PII restrictions. Operational, proprietary, and research datasets are often able to be shared for project related purposes including performance monitoring or modeling if data licensing agreements are respected.

Dataset access levels are differentiated by proprietary data, which are restricted by license or usage agreements, versus protected datasets, that require access controls to preserve privacy and security. This section describes access levels at four levels. Data access levels are defined as:

- **Open** – Data that can be used by the public with no or limited licensing restrictions. These data are available to the public without needing to request permissions and will be provided to the USDOT-managed Public System. These can include anonymized or aggregated versions of private datasets, where the anonymization and aggregation processes remove any PII from the data.

- **Private** – Data that cannot be shared with external users. Access to these data is limited and only granted with project team approvals. Some subcategories require IRB approval for access in addition to project team approval. Subcategories:

  - **Operational** – Data used in realtime applications and operations of the system. The data may contain licensed data restricted by usage agreements.

  - **Proprietary –** Licensed data from third parties or commercial business interests (CBI). These data may be used for planning or operational purposes. Any access to the data is determined by usage agreements between the parties.

  - **Research** – Data that are available for research for users that meet requirements. Users of the data and their associated access protocols must meet IRB requirements before gaining access to the data. These datasets may originally have had PII, but have been sanitized and compiled to remove PII before use in machine learning and research purposes that do not require PII. However, since the data may retain trace elements and some data may be proprietary, the data is subject to restrictions even after it has been sanitized.

  - **PII Certification** – Data that have PII included in the dataset. The access to this data is as restrictive as possible to protect the PII based on IRB-approved processes. Data in this category should have an operational purpose that justifies its storage.

Raw data are stored and archived to ensure data integrity. Data used in active implementation of the system will appear in online working datasets that are updated and queried in realtime.

PII data protection is governed by GA Tech secure sever systems data management protocols:

- Transmission of encrypted PII data to the GA Tech interface server.

- Decryption of data on the GA Tech Secure Server.

- Elimination of remote access to the Secure Server.

- Direct hard-wire connections between the Secure Server and the data analysis terminals located in the Secure Data Lab on the first floor of the SEB Building.

- Secure Data Lab door security that requires cardkey access.

- Login ID and password protection to terminals and the secure server.

- Layered protection within the secure server data storage system to limit access to different datasets to specific users.

Access to protected data is governed by non-disclosure agreement (NDA) and approved IRB protocols:

- IRB approval is required for any access to PII data (amendments may be submitted to facilitate supplemental data access and uses while ensuring continued privacy protection).

- Individual access to the data requires card key access to the Secure Data Lab, login ID and password access to the Secure Data Lab terminals, and pre-approved login ID and password access to the secure server, login ID and password access to each data folder.

- Access to PII data can only be accomplished through dedicated direct-connect terminals located in the Secure Data Lab on the first floor of the Sustainable Education Building (no remote access).

2.5.1 Private Datasets

The restriction of datasets occurs on two levels: *proprietary* datasets constrained by license agreements and *protected* datasets that are restricted because they contain PII. **Table 6** describes the classification of every dataset as Protected, Proprietary, or Open and identifies the basis for private data classification. All proprietary data are stored on the secure server and working datasets are generated to which access is limited to project-specific modeling and data processing routines. Licensed data may be made available by the data owners to third parties. Most agencies that limit access to their data require the execution of a data user agreement to ensure data integrity and track data usage (at no fee). Data licensed from private companies generally contains intellectual property and commercial value. Access to these data typically requires the execution of a data license and fees are likely required to obtain data access. Use of private data is minimized in this project, with the focus being on the use of open data.

The datasets are described by the following columns:

- **Dataset Identifier (ID):** unique identifier related to every dataset or dataset subset.
- **Dataset Name:** the name or title for the dataset.
- **Access Levels:** describes how individuals may gain access or how the datasets may be used. Values include: PII Certified, Research, Proprietary, Operational and Open.
- **Reason(s) Data are Private:** describes why the data are designated as proprietary or protected. General reasons are stated above.
- **Safeguarding Methods and Processes:** describes the method used to secure and gain access to data.

Table 6. List of Privacy and Access Levels Assigned to Datasets

ID	Dataset Name	Access Level	Reason(s) Data are Private	Safeguarding Methods and Processes
3	Whole Road Network	Open	NA	NA
5	STM Network	Open	NA	NA
6	NaviGAtor Network	Open	NA	NA
8	OSM Network	Open	NA	NA
10	Sidewalk Network	Open	NA	NA
11	Indoor Pathways	Open	NA	NA

ID	Dataset Name	Access Level	Reason(s) Data are Private	Safeguarding Methods and Processes
15	NaviGAtor Data	Open	Subject to GDOT usage agreement.	Although data are open, access to data is restricted by data usage agreement with GDOT. Third-parties typically must execute a user agreement with the data owner to access data.
20	Modeled Future Operating Conditions	Research / Proprietary	Derived from proprietary data sources.	Access to licensed data may be granted by the data owner. Third parties typically must execute a user agreement with the data owner to access data.
25	Network Impedance API	Operational; Open	NA	NA
26	Roadway Design and Condition Data	Open	NA	NA
27	Roadway Intersection Design and Condition Data	Open	NA	NA
29	Pedestrian Intersection Asset Design and Condition Data	Operational	Subject to agency data license and usage restrictions	NA
30	Building Pathway Asset Design and Condition Data	Open	NA	NA
31	Building Wayfinding Asset Design and Condition Data	Open	NA	NA
32	Transit Stop Asset Design and Condition Data	Open	NA	NA

ID	Dataset Name	Access Level	Reason(s) Data are Private	Safeguarding Methods and Processes
33	Transit Vehicle Asset Design and Condition Data	Operational	Subject to agency data license and usage restrictions	Access to licensed data may be granted by the data owner. Third parties typically must execute a user agreement with the data owner to access data.
34	GTFS (Ride Gwinnett)	Open	NA	NA
35	GTFS (MARTA)	Open	NA	NA
36	GTFS Realtime (Ride Gwinnett)	Open	NA	NA
37	GTFS Realtime (MARTA)	Open	NA	NA
39	BSM	Operational	NA	NA
44	Ped-X	Operational	NA	NA
45	Trip options	Operational (realtime, transmitted, deleted), PII Certification (raw data)	Customer names and other traveler destinations such as home/work locations, and daycare/school locations, all constitute PII. Turn-by-turn directions reveal address locations.	Data protection governed by Georgia Tech secure sever systems data management protocols. Access conditions governed by NDA and approved IRB protocols.
46	VRU categories	Open	NA	NA

ID	Dataset Name	Access Level	Reason(s) Data are Private	Safeguarding Methods and Processes
51	Mobile App Logs	PII Certification (raw data); Research (when PII is removed)	Customer names and other sensitive trip-related data (e.g., home/work locations, healthcare visits and daycare/school locations), all constitute PII. Trip logs reveal address locations.	Data protection governed by GA Tech secure sever systems data management protocols. Access conditions governed by NDA and approved IRB protocols.
52	Traverse Data	PII Certification	Customer names, home/work locations, and daycare/school locations, all constitute PII. Traverse data reveal address locations.	Data protection governed by GA Tech secure sever systems data management protocols. Access conditions governed by NDA and approved IRB protocols.
53	Trip Feedback Reports	PII Certification (raw data); Open (when PII is removed)	Customer names, home/work locations, and daycare/school locations, all constitute PII. Trip logs reveal address locations.	Data protection governed by GA Tech secure sever systems data management protocols. Access conditions governed by NDA and approved IRB protocols.
58	ST-CTN Performance Measures Data	Open Data	NA	NA
59	STM Communication Logs	Open Data	NA	NA
60	STM Impedance Calculation Logs	Open Data	NA	NA
64	Ridership: Fixed Route	Open Data	NA	NA

ID	Dataset Name	Access Level	Reason(s) Data are Private	Safeguarding Methods and Processes
65	Ride Gwinnett Complaint Log	Research	The data contain data that must be secured.	Data protection governed by GA Tech secure sever systems data management protocols. Access conditions governed by NDA and approved IRB protocols.

The Data Privacy Plan (DPP) can be referenced for a full description of the privacy and security measures of the ST-CTN project to protect user privacy and mitigate risk of threat to users, data privacy, and system hardware. The DPP includes information on access requirements, risk assessment, and security policies and procedures.

2.6 Relationship to Performance Measures

The datasets identified within this document will support the measurement of the ST-CTN system performance as is fully described in the Phase 2 Performance Measurement and Evaluation Support Plan (PMESP). Performance measures are organized by use case: complete trip (use case 1) and connected vehicle (use case 2). Complete trip and CV performance measures are used to evaluate the success of the ST-CTN system from the perspective of the end user (traveler) and focus on the user experience, safety, mobility, and accessibility. Greater impacts to the community, population, and system may be realized over time and deployment expanded; however, initial deployment and evaluation is anticipated to be primarily focused on the individual end user.

A summary of each performance measure for the ST-CTN system is provided below. **Table 7** provides the associated datasets, data owner, steward, and custodian for each measure. Each performance measure and associated metrics have been given a unique identifier with the following nomenclature:

AB-CD-E. F, where:

- AB = Performance measurement category
    - CT = Complete Trip (Use Case 1)
    - CV = Connected Vehicle (Use Case 2)
- CD = Measure or metric
    - PM = Performance Measure
    - ME = Metric
- E = Performance Measure ID
- F = Metric ID

CT-PM-1 Enhance Traveler Experience. This performance measure will assess the user's complete trip travel experience while using the ST-CTN system. The measure will be used to evaluate the system's achievement of Goal 1 (described further in PMESP) which is to improve the traveler's experience throughout their complete trip. Travel experience surveys, unique user log-ins, anonymized user data, and Ride Gwinnett complaint logs will be used to understand the impact of the ST-CTN system on the travelers' complete trip travel experience.

CT-PM-2 Improve Accessibility. This performance measure will evaluate the impact that the ST-CTN system has on the traveler's independence and ability to access employment and other types of trips with use of the ST-CTN system. Trip types such as employment, education, shopping, leisure activities, and other essential trips will be considered. The measure will evaluate the system's achievement of Goal 1 (described further in PMESP) which is to improve the traveler's experience throughout their complete trip. In addition, this measure will evaluate the system's achievement of Goal 4 (described further in PMESP) which is to improve the traveler's accessibility. Travel experience surveys will be used to understand the impact of the ST-CTN system on the travelers' accessibility.

CT-PM-3 Enhance Complete Trip Pedestrian Safety. This performance measure will evaluate the impact of the ST-CTN system to enhance pedestrian safety and driver awareness. Analysis of pedestrian incident data and pedestrian routing within the project area will be done to evaluate the system. Pedestrian incident data tend to be extremely sparse and unreliable. The ST-CTN project team intends to collect these data throughout the life of the project, however, statistically significant analysis is not expected to be possible during the timeframe of this project and there is not a specific metric related to reported pedestrian incidents. This performance measure evaluates Goal 2 (described further in PMESP) which is to improve safety and increase awareness for ST-CTN system users. Travel surveys and crash data will be analyzed within the study area.

CT-PM-4 Enhance Fixed-Route Transit. This performance measure will evaluate the impact of the ST-CTN system to transit ridership. The purpose of this measurement is to collect and analyze data about fixed-route transit ridership to determine if there is a mode shift due to the ST-CTN system. A mode shift from paratransit to fixed-route service with the use of the ST-CTN system could demonstrate a traveler's improved experience and increased safety, reliability, mobility, and/or accessibility. This performance measure evaluates Goal 3 (described further in PMESP) which is to improve the traveler's experience throughout their complete trip and improve safety, reliability, mobility, and accessibility for ST-CTN system users. Ride Gwinnett ridership data will be analyzed within the study area.

CV-PM-1 Enhance Safety and Awareness with Connected Vehicles. This performance measure will evaluate the impact of the ST-CTN system to pedestrian safety and vehicular awareness. Analysis of pedestrian signalized intersection crossing times and broadcasted CV PSM messages will support the evaluation. This performance measure evaluates Goal 2 (described further in PMESP) which is to improve safety and awareness for ST-CTN system users.

CV-PM-2 Improve Transit Reliability. This performance measure will assess transit data to determine the impact of ST-CTN on transit reliability. The purpose of this measure is to assess changes in transit service performance over time. This performance measure evaluates Goals 3 and 4 (described further in PMESP) which are to improve reliability, mobility, and accessibility for ST-CTN system users.

Table 7. ST-CTN Performance Measures and Associated Datasets

PM ID	DMP Data ID	Dataset Name	Data Owner	Data Steward	Data Custodian
CT-PM-1					
	25	Network Impedance API	STM	STM (GA Tech)	GA Tech
	26	Roadway Design and Condition Data	STM	STM (GA Tech)	GA Tech
	27	Roadway Intersection Design and Condition Data	STM	STM (GA Tech)	GA Tech
	46	VRU Categories	ARC	ARC (GA Tech)	GA Tech
	51	Mobile App Logs	ARC	ARC (IBI)	IBI
	52	Traverse Data	GA Tech	GA Tech	IBI/GA Tech
	53	Trip Feedback Reports	ARC	ARC	IBI/GA Tech
	58	ST-CTN Performance Measures Data	G-MAP / GA Tech	GA Tech	GA Tech
	59	STM Communication Logs	G-MAP / GA Tech	GA Tech	GA Tech
	60	STM Impedance Calculation Logs	G-MAP / GA Tech	GA Tech	GA Tech
	65	Ride Gwinnett Complaint Log (Baseline)	Ride Gwinnett	Ride Gwinnett	Ride Gwinnett
CT-PM-2					
	46	VRU Categories	ARC	ARC (GA Tech)	GA Tech
	53	Trip Feedback Reports	ARC	ARC	IBI/GA Tech
	58	ST-CTN Performance Measures Data	G-MAP / GA Tech	GA Tech	GA Tech
CT-PM-3					
	25	Network Impedance API	STM	STM (GA Tech)	GA Tech
	26	Roadway Design and Condition Data	STM	STM (GA Tech)	GA Tech

PM ID	DMP Data ID	Dataset Name	Data Owner	Data Steward	Data Custodian
	27	Roadway Intersection Design and Condition Data	STM	STM (GA Tech)	GA Tech
	46	VRU Categories	ARC	ARC (GA Tech)	GA Tech
	52	Traverse Data	GA Tech	GA Tech	IBI/GA Tech
	53	Trip Feedback Reports	ARC	ARC	IBI/GA Tech
	58	ST-CTN Performance Measures Data	G-MAP / GA Tech	GA Tech	GA Tech
CT-PM-4					
	25	Network Impedance API	STM	STM (GA Tech)	GA Tech
	26	Roadway Design and Condition Data	STM	STM (GA Tech)	GA Tech
	27	Roadway Intersection Design and Condition Data	STM	STM (GA Tech)	GA Tech
	46	VRU Categories	ARC	ARC (GA Tech)	GA Tech
	52	Traverse Data	GA Tech	GA Tech	IBI/GA Tech
	53	Trip Feedback Reports	ARC	ARC	IBI/GA Tech
	58	ST-CTN Performance Measures Data	G-MAP / GA Tech	GA Tech	GA Tech
	64	Fixed-Route Transit Ridership (Baseline)	Ride Gwinnett	Ride Gwinnett	Ride Gwinnett
CV-PM-1					
	15	NaviGAtor Data	GDOT	GDOT	GA Tech
	25	Network Impedance API	STM	STM (GA Tech)	GA Tech
	25	Network Impedance API	STM	STM (GA Tech)	GA Tech
	26	Roadway Design and Condition Data	STM	STM (GA Tech)	GA Tech
	39	BSM	GDOT/GCDOT	GDOT/GCDOT	GA Tech
	44	Ped-X	ARC /GCDOT	ARC /GCDOT	IBI/GA Tech

PM ID	DMP Data ID	Dataset Name	Data Owner	Data Steward	Data Custodian
	46	VRU Categories	ARC	ARC (GA Tech)	GA Tech
	52	Traverse Data	GA Tech	GA Tech	IBI/GA Tech
	53	Trip Feedback Reports	ARC	ARC	IBI/GA Tech
	58	ST-CTN Performance Measures Data	G-MAP / GA Tech	GA Tech	GA Tech
CV-PM-2					
	36	GTFS Realtime (Ride Gwinnett) (Baseline)	Ride Gwinnett	ARC (IBI)	IBI
	46	VRU Categories	ARC	ARC (GA Tech)	GA Tech
	51	Mobile App Logs	ARC	ARC (IBI)	IBI
	53	Trip Feedback Reports	ARC	ARC	IBI/GA Tech
	58	ST-CTN Performance Measures Data	G-MAP / GA Tech	GA Tech	GA Tech

2.6.1 Baseline Data

Baseline data will be limited for the use of measuring ST-CTN system performance. The ST-CTN project team will rely heavily on survey and time series methods to evaluate the ST-CTN system. Baseline data will be collected to establish the conditions in the project area prior to the implementation of the system, as well as be used to evaluate changes in response to system implementation. **Table 8** provides a summary of baseline data that will be collected, associated performance measures, and how the baseline data will be used.

Table 8. ST-CTN Baseline Data for Performance Measures

DMP Data ID	Dataset Name	PM ID	Baseline Data Use
36	GTFS Realtime (Ride Gwinnett) (Baseline)	CV-PM-2 Improve Transit Reliability	GTFS Realtime (Ride Gwinnett) data will be used to establish a baseline On-Time Performance (OTP). Following the implementation of the ST-CTN system, OTP will be assessed to see if the system increased the reliability of Gwinnett County fixed-route transit with enhanced TSP.

DMP Data ID	Dataset Name	PM ID	Baseline Data Use
64	Fixed-Route Transit Ridership (Baseline)	CT-PM-4 Enhance Fixed Route Transit	Ridership data will be used to establish a baseline of fixed-route transit use within the network to determine if the implementation of the system increased ridership.
65	Ride Gwinnett Complaint Log (Baseline)	CT-PM-1 Enhance Traveler Experience	Complaint logs will be reviewed to determine if the ST-CTN system reduced the number of complaints throughout the Ride Gwinnett network.

3 Data Standards

3.1 Data Standards

Each dataset is specified and encoded using an open, standard or proprietary format. This section identifies the standards used to specify each dataset. **Table 9** lists the dataset, specified standard, and the rationale for using the standard.

Columns listed in **Table 9** include:

- **Dataset Identifier (ID):** unique ID that corresponds to **Table 3**.
- **Dataset Name**: dataset title that corresponds to **Table 3** dataset name.
- **Data Standard (Standard ID)**: the name(s) of the data standard(s) in which the data are made available to the USDOT. The data standard is notated with a Standard ID that references entries in the **Table 12. Comprehensive List** of Standards (**Table 12.** Comprehensive List of Standards). **Appendix B** contains the details of the standard, standard profile, (PICS, or ICD used to describe the data) and on-line reference to obtain the standard.
- **Data Standard Rationale**: explanation for use of the chosen data standard and preferred format.

Table 9. Dataset Standards and Rationale

ID	Dataset Title	Data Standard(s) Name	Data Standard Rationale
8	OSM Network	OSM	Required for use with the OTP
12	GTFS Transit Network	GTFS + other formats	Industry standard for exchanging transit schedule data
25	Network Impedance API	OSM	Required for use with the OTP and becoming an industry standard for exchanging attribute and feature information on ROW and PROW
34	GTFS (Ride Gwinnett)	GTFS	Industry standard for exchanging transit schedule data
35	GTFS (MARTA)	GTFS	Industry standard for exchanging transit schedule data

ID	Dataset Title	Data Standard(s) Name	Data Standard Rationale
36	GTFS Realtime (Ride Gwinnett)	GTFS Realtime	Industry standard for exchanging transit event, realtime, and arrival prediction data
37	GTFS Realtime (MARTA)	GTFS Realtime	Industry standard for exchanging transit event, realtime and arrival prediction data
39	PSM	SAE J2735 + J2945 Part 9	Industry standard for CV message sets and produced by existing GDOT deployment

A comprehensive list of data standards and specifications are listed in **Appendix B**. The Data Standard names refer to the Standard Name in **Table 12**. Additional information about encoding, security/encryption, and communications protocols are included in the ICD.

3.2 Versioning

Datasets will be generated on a timely basis as previously specified in **Table 5**, frequency of update. Each file or database storage system will include version information embedded in the storage container. Depending on the dataset, the data will include published date, activation date, and deactivation or expiration date. The versioning information will be included in metadata files that are associated with the dataset.

Derived datasets, such as aggregated data or summary data, will include the duration for which the datasets span as well as references for the datasets. In addition, the publication date will be identified as the version for the dataset.

The system cannot version datasets that are imported from other sources, however, the system will log the date and time the dataset was ingested or when a transaction occurs.

3.3 Metadata

The USDOT requires metadata with each dataset submission and project information to support research and enable search and discovery. This section describes the definitions for entries into the Metadata files. These metadata types describe information about the datasets, their source(s), processes, quality, restrictions, and usage.

3.3.1 Metadata Types

Several datasets contain formats that describe their source, ownership, usage, processing, quality, and lineage. For example, the GTFS specification includes two files (feed_info.txt and attributions.txt, see [GTFS]) that contain appropriate metadata information. The metadata types described in this section will be used for datasets derived from the ST-CTN project that do not have existing metadata or versioning formats.

1. **Discovery** – Metadata that are used to allow other users to find and work with the data. The types of metadata will include:
    a. **Responsibilities:** Who publishes, owns, and produces values for the datasets.
    b. **Permissions:** What user role is required to view the dataset to ensure privacy and security of the data is maintained.
    c. **Access:** Where the dataset is stored and how it is accessed.
    d. **Abstract:** What is included in the dataset and what is its intended purpose.
    e. **Lineage:** How the dataset was created, derived, or processed (summary level). This information points to the technical metadata file that describes the processes in more detail.
    f. **Versioning:** When the dataset was created, produced, or derived.
    g. **Publication:** When the dataset was published (optional).
    h. **Currency:** What is the currency of the dataset (when were the data active, from activation to deactivation).
    i. **Licensing:** What are the restrictions on the data use including copying, publishing, distributing, transmitting, citing, or adapting the data. This field may contain a link to the licensing terms and agreements document.

2. **Technical Metadata** – Data that are used to provide technical details on the data.
    a. **Data Schema** – Metadata that document the exact fields, data elements or objects in the data including, field name, description, data type, and notes. This information will be at the asset level, describing the contents of the data or dataset generally, in accordance with the Project Open Data Metadata Schema another appropriate standard.
    b. **Data Collection Method** – Metadata that document the data collection method used to acquire the data. The information includes the quality and accuracy of the collection devices and methods.
    c. **Data Processing** – Metadata that document any data processing that was done to the data from the data inception (when the data was produced) to when it was processed for the USDOT.
    d. **Data Impact Log** – Metadata that provide information on any changes to data during the collection period. Any time the data change in a unique way that is not expected in the experimental design either by internal or external forces is documented here. Some examples of possible triggers for updates include:
        i. Internal Events
            1. Data collection start and end
            2. Testing
            3. Changes in the deployment scope
            4. Software updates that modify data fields
            5. Sensor disruptions, either communication-based or hardware-based
        ii. External Events
            1. Volume increases or reductions
            2. Construction and/or work zones
            3. Special events in area
            4. Changes in roadway or sidewalk configurations

        This impact log will include information on the date and time of the impact, type of trigger categories, particularly those related to schema/data field or data collection method changes or gap/disruptions to collection.

        - If a schema changes, the previous and current data field and replacement impact will be described.

- If a data collection or processing change, the impact to the dataset quality or data collection update frequency will be described.
- If gap or disruption, the duration of the disruption will be described.

3.3.2 Metadata Structure

The Metadata Structure will be delivered in a flat file format encoded as JSON or comma-separated values (CSV).

A compressed set of files that is composed of the following embedded folders will be provided:

ST-CTN project file – contains a summary of the project information including:

- Discovery information (as defined in **Section 3.3.1**, limited to project file publication)
- Folders containing metadata and the related metadata formats used (this will identify if the data use a standard format rather than the DMP Metadata Types)
- Link to published DMP

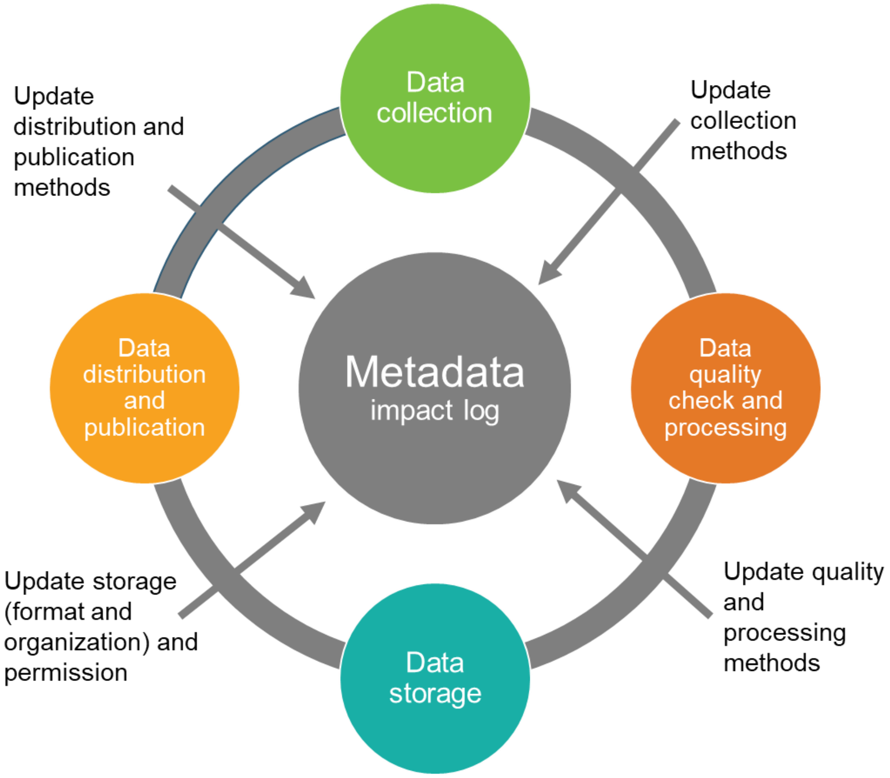
Dataset metadata file (folders) – Discovery and technical metadata information for each dataset. The core datasets will be included in this set of files because they may be collected or derived using different collection and processing methods. The core dataset files may include sidewalk and pathway information, signal system and RSU asset data and stop information (from GTFS). The associated files per dataset will include:

- Discovery information (as defined in **Section 3.3.1** or using native specification format)
- Schema or Standard Profile (if not an open specification or standard) – includes the version used to generate the referenced dataset
- Data collection and processing information
- Impact Log

3.3.3 Metadata Update Process

A new metadata file will be generated for each dataset that is published. The update process will follow the data curation process as each curation process occurs (depicted in **Source:** ARC, 2022

Figure 6). the methods will be verified and changes documented to relevant metadata fields, as well as logged in the Impact Log.


Data collection
Update collection methods
Data quality check and processing
Update quality and processing methods
Data storage
Update storage (format and organization) and permission
Data distribution and publication
Update distribution and publication methods
Metadata
impact log

Source: ARC, 2022

Figure 6. Metadata Update Process

This update process includes:

- Updating collection methods during and following the data collection stage.
- Updating quality and processing methods following the data quality checking and processing stage.
- Updating data storage including formats, organization, and storage methods. This also includes updating information on permission levels (note: the permissions are generated at this level, because the database management system will enforce a major portion of the data security and privacy provisions).
- Updating distribution and publication methods (e.g., APIs or flat file formats) for datasets.

Appendix A. Terms and References

This section provides acronyms, abbreviations, and glossary of terms used throughout the DMP.

Acronyms

ABM - Activity Based Model

APC – automated passenger count

API – application programming interface

ARC – Atlanta Regional Commission

ATIS – Advanced Traveler Information System

AVL – automatic vehicle location

BSM – basic safety message

CBI – commercial business interest

CDP – Connected Data Platform

ConOps – Concept of Operations

CSV – comma-separated value

CV – connected vehicle

CV1K – Regional Connected Vehicle Infrastructure Deployment Program

C-V2X – Cellular – Vehicle to Everything

CVTMP – Connected Vehicle Technology Master Plan

DSRC – Dedicated Short-Range Communication

DMP – Data Management Plan

DOR – Department of Revenue

DPP – Data Privacy Plan

EDG – Enterprise Data Governance

EDS – Enterprise Data Steward

EX ID – Data Exchange ID

FHWA – Federal Highway Administration

FTA – Federal Transit Administration

G-MAP – Georgia Mobility and Accessibility Planner

GA Tech – Georgia Institute of Technology

Ride Gwinnett / GCT – Gwinnett County Transit

GDOT – Georgia Department of Transportation

GTFS – General Transit Feed Specification

GTRI – Georgia Tech Research Institute

ICD – Interface Control Document

IE – independent evaluator

IOO – infrastructure owner and operator

IRB – Institutional Review Board

ITS – Intelligent Transportation Systems

JPO – Joint Program Office

JSON – JavaScript Object Notation

KPI -- key performance indicators

LEP – limited English proficiency

MARTA – Metropolitan Atlanta Rapid Transit Authority

MOU – memoranda of understanding

MVP – Minimum Viable Project

NDA – non-disclosure agreement

OBU – onboard unit

OSM – OpenStreetMap

OSS – Open Source Software

OST – Office of the Secretary

OTP – Open Trip Planner

PACE – Partnership for an Advanced Computing Environment

PED-SIG – Mobile Accessible Pedestrian Signal System

PII – personally identifiable information

PMD – Performance Management Dashboard

PROW – public right of way

REST – representational state transfer

RSU – roadside unit

ST-CTN – Safe Trips in a Connected Transportation Network

STM – space time memory

TMC – traffic management center

TPI – transit pedestrian indication

TSP – transit signal priority

TSR – transit stop request

USDOT – U.S. Department of Transportation

VRU – vulnerable road user

XML – extensive markup language

Glossary

The following table provides a summary of terms by category used throughout the DMP.

Table 10. Glossary

Category	Term	Definition
Dataset Type	Assets	Asset information about facility and Automated Traffic Management System devices-- sensors, signals, comm including electronic signs, PED-X signals, etc.
Dataset Type	Crowdsource	Data generated by the "crowd"
Dataset Type	CV	Connected vehicle produced dataset
Dataset Type	Demographics	Data about the structure of populations
Dataset Type	Land Use	Human use of land
Dataset Type	Mobility Service API	Application programming interfaces (API) or services that are pushed or pulled by an application.
Dataset Type	Network	Infrastructure characteristics and topological connectivity including right of way, PROW (sidewalk, crosswalk, bike lanes/paths)
Dataset Type	Network operating conditions	Condition and status of network infrastructure including planned and unplanned events: incidents, special events, work zones, and other impacts.
Dataset Type	System-Customer Performance	Datasets that require archiving and use by performance measures
Dataset Type	Transit	Transit network (routes) and realtime conditions/event data
Dataset Type	VRU Modes	Types of VRUs; used to describe categories of default impedance values

Category	Term	Definition
Dataset Type	Weather	Weather data
Dataset Structure	Link, Node	Link refers to a roadway or pathway. Node refers to an intersection or junction. Links and nodes are used to describe the structure of a path within the context of trip planning models, OSM, and OTP.
Access Level	PII Certification	Data that have PII included in the dataset. The access to this data should be as restrictive as possible to protect the PII based on IRB-approved processes. Data in this category should have an operational purpose that justifies its storage.
Access Level	Private	Data that cannot be shared with external users. Access to these data is limited and only granted with IRB and Project Team approvals. Private data include four subcategories: operational, proprietary, research and PII Certification.
Access Level	Proprietary	Licensed data from third parties or CBIs. These data may be used for planning or operational purposes. Any access to the data is determined by usage agreements between the parties.
Access Level	Operational	Realtime and other data used in the applications and operations of the system. The data may contain licensed data restricted by usage agreements.
Access Level	Research	Data that are available for research, but users of the data must meet IRB requirements before gaining access to the data. These datasets may have PII. These datasets are compiled for machine learning and research purposes that do not contain PII.
Access Level	Open	Data that can be used by the public with no or limited licensing restrictions. These data are available to the public without needing to request permissions and will be provided to the USDOT-managed Public System. These will be anonymized or aggregated versions of private datasets to protect PII.
Collection Method	Derived	Data are derived from one or more sources for summary or fusion purposes.
Collection Method	External Input	Data are ingested from a third-party source (the ingestion process may be through a digital interface or through manual processes).
Collection Method	Collect / forward	Data created, collected, forwarded and stored. These include user-input transactions (e.g., between APIs), web forms, user tracking methods (e.g., trace data from mobile phones).

References

The following table lists the documents that were used to support the development of the ST-CTN Phase 2 DMP document. References to these documents are identified with the acronym provided in brackets.

Table 11. References

ID	Referenced Documents
[ConOps]	Atlanta Regional Commission. Deliverable Task 2 Concept of Operations. Atlanta: U.S. Department of Transportation. (2021).
[CV1K]	Georgia Department of Transportation. "The Regional Connected Vehicle Program Scope of Work." Atlanta: Georgia Department of Transportation.
[CVTMP]	AECOM. "Gwinnett County Connected Vehicle Technology Master Plan (CVTMP)." Duluth: Gwinnett County Department of Transportation. (2019).
[DMP]	Atlanta Regional Commission. Deliverable Task 3 Data Management Plan. Atlanta: U.S. Department of Transportation. (2021).
[DPP]	Georgia Department of Transportation. Deliverable Task 2C Data Privacy Plan. Atlanta: U.S. Department of Transportation. (2023).
[GTFS]	GTFS. General Transit Feed Specification Reference. Washington D.C.: GTFS. (2019).
[ICD]	Georgia Department of Transportation. Deliverable Task 2B Interface Control Document. Atlanta: U.S. Department of Transportation. (2023).
[IPFP]	Atlanta Regional Commission. Deliverable Task 10 Institutional, Partnership, and Financial Plan. Atlanta: U.S. Department of Transportation. (2022).
[PMESP]	Atlanta Regional Commission. Deliverable Task 5 Performance Measurement and Evaluation Support Plan. Atlanta: U.S. Department of Transportation. (2021).
[SySR]	Atlanta Regional Commission. Deliverable Task 6 System Requirements Specification. Atlanta: U.S. Department of Transportation. (2021).
[VPFP]	Guensler, R., Y. Xu, V. Elango. Value Pricing Fellowship Project. Atlanta: Georgia Department of Transportation. (2013).

Appendix B. Comprehensive List of Standards Used

The Comprehensive list of standards used provides the set of profiles, message sets, and data dictionaries that are used by the datasets to organize and encode the datasets, and define the data meaning, formats, and enumerated values of the data.

- **Standard Identifier:** unique identifier assigned to the identifier in this document.
- **Standard Profile:** the name of the profile that customizes provisions of one or more message or data standards.
- **Standard Name(s):** the name of the data standard in which the data are made available to the USDOT.
- **Data Standard Digital Object Identifier(s) (DOI):** the DOI of the standard for the data. A URL to the data standard(s). "N/a" indicates that a reference is not available on-line.
- **Open or Proprietary:** Indicates whether the data standard is "Open" or "Proprietary." Open includes specifications published using the creative commons license or licensed by a consensus-based standards organization.
- **File Format Used:** the encoding and file format approach used. Values include CSV, JSON, XML, or other.

Table 12. Comprehensive List of Standards

Std ID	Standard Profile	Std Name(s)	DOI / URL	Open / Proprietary	File Format Used
STD-1	GTFS	GTFS	https://gtfs.org/reference/static	open	CSV
STD-2	GTFS Realtime	GTFS Realtime	https://gtfs.org/reference/realtime/v2/	open	GTFSrealtime.proto

Std ID	Standard Profile	Std Name(s)	DOI / URL	Open / Proprietary	File Format Used
STD-5	OSM	OSM	https://www.openstreetmap.org/about	open	Geographic format (can be exported to multiple standard formats including Shape Files)
STD-6	PSM	SAE J2735 and J2945/9	J2735SET_202007 -- https://www.sae.org/standards/content/j2735set_202007/ J2945_201712 -- https://www.sae.org/standards/content/j2945_201712/	open	XML

Appendix C. Dataset Impact Log

This section describes the changes to datasets as described in Section 1.1.2. Table 13 includes a log of datasets that were changed from the last version (i.e., Version 1) to this current version. The table includes the dataset identification number and name, current state (initial, updated, removed, replaced), the version in which it was last specified, and information on changes to the dataset or its characteristics, when and what was changed (e.g., table and item in table).

For example, if the Ride Gwinnett GTFS dataset was changed because a new version of the GTFS standard was used, then the table would include two entries for the Ride Gwinnett GTFS, the first with a state of “initial” and the Phase 1 version (1.0), and the second with the same ID and name, but with a state of “updated”, DMP Version of 1.0, DMP Update of 2.0 (for Phase 2), and Reason would include **Table 3** row # and a statement that states “dataset now conforms to current standard version”.

The current Dataset Impact Log lists the changes from the Phase 1 DMP to the Phase 2 DMP. The log includes datasets that were removed due to duplication or are not stored by the ST-CTN project. As mentioned in the dataset naming section, dataset IDs will not be reused. To that end, the current set of valid dataset identifiers stored by the system are not sequential.

Table 13. Dataset Impact Log

Dataset ID	Dataset Name	Status	DMP (last)	DMP Update (current)	Reason
1	Parcel-level Land Use Data	Removed	1	2	Data are not stored. Used only as metatags in whole roads network.
2	Building Address and Landmark Data	Removed	1	2	Data are not stored. Used only as metatags in whole roads network.
4	ABM Network	Removed	1	2	Data are not stored. Used only as metatags in whole roads network.
6	NaviGAtor Network	Removed	1	2	Data are not stored. Used only as metatags in whole roads network.
7	SRTA Managed Lane Network	Removed	1	2	Data are not stored. Used only as metatags in whole roads network.
9	Licensed Networks	Removed	1	2	Data are not stored. Used only as metatags in whole roads network.
12	GTFS Transit Network	Removed	1	2	Redundant with Dataset IDs 36 and 37
13	TransitSim Network	Removed	1	2	Redundant with Dataset IDs 36 and 38

Dataset ID	Dataset Name	Status	DMP (last)	DMP Update (current)	Reason
14	BikewaySim Network	Removed	1	2	Not used for ST-CTN project
16	SRTA Data	Removed	1	2	Data are not stored.
17	SRTA Tolling Data	Removed	1	2	Data are not stored.
18	Subscription Roadway Operating Condition Data	Removed	1	2	Data are not stored.
19	Historic Roadway Operating Condition Data	Removed	1	2	Data are not stored.
21	Waze Alerts	Removed	1	2	Data are not stored.
22	GDOT traffic management center (TMC) Incident Data	Removed	1	2	Data are not stored.
23	GDOT TMC Special Event Data	Removed	1	2	Data are not stored.
24	GDOT TMC Work-Zone Data	Removed	1	2	Data are not stored.
28	Pedestrian Intersection Asset Design and Condition Data	Removed	1	2	Redundant with Dataset ID #29
38	GTFS-Flex	Removed	1	2	Not used for ST-CTN project
40	PSM	Removed	1	2	Data are not stored.
41	SPaT	Removed	1	2	Data are not stored.
42	MAP	Removed	1	2	Data are not stored. Used only as metatags in whole roads network.
43	Signal Status Message Exchange	Removed	1	2	GDOT implemented different signal system that does not use NTCIP
47	Weather data	Removed	1	2	Not included in prediction engine
48	Customer Demographic Data	Removed	1	2	No longer needed
49	Household level Licensed demographic data	Removed	1	2	No longer needed
50	Household level Vehicle Registration Data	Removed	1	2	No longer needed
54	Trip Crowdsource Reports	Removed	1	2	Removed crowdsourcing from project scope
55	Geocode. earth API	Removed	1	2	Data are used but not stored
56	Business Level Licensed Facility Data	Removed	1	2	No longer needed

Dataset ID	Dataset Name	Status	DMP (last)	DMP Update (current)	Reason
57	MOVES-Matrix Energy Consumption and Emission Rates	Removed	1	2	No longer needed
61	Secure MU Gateway event logs	Removed	1	2	No longer needed
62	Pedestrian Crash Data	Removed	1	2	Data will be used in performance measurement but not stored on server
63	Pedestrian Incidents Police Reports	Removed	1	2	Data will be used in performance measurement but not stored on server
29	Pedestrian Intersection and Pathway Asset Design and Condition Data	Updated	1	2	Updated name to merge dataset ID #28 with 29.
Table 3	Several datasets	Edited	1	2	Edited descriptions for clarity in Table 3 Summary Table
Table 3	Several datasets	Edited	1	2	Edited and added subset descriptions for clarity in Table 3 Summary Table
Table 4	Several datasets	Updated	1	2	Updated Frequency of several datasets from continuous to as needed, monthly or daily.

U.S. Department of Transportation
ITS Joint Program Office-HOIT
1200 New Jersey Avenue, SE
Washington, DC 20590

Toll-Free “Help Line” 866-367-7487
www.its.dot.gov

FHWA-JPO-22-974

U.S. Department of Transportation