# Phase 1 Data Management Plan (DMP)

## University of Washington ITS4US Deployment Project

https://www.its.dot.gov/index.htm

**Final Report—September 23, 2021**
**FHWA-JPO-21-869**

COMPLETE TRIP
ITS4US

U.S. Department of Transportation

Produced by University of Washington ITS4US Deployment Program, Phase 1
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

## Technical Report Documentation Page

| 1. Report No. | 2. Government Accession No. | 3. Recipient's Catalog No. |
|---|---|---|
| **FHWA-JPO-21-869** | N/A | N/A |

| 4. Title and Subtitle | 5. Report Date |
|---|---|
| Phase 1 Data Management Plan (DMP)—University of Washington ITS4US Deployment Project | September 23, 2021 |
| | **6. Performing Organization Code** |
| | N/A |

| 7. Author(s) | 8. Performing Organization Report No. |
|---|---|
| Anat Caspi, University of Washington; Mark Hallenbeck, University of Washington, Erin Flanigan, Cambridge Systematics; Mark Jensen, Cambridge Systematics; Alice Marecek, Cambridge Systematics; Adam Danczyk, Cambridge Systematics; Guilherme Leao, Cambridge Systematics | |

| 9. Performing Organization Name and Address | 10. Work Unit No. (TRAIS) |
|---|---|
| University of Washington<br>4333 Brooklyn Ave NE<br>Box 359472<br>Seattle, WA 98195-9472 | N/A |
| | **11. Contract or Grant No.** |
| | 693JJ321C000004 |

| 12. Sponsoring Agency Name and Address | 13. Type of Report and Period Covered |
|---|---|
| U.S. Department of Transportation<br>ITS Joint Program Office<br>1200 New Jersey Avenue, SE<br>Washington, DC 20590 | N/A |
| | **14. Sponsoring Agency Code** |
| | HOIT-1 |

**15. Supplementary Notes**

Kate Hartman, COR

**16. Abstract**

This document discusses the Data Management Plan (DMP) for the University of Washington (UW) ITS4US Deployment Project. This project currently is in Phase 1 planning and systems engineering development, in which the preliminary idea is developed into a structured concept that is suitable for further design, building, testing, and operation. The structured concept will include identifying specific performance measures, targets, and capabilities associated with performance monitoring and performance management.

This DMP is informed by a Concept of Operations (ConOps) for the proposed system, which bridges between the needs that motivated the project and the specific technical requirements from which the project is built. The DMP discusses data management, analysis, storage, and other data-related topics that are relevant to the United States Department of Transportation's (USDOT) data requirements.

This document is intended to be a living document. While preliminary recommendations are identified herein, many of these decisions regarding data storage, management, standards, schema, and descriptions will be worked out as part of development in Phase 2. It is anticipated that many items will be adjusted, but the underlying intent of the proposed data management will be preserved.

| 17. Keywords | 18. Distribution Statement |
|---|---|
| ITS4US; Complete Trip; Deployment; ITS; Intelligent Transportation Systems; Data Management Plan; Accessibility; Sidewalks; Navigation Software; Data Standards | N/A |

| 19. Security Classif. (of this report) | 20. Security Classif. (of this page) | 21. No. of Pages | 22. Price |
|---|---|---|---|
| N/A | N/A | 46 | N/A |

**Form DOT F 1700.7 (8-72)** **Reproduction of completed page authorized**

# Revision History

| Name | Date | Version | Summary of Changes | Approver |
|---|---|---|---|---|
| Mark Hallenbeck, University of Washington; Mark Jensen, Cambridge Systematics | 26 July 2021 | 1 | Initial Draft | |
| Mark Hallenbeck, University of Washington; Mark Jensen, Cambridge Systematics | 23 August 2021 | 2 | Final | |
| Mark Hallenbeck, University of Washington; Mark Jensen, Cambridge Systematics | 23 September 2021 | 2.1 | Minor revisions made in response to five USDOT comments. | |

# Table of Contents

## List of Tables



## List of Figures

# 1. Introduction

This document discusses the Data Management Plan (DMP) for the University of Washington (UW) ITS4US Deployment Project. This project currently is in Phase 1, Concept Development, in which the preliminary idea is developed into a structured concept that is suitable for further design, building, testing, and operation. The structured concept will include identifying specific performance measures, targets, and capabilities associated with performance monitoring and performance management.

This DMP is informed by a Concept of Operations (ConOps) for the proposed system, which bridges between the needs that motivated the project and the specific technical requirements from which the project is built. The DMP discusses data management, analysis, storage, and other data-related topics that are relevant to the University of Washington's ITS4US deployment and the United States Department of Transportation's (USDOT) data requirements.

This document is intended to be a living document. While preliminary recommendations are identified herein, many of these decisions regarding data storage, licensing, management, standards, schema, and descriptions will be worked out as part of development in Phase 2. It is anticipated that many items will be adjusted, but the underlying intent of the proposed data management will be preserved.

A glossary and acronym list are provided in the appendix.

## 1.1. Project Background

The proposed project discussed in this DMP is the UW ITS4US Deployment Project, referred to as the ***Transportation Data Equity Initiative*** (TDEI), and it represents one of the Phase 1 Complete Trip—ITS4US Deployment Program projects selected by the USDOT to showcase innovative business partnerships, technologies, and practices that promote independent mobility for all travelers regardless of location, income, or disability. The UW ITS4US Deployment Project aims to increase the availability of pedestrian pathway and flexible transit data to travelers—primarily those underserved by the system as a result of disability or travel-specific preferences—by implementing sustainable data infrastructure, standards, and accessible mobility applications. By working in an active community of practice, developers hope to implement and adopt collaborative data standards that can be easily used by data contributors and application developers to create useful pathway information. This system design will ultimately be scalable to allow for ease of adoption in any geographic location without substantial overhaul of the structure, all while allowing various types of contributing groups a clear means to contribute valuable infrastructure data.

Further details on the project's background, motivations, and approach are discussed in the UW ITS4US Deployment Project Phase 1 ConOps.[1]

[1] Federal Highway Administration (FHWA), *Phase 1 Concept of Operations (ConOps)—University of Washington ITS4US Deployment Project*, FHWA-JPO-21-861, June 28, 2021.

# 2. Project Overview

A summary about this research project, its goals, and how the project data will help achieve USDOT's research goals is below:

- **Project Title:** UW ITS4US Deployment Project—Transportation Data Equity Initiative (TDEI).
- **Project Goals and Objectives:** This project has several main goals:
    - Work with three existing standards committees to extend and update existing, early-stage international data standards: OpenSidewalks, General Transit Feed Specification (GTFS) for demand-responsive or paratransit service (GTFS-Flex), and General Transit Feed Specification for pathways linking together locations within stations (GTFS-Pathways). These three data standards enable the consistent collection and reporting of data that provide the underlying information needed by the currently underserved target populations—people with disabilities, older adults, and individuals with low income—to efficiently travel.
    - Develop a series of tools that help agencies, jurisdictions, and other stakeholders collect the data that can be stored with these refined data standards. These tools are needed to lower the cost and improve the quality and consistency of those data collection efforts to increase the availability of the data.
    - Develop tools, policies, and procedures that allow sharing and governance of the collected data. The tasks performed will enable effective and efficient vetting, aggregation, management, and fusion of the data that participating agencies, jurisdictions, and other stakeholders collect. This portion of the project also includes tasks required to enable and manage the sharing of those data with application developers that write software to deliver requested travel information.
    - Develop an open-data repository to contain the data to be shared within the six counties that represent the geographic boundaries for this ITS4US project. The data repository will be developed to illustrate how these data can be collected, stored, governed, updated, and maintained over time, and then served upon request to application developers.
    - Develop three example applications that use the collected data. The three applications are intended to demonstrate three very different uses of the data that are collected, maintained, and made available to application developers as a result of the other four aspects of this project.

    The data collected by this project will make pathway and service attributes that currently are not widely available digitally visible, thereby, allowing travelers with specific travel preferences to utilize applications that recommend routes with attributes that satisfy those preferences. This will increase the accessibility for travelers who are not served (or served in a limited capacity) because this information currently is not available.
- **Project Description:** The project will produce a scalable architecture to facilitate data collection, processing/storage, and distribution of relevant pathway and transit (fixed-route and on-demand) service information. Although the end-users of the system will include travelers with specific preferences who benefit from improved information on pedestrian and transit services, a core component of this project is to establish a low barrier of entry for

prospective data contributors and application developers to facilitate widespread adoption of their applications across a wide geographic range. This project was envisioned as part of ongoing standards workshops that have sought to create usable standards to describe these environments. This project will ultimately develop the framework to collect, process, and distribute this information, allowing other interested groups to advance this effort.

- **Project Performance Measurements:** The UW ITS4US project's goals fall into five basic categories: Data Standards, Data Generation, Data Vetting, Data Provisioning/Publication, and Demonstration Application Performance. Performance measures and the data required by those measures come from a number of sources. The evaluation of the data standards is dependent on the opinions of the stakeholders that depend on the data, which will be obtained via surveys to be performed during the project, along with the extent to which data are present in the database. This measures the degree to which the data standards have been adopted and, when tracked over time, show how that acceptance is growing over time. In addition to using the amount of data generated (as noted above), Data Generation will be evaluated primarily by using quality assurance statistics, which are generated by the data collection/generation software, data vetting, and data upload activities. For example, the AI/ML software used to generate initial routable sidewalk networks will also report confidence levels associated with those data. Similarly, the multi-faceted data vetting programs being constructed and implemented will provide statistics about the quality of the data, either because the data pass those tests or because data need to be replaced. Outcome statistics from the vetting process will be stored and reported as part of that process because those measures are part of the process intended to maintain the quality of the dataset. The central database system will also generate and store performance logs that describe the number of and source for data queries, the response time associated with those queries, and any errors which occur when responding to those queries. Next, the project team will perform a number of lab and field tests that examine the performance of the demonstration applications with emphasis on the Multimodal AccessMap application, since that is a demonstration application being developed by the project team. These tests will capture data on the performance of the applications from recruited test subjects using software developed previously by the UW's Taskar Center. These data are not part of the data repository, but are a separate evaluation dataset. Finally, the project team will perform a number of stakeholder surveys to obtain input on the perceived performance of the system by the various stakeholder groups. These data too are not part of the primary project repository, but will be retained separately, specifically for project evaluation, and will be shared with the USDOT.

  More details are provided as part of the Performance Measurement Plan.

## 2.1. Change Control

Modifications to the DMP will be noted as part of the revision history log included in this document. Each noted revision will include a date, a revision number, and notation of which sections of the DMP received an update. An updated version of the DMP will be posted in place of the current DMP that is available. Depending on how the change control for Phase 2 is set up, this may be accompanied by an email notification to relevant project partners.

## 2.2. Relevant Sources

References to data standards and schema, as well as other development information, are provided below and in the UW ITS4US Deployment Project Phase 1 ConOps:[2]

- GTFS-Flex document (ongoing), https://docs.google.com/document/d/1v7jHgahgttthGXOtVrxMLXP_EE6PTcJtGnsO05k2cqk/edit.
- GTFS-Flex GitHub site, https://github.com/MobilityData/gtfs-flex.
- GTFS-Pathways document (ongoing), https://docs.google.com/document/d/1qJOTe4m_a4dcJnvXYt4smYj4QQ1ejZ8CvLBYzDM5lyM/edit#heading=h.edxt3s6om1lm.
- GTFS-Pathways GitHub site, https://github.com/google/transit/pulls?q=is%3Apr+pathways.
- OpenSidewalks website, https://tcat.cs.washington.edu/opensidewalks-2/.
- OpenSidewalks GitHub site, https://github.com/OpenSidewalks/OpenSidewalks-Schema.

## 2.3. Data Schedule

**Table 1** illustrates the anticipated schedule for data-related events, particularly those that include delivery to and discussion with the USDOT. This schedule is subject to updates, which will likely occur as insights are determined as part of the Phase 2 design effort.

**Table 1. Transportation Data Equity Initiative data-related schedule.**

| ID | Event Title | Description | Date |
|---|---|---|---|
| 1 | Draft DMP is delivered to USDOT. | Initial Draft DMP with basic information known at the time of writing. | July 2021 |
| 2 | Final DMP Phase 1. | DMP is updated with USDOT comments addressed. | August 2021 |
| 3 | Updates to DMP as a result of Phase 2 Design. | DMP is updated based on details identified in design. | Phase 2 |
| 4 | Initial data samples provided to USDOT. | Initial data samples are created, validated, and submitted to USDOT for review. | Phase 2 likely, but may be early in Phase 3 |

[2] The UW ITS4US Deployment Project Phase 1 Concept of Operations report can be found on https://www.its.dot.gov/its4us/index.htm.

| ID | Event Title | Description | Date |
|---|---|---|---|
| 5 | Initial meeting with USDOT data team to review data. | Meeting to review data with USDOT and walk through the data schema and DMP. | Phase 2 likely, but may be early in Phase 3 |
| 6 | Baseline data collection starts. | Initial collection of data on current conditions starts. | Phase 2 likely, but may be early in Phase 3 |
| 7 | DMP updated. | DMP may need to be updated multiple times, incorporating changes from testing and sample data schema, with a final update at the end of the project. | Phase 3 |
| 8 | Baseline data provided to USDOT. | Complete baseline datasets are uploaded to USDOT and the Independent Evaluation (IE). | Phase 3 |
| 9 | Month of testing of applications begins. | Initial upload after datasets is collected through testing. | Phase 3 |
| 10 | Data transferred to USDOT. | Semi-annual updates of after case data are provided to USDOT and IE. | Phase 3 |
| 11 | Data review. | Ongoing Data review conducted with USDOT and IE to ensure the datasets are complete. | Phase 3 |
| 12 | Draft Final Analysis Report submitted. | Draft Final Analysis Report submitted to USDOT. | Phase 3 |
| 13 | Final Analysis Report submitted. | Draft Final Analysis Report submitted to USDOT. | Phase 3 |

# 3. Data Overview

This section provides an overview of the data that will be collected as part of the UW ITS4US Deployment Project. A context diagram of the proposed TDEI system was developed as part of the UW ITS4US ConOps document. For reference, the context diagram is shown in **figure 1**. Section 3.1 provides a summary of the data needs for this project, referring to the context diagram developed as part of the ConOps document, and then separately identifies (via callouts in **Figure 2**) the key data processes that make up the system. Section 3.2 provides a data overview of the respective data flows (called out in **Table 2** in reference to **Figure 2**) that will move between system components. This DMP covers data that is sent to, stored, or transmitted from the proposed data repository through the processes developed by the UW team within the central system. The DMP does not deal with data that the data contributors have not submitted to the system, nor data collected by demonstration applications that is not submitted to the data repository in the form of a query.

## 3.1. Data Needs Summary

As described in the ConOps document, each subsystem and its components, in the context of the ITS4US Program, can be subdivided into several different efforts. These efforts include the following:

- Components that the UW team will directly develop and test, which primarily include the data validation and data services technologies that are the focal point of this project. In the context of **Figure 1**, these components include the data processing pipelines, the data repository itself, and the service pipelines.
- Components that the UW team will assist in developing to encourage data contributions, namely tool sets through which data providers will be encouraged to submit data. In the context of **Figure 1**, these tool sets will serve groups such as municipal governments, transit agencies, and other data providers.
- Components that represent software demonstrations whose development the UW team will support to illustrate the success of the pipelines. These include the three applications that have been vetted to provide the services needed by underserved end users. In the context of **Figure 1**, these components include the Taskar Center for Accessible Technology's (TCAT) (University of Washington) Multimodal AccessMap, Microsoft's Soundscape, and Unity Technologies' Digital Twin.
- Other components that provide data, such as data service providers. In the context of **Figure 1**, these include weather and other data services.

Various data flows in this context diagram will facilitate the overall system; and various components within subsystems will serve different roles to process, transform, and/or store these data. To summarize this, **Figure 2** illustrates the types and nature, scope, and scale of data that are expected to flow among the entities, offering a high-level view of all data moving in the proposed system and the entities exposed to those data. Call-outs in **Figure 2** refer to data flows that are outlined in **Table 2**.

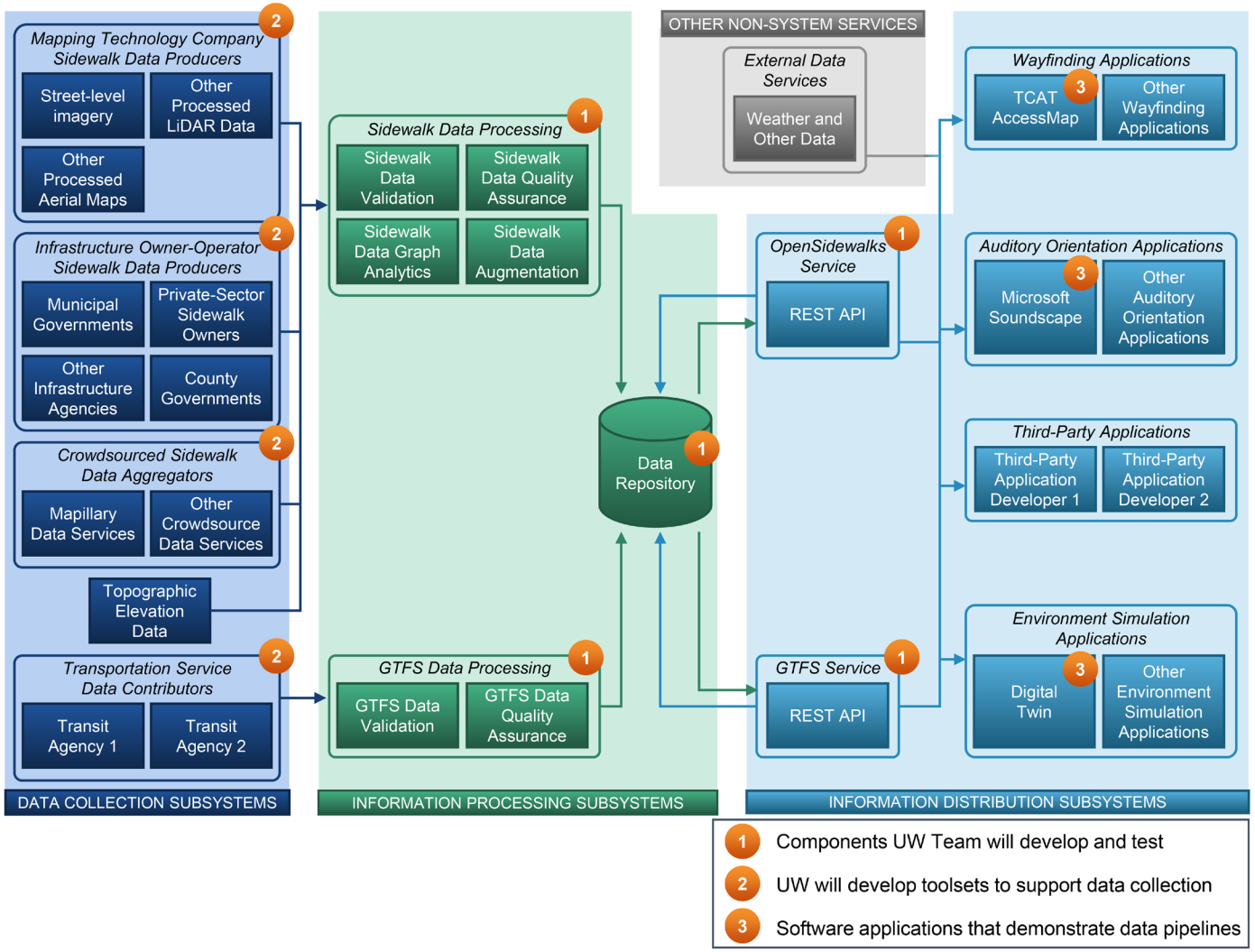


**Figure 1. Diagram. Context diagram for the proposed Transportation Data Equity Initiative system.**

*Source: University of Washington and Cambridge Systematics.*

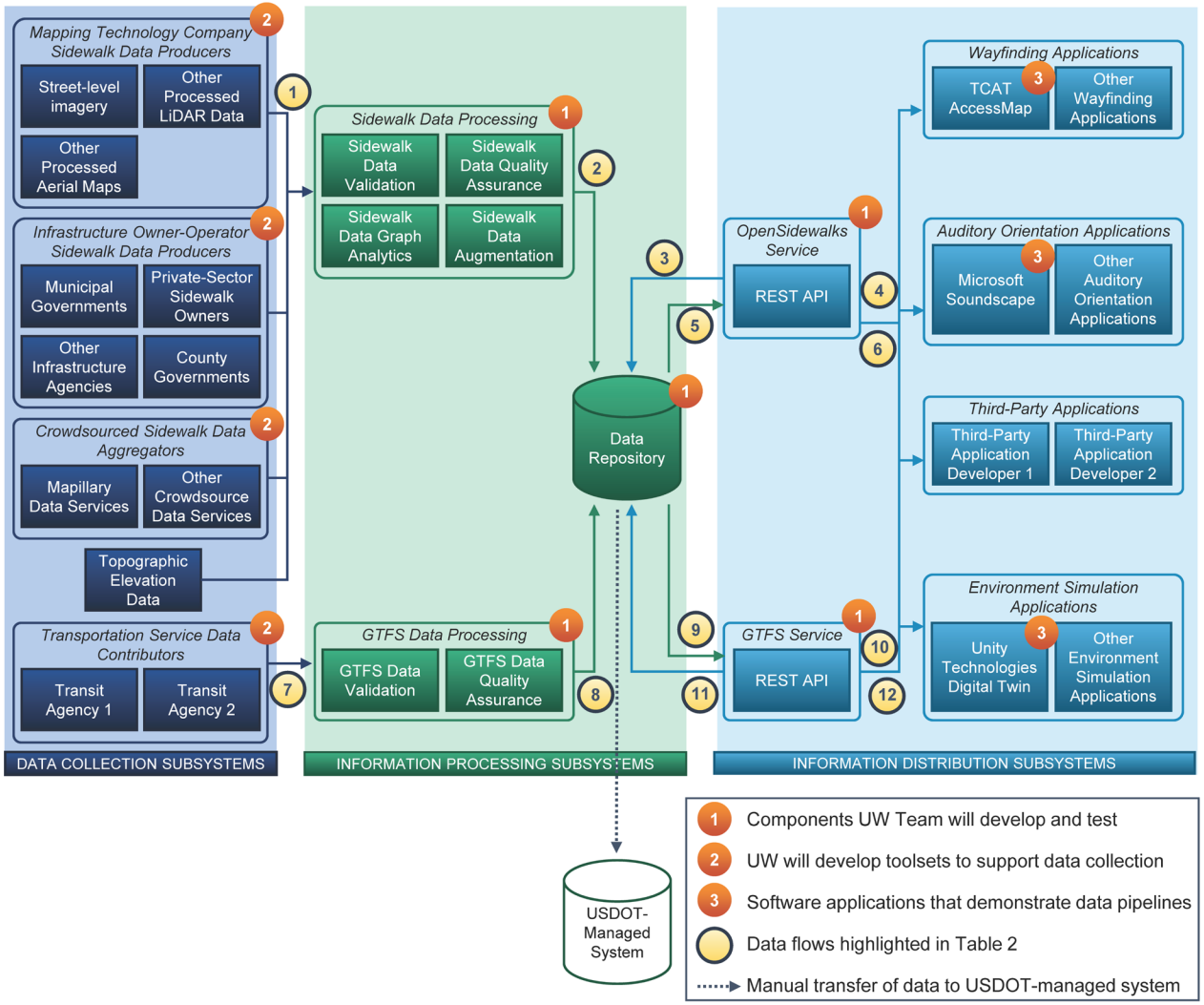


**Figure 2. Diagram. Context Diagram with data flow call-outs.**

*Source: University of Washington and Cambridge Systematics.*

**Table 2. Transportation Data Equity Initiative data overview table.**

| Dataset Title | Description | Type | Collection Method | Data File Format(s) |
|---|---|---|---|---|
| **Sidewalk Data (Data Flow #1)** | These data consist of attributes of the pedestrian infrastructure. | Numerical data, text sequences, positional data (latitude, longitude) | Currently collected through manual entry by local entities, including infrastructure owner-operators and local community crowdsourcing groups. Potential for automated collection by transforming other data sources is expected in the future. | Extensible Markup Language (XML) OpenStreet Map (.osm) file |
| **Graphed Sidewalk Data (Data Flow #2)** | These data consist of attributes of the pedestrian infrastructure transformed into a graph network. These data are validated and may be filtered. | Numerical data, text sequences, positional data (latitude, longitude) | Transformed sidewalk data that are received by the proposed system. | XML OpenStreet Map (.osm) file |
| **Fixed-Route Transit Service Data (Data Flow #7)** | These data consist of information on fixed-route transit service, including routes, bus stops, and other attributes. | Numerical data, text sequences, positional data (latitude, longitude) | Currently collected through manual entry by transit agencies or automatically through Geographic Information System (GIS) files and scheduling software. | GTFS Comma-Separated Values (.csv) file |
| **On-Demand Transit Service Data (Data Flow #7)** | These data consist of information on on-demand transit service, including operating hours, contact information, and other attributes. | Numerical data, text sequences, positional data (latitude, longitude) | Collected manually by transit agencies. | GTFS Comma-Separated Values (.csv) file |

| Dataset Title | Description | Type | Collection Method | Data File Format(s) |
|---|---|---|---|---|
| **Transit Station Pathway and Attribute Data (Data Flow #7)** | These data consist of attribute data on pathways within a transit station and other attributes. | Numerical data, text sequences, positional data (latitude, longitude) | Collected manually by transit agencies. | GTFS Comma-Separated Values (.csv) file |
| **Validated Fixed-Route Transit Service Data (Data Flow #8)** | These data consist of information on fixed-route transit service, including routes, bus stops, and other attributes. These data are validated and may be filtered. | Numerical data, text sequences, positional data (latitude, longitude) | Currently collected through manual entry by transit agencies. | GTFS Comma-Separated Values (.csv) file |
| **Validated On-Demand Transit Service Data (Data Flow #8)** | These data consist of information on on-demand transit service, including operating hours, contact information, and other attributes. These data are validated and may be filtered. | Numerical data, text sequences, positional data (latitude, longitude) | Collected manually by transit agencies. | GTFS Comma-Separated Values (.csv) file |
| **Validated Transit Station Pathway and Attribute Data (Data Flow #8)** | These data consist of information on pathways within a transit station and other attributes. These data are validated and may be filtered. | Numerical data, text sequences, positional data (latitude, longitude) | Collected manually by transit agencies. | GTFS Comma-Separated Values (.csv) file |
| **Pathway Request Data (Data Flow #4 for sidewalks and Data Flow #10 for transit service and/or stations)** | These data consist of end-user-initiated requests for trip information, including an origin, a destination, and any required waypoints along the way. Query is made to both the OpenSidewalk Service and the GTFS service in accordance with user's preferences. | Numerical data, text sequences, positional data (latitude, longitude) | Collected via applications through user entry and request. | JavaScript Object Notation (.json) file |

| Dataset Title | Description | Type | Collection Method | Data File Format(s) |
|---|---|---|---|---|
| **Filtered Pathway Request Data (Data Flow #3 for sidewalks and Data Flow #11 for transit service and/or stations)** | These data consist of end-user-initiated requests for trip information, including an origin, a destination, and any required waypoints along the way. Data are filtered via the Application Program Interface (API). Data requests are made from the service to the data repository for data of relevance. | Numerical data, text sequences, positional data (latitude, longitude) | Collected via applications through user entry and request. | JavaScript Object Notation (.json) file |
| **Pathway Response Data (Data Flow #5 for sidewalks and Data Flow #9 for transit service and/or stations)** | These data consist of pathway information, which is produced in response to a query and provided to the service. | Numerical data, text sequences, positional data (latitude, longitude) | Collected automatically in response to a query. | JavaScript Object Notation (.json) file |
| **Filtered Pathway Response Data (Data Flow #6 for sidewalks and Data Flow #12 for transit service and/or stations)** | These data consist of pathway information, which is filtered by the service and provided to the requesting application. | Numerical data, text sequences, positional data (latitude, longitude) | Collected automatically in response to a query. | JavaScript Object Notation (.json) file |

At a very high level, the following data flows, which correspond to the numbers shown in **Figure 2**, must be present for the proposed system. Breakouts on specific data flows are discussed in **Table 2**:

- **Sidewalk Data Contribution to the Data Repository (Data Flows #1 and #2):** The proposed system needs data contributors—with first-hand knowledge or data on sidewalk infrastructure—to provide information on sidewalk pathways and relevant accessibility attributes that can be added to a larger pool of sidewalk network data. Several types of entities may have information on sidewalk attributes or other characteristics of the pedestrian built-environment. These will generally include infrastructure owner-operators (public and/or private), crowdsourced sidewalk data aggregators that report their observations on sidewalks, and mapping technology companies that utilize advanced analytics on aerial imagery or light detection and ranging (LiDAR) data to generate observations on sidewalk attributes. In this proposed system, all data from these data contributors will be converted into the OpenSidewalks data format via the tools developed by UW and submitted those to the proposed system. Contributed sidewalk data will be screened and checked for accuracy by validators before being deposited in the data repository, primarily to validate that the data are compliant with the OpenSidewalks standard. A quality assurance check will be done for each data contribution to ensure that the contribution is reasonable; this process may be automated or manual (human verified) and is anticipated to evolve on the basis of available technology at the time of the project. The quality assurance check will compare the contributed data with other reliable sources; for example, a contributed sidewalk segment will be verified by using aerial imagery to confirm that the path exists. After quality assurance is complete, this component will transform the data into a graph segment and determine how it augments the existing graphed sidewalk network. These data—now filtered by the previous checks—will then flow to the data repository for storage. That repository will comprise a fully comprehensive sidewalk network (based on data that have been submitted).
- **Transit Service Data Contribution to the Data Repository (Data Flows #7 and #8):** The proposed system needs data contributors—with first-hand knowledge or data on their offered transit services and station pathways—to provide information on these services and stations that can be utilized as part of a complete trip. Entities with this information will primarily include transit service providers that operate transit or maintain station infrastructure to support transit services. In the proposed system, these data contributors will transform their data into the GTFS data format or, where applicable, a comparable extension. Contributed transit data will be screened for accuracy before being deposited in the data repository to validate that the data are compliant with the GTFS standard and its extensions. A quality assurance check will be done for each data contribution to ensure that the contribution is reasonable; this process may be automated or manual (human verified) and is anticipated to evolve on the basis of the available technology at the time of deployment. These data—now filtered by the previous checks—will then flow to the data repository for storage. That repository will comprise various network links (based on data that have been submitted).
- **Sidewalk Data Distribution to the Applications (Data Flows #3, #4, #5, and #6):** The proposed system needs to provide data upon request for a relevant geographic area of interest. At the application level, the end user will enter a preferred origin, destination, and trip-specific travel preferences. The application will retain the trip-specific travel preferences and query the proposed system for sidewalk options by utilizing the user's requested origin and destination for the trip. The OpenSidewalks data service will receive, and verify via the application's descriptive metadata, the request from the application through its secure API and pull all relevant data from the data repository. The relevant OpenSidewalks data will be

sent in the JavaScript Object Notation (JSON) data format to the application that made the original request. The application will utilize these data to identify paths (including any paths that involve a transit option, if requested or available); screen those paths based on the user's trip-specific travel preferences; and provide a recommended route to the end user.

- **Transit Service Data Distribution to the Applications (Data Flows #9, #10, #11, and #12):** The proposed system needs to provide data upon request for a relevant geographic area of interest. At the application level, the end user will enter a preferred origin, destination, and trip-specific travel preferences. The application will retain the trip-specific travel preferences and query the proposed system for sidewalk options by utilizing the user's requested origin and destination for the trip. The GTFS data service will receive the request from the application through its API and pull all relevant data from the data repository. The relevant data will be sent in the GTFS data format to the application that made the original request. The application will utilize these data to identify paths (including any paths that involve a sidewalk path option, if requested or available); screen those paths based on the user's trip-specific travel preferences; and provide a recommended route to the end user.

These data movements will allow the proposed system to operate. Data that are retained for performance measures will likely be collected from each of these data movements and ultimately stored in the data repository for administrative use. The proposed system aims to not collect Personally Identifiable Information (PII), although some Locational PIIs are noted in some of the data flows. This will be discussed in later sections. Data that may potentially qualify as PII (e.g., trip-specific travel preferences) will be retained on the user's mobile application, so no PII scrubbing will be required.

## 3.2. Data Overview

**Table 2** provides a data overview of all datasets that are envisioned as part of the proposed project. Referenced data flows under the Dataset Title refer to the numbers shown in **Figure 2**. **Table 2** discusses the following:

- **Dataset Title:** The title of the dataset.
- **Description:** Description of the dataset.
- **Type/Scale:** The anticipated type or scale of the dataset.
- **Collection Method:** How the data will be collected, as well as whether through a manual or automatic process.
- **Data File Format(s):** The anticipated data file format. Note that this may be subject to change as the system is developed.

Data flows will be kept at a high level as part of this DMP with detailed data flows currently under development. Detailed data flows and associated diagrams will be included as part of Phase 2.

# 4. Data Stewardship

This section provides details regarding data stewardship for the proposed system. Data stewardship involves proper data management throughout the data's lifecycle, including maintenance of data quality and safeguards. This section discusses the data owner and stewardship; respective access levels; and policies regarding reuse, redistribution, and derivation.

This section focuses on data stewardship in the context of the proposed system for the duration of the UW ITS4US Deployment Project. Stakeholders during the ConOps document walkthrough noted interest in two-directional feedback loops that could share data between the data repository and external groups, but that activity is outside the scope for the UW ITS4US Deployment Project. Data stewardship covered in this section applies only to data that are contributed to the proposed data repository for use in this proposed system.

## 4.1. Data Owner and Stewardship

**Table 3** defines the data owner and data steward for the datasets that will be part of the proposed system. In general, the USDOT Complete Trip—ITS4US Deployment Program will be the owner of all data collected, processed, and distributed as part of this USDOT-funded research project. The Washington State Transportation Center (TRAC) at the University of Washington and the TCAT will be the data stewards, composing the primary group developing the proposed system and coordinating with associated third-party entities. The Intelligent Transportation Systems Joint Programs Office (ITS JPO) is the USDOT sponsor for these datasets.

**Table 3** discusses the following information:

- **Dataset Title:** The title of the dataset.
- **Data Owner:** The anticipated dataset owner, which is the person or organization with the authority, ability, and responsibility to access, create, modify, store, use, share, and protect the data.
- **Data Steward:** The anticipated dataset steward, which, 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. The data steward also may 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 USDOT sponsor for this/these dataset(s).

**Table 3. Data owner and steward information.**

| Dataset Title | Data Owner | Data Steward | Federal Sponsor |
|---|---|---|---|
| **Sidewalk Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **Graphed Sidewalk Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **Fixed-Route Transit Service Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **On-Demand Transit Service Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **Transit Station Pathway and Attribute Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **Validated Fixed-Route Transit Service Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **Validated On-Demand Transit Service Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **Validated Transit Station Pathway and Attribute Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **Pathway Request Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **Filtered Pathway Request Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **Pathway Response Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **Filtered Pathway Response Data** | The University of Washington | TRAC and TCAT | ITS JPO |
| **Other Data Requests and Response** | The University of Washington | TRAC and TCAT | ITS JPO |

# 4.2. Access Level

The proposed system has two access levels:

1. **Open: Data that can be used by the public with no or limited licensing restrictions. This data is available to the public without needing to request permissions and will be provided to the USDOT-managed Public System. These may be anonymized or aggregated version of private datasets to protect PII.** Permission to access these data will be available across the entire program's scope. Any user may request access to the dataset in question, in accordance with the OPEN Government Data Act. This permission will include data that are non-PII and non-confidential business information (CBI), but will include no other information that threatens the privacy of an individual or group. At this time, this will predominantly include contributors that submit sidewalk or transit station/service data.
2. **Private (Research): Data that cannot be shared with external users. Access to these data is limited and only granted with Institutional Review Board (IRB) and Project Team approvals.** Permission to access these data will be limited to program managers. A specific concern will be the presence of PII, which includes potential PII, actual PII, locational PII, and sensitive PII, as well as other potential information that threatens the privacy of an individual or group. The proposed system will not capture CBI. The data collected by the proposed system may contain locational PII, which is captured by routing and pathway requests. This data is considered research private data, meaning it is available for research, but users of the data must meet IRB requirements before gaining access. These data will require controlled access, as described more in **Table 4**.

Other private data subcategories include PII data and proprietary data, which do not apply to this proposed system because no actual PII, licensed data from third parties, or CBI is used.

## 4.2.1. Private Datasets

This section lists the datasets that will be private and not shared with the public. For these datasets, the Fair Information Practice Principles will be followed. Users will be notified of the authority doing the collection, and the ultimate purpose, and will consent to the information being collected from them. Only necessary and relevant PII will be collected, and such data will only be used for the original purpose of the proposed system. The security of PII data will be protected, and any individuals who use or collect the data will be accountable for them. All other datasets not listed in this section are considered open and will be available to the public. The information listed here is the initial assignment. Future review may allow for more data to be made public under certain conditions or processes. **Table 4** provides a list of the private datasets, the reasons why the data will be private, and the associated safeguarding methods and processes.

**Table 4. Private datasets.**

| Dataset Title | Reason(s) the Data is Private | Safeguarding Methods and Processes |
|---|---|---|
| **Pathway Request Data (Data Flow #4 for sidewalks and Data Flow #10 for transit service and/or stations)** | Locational PII exists when a user requests pathway data for origins and destinations. | Data contain locational PII for the user's origin and destination, which can be augmented with other public data to identify a person. These data are necessary to provide good quality service and will have to be transmitted to the proposed system. To safeguard these data, the data themselves (i.e., geographical coordinates) will be stripped of several decimal points when stored for performance measures to decrease the resolution of the location. |
| **Filtered Pathway Request Data (Data Flow #3 for sidewalks and Data Flow #11 for transit service and/or stations)** | Locational PII exists when the service requests pathway data from the data repository for a user's specific origins and destinations. | Data contain locational PII for the user's origin and destination, which can be augmented with other public data to identify a person. These data are necessary to provide good quality service, and will have to be transmitted to the proposed system. To safeguard these data, the data itself (i.e., geographical coordinates) may be stripped of several decimal points when stored for performance measures to decrease the resolution of the location. |
| **Pathway Response Data (Data Flow #5 for sidewalks and Data Flow #9 for transit service and/or stations)** | Locational PII may exist when segments of the network are pulled from the data repository by the service for a user's specific request. | Data include network information for a specific user, which may be tied to an original origin-destination request for a specific user account and lead to locational PII being shared. These data are necessary to provide good quality service and will have to be transmitted to the proposed system. Limited safeguarding can be done for this particular dataset, but data stripping of the pathway request data (filtered or unfiltered) will decrease the resolution of the location. |

| Dataset Title | Reason(s) the Data is Private | Safeguarding Methods and Processes |
|---|---|---|
| **Filtered Pathway Response Data (Data Flow #6 for sidewalks and Data Flow #12 for transit service and/or stations)** | Locational PII may exist when segments of the network are sent from the service to an application for a user's specific request. | Data include network information for a specific user, which may be tied to an original origin-destination request for a specific user account and lead to locational PII being shared. These data are necessary to provide good quality service and will have to be transmitted to the proposed system. Limited safeguarding can be done for this particular dataset, but data stripping of the pathway request data (filtered or unfiltered) will decrease the resolution of the location. |

### 4.2.2. Access Request

A full IRB process will occur as part of the Phase 1 tasks to verify that these approaches meet the requirements for storing the necessary data while protecting user privacy.

### 4.2.3. Related Tools, Software, and/or Code

Different datasets may require different tools to access, view, and utilize the data. This is usually driven by the file format of the data standard that is being used. The exact tools or software available to access the data may differ by the time Phase 2 occurs and should be reviewed at that time to determine applicability, but it is not anticipated that any of these data will require unusual tools or software.

Comma-separated value (CSV), OpenStreetMap (OSM), and JSON files are plain text format files that can be read by most tools. No special formats are used that require specific tools or software.

For all CSV files, these data will require Microsoft Excel or a comparable data viewing tool that can read CSV files.

For all OSM files, a software that can read OSM data will be necessary in order to view the data. The data are structured in XML format in the form of "nodes" (points); "ways" (connections); and "relation" (street and object properties, such as tags). The OpenStreetMap project uses a Web interface for displaying and exporting OSM data, but a handful of specialized programs can be used to open OSM files, including Quantum Geographic Information System (QGIS), an open-source GIS software.[3]

[3] Quantum Geographic Information System: https://www.qgis.org/en/site/.

For JSON files, their standard data interchange format permits them to be accessed through almost any text editor, including Microsoft Notepad and Apple TextEdit.

### 4.2.4. Relevant Privacy and/or Security Agreements

The proposed system will operate as a public-facing service, in which data contributors can voluntarily submit data in exchange for a potentially wider user audience. Therefore, no relevant privacy and/or security agreements are anticipated to be necessary as part of the ITS4US Program development. In future scenarios that follow the completion of the ITS4US Program, a commercial adopter may acquire the system as part of its service offering. In that case, privacy and security agreements may need to be revisited, depending on that organization's policies and procedures.

## 4.3. Reuse, Redistribution, and Derivative Products Policies

**Table 5** provides the applicable licenses for each of the datasets and specifies whether an open license is utilized. It outlines the following information:

- **Dataset Title:** The title of the dataset.
- **License Used:** Applicable licenses that could be used for this dataset, with definition given if this is an open license or a non-open license.
- **Reasons for a Non-Open License:** The reason that a non-open license is recommended for this particular dataset.

As discussed in Section 3.1, sidewalk data will be converted to the OpenSidewalks data format and would follow that standard's file format, which is an OSM file format, before submittal to the proposed system. These data will be available through OSM's open database. Data exported from OSM are licensed under the Open Data Commons Open Database License (ODbl) Version 1.0. GTFS data, including extensions for pathways and on-demand transportation services, are collected and reported manually by individual transit agencies. Unless otherwise noted by a specific agency, these data will be shared under an open license, the Creative Commons Attribution 4.0 License.

**Table 5. Reuse, redistribution, and derivative products policies.**

| Dataset Title | License Used | Reason(s) for Non-Open License |
|---|---|---|
| **Sidewalk Data (Data Flow #1)** | Open Data Commons ODbL Version 1.0 | Not Applicable (N/A) |
| **Graphed Sidewalk Data (Data Flow #2)** | Open Data Commons ODbL Version 1.0 | N/A |

| Dataset Title | License Used | Reason(s) for Non-Open License |
|---|---|---|
| **Fixed-Route Transit Service Data (Data Flow #7)** | Creative Commons Attribution 4.0 License | N/A |
| **On-Demand Transit Service Data (Data Flow #7)** | Creative Commons Attribution 4.0 License | N/A |
| **Transit Station Pathway and Attribute Data (Data Flow #7)** | Creative Commons Attribution 4.0 License | N/A |
| **Validated Fixed-Route Transit Service Data (Data Flow #8)** | Creative Commons Attribution 4.0 License | N/A |
| **Validated On-Demand Transit Service Data (Data Flow #8)** | Creative Commons Attribution 4.0 License | N/A |
| **Validated Transit Station Pathway and Attribute Data (Data Flow #8)** | Creative Commons Attribution 4.0 License | N/A |
| **Pathway Request Data (Data Flow #4 for sidewalks and Data Flow #10 for transit service and/or stations)** | Institute of Education Sciences (IES)/National Center for Education Statistics (NCES) Restricted-Data License | Open-license is not used because of privacy concerns. |
| **Filtered Pathway Request Data (Data Flow #3 for sidewalks and Data Flow #11 for transit service and/or stations)** | IES/NCES Restricted-Data License | Open-license is not used because of privacy concerns. |
| **Pathway Response Data (Data Flow #5 for sidewalks and Data Flow #9 for transit service and/or stations)** | IES/NCES Restricted-Data License | Open-license is not used because of privacy concerns. |
| **Filtered Pathway Response Data (Data Flow #6 for sidewalks and Data Flow #12 for transit service and/or stations)** | IES/NCES Restricted-Data License | Open-license is not used because of privacy concerns. |

## 4.4. Data Storage and Retention

This section describes the data storage systems that will be used to store the project's data, the details of those data systems, and the duration for which data will be stored in each system. At a high level, three data storage systems will be required to support the needs of this initiative:

1. **Proposed System Data Repository:** This repository represents the location where data will be stored to facilitate primary operation of the proposed system. As the central focal point, it will ingest nearly all the identified datasets as part of its operation. Data stored in the proposed system's data repository will include both public and restricted datasets; therefore, the data repository's storage will control access at the user level. Allowable users will include the UW team, which will oversee the effort.
2. **USDOT Data Repository:** This represents the USDOT-managed data storage for purposes of providing allowable data to the public or to approved users. It will include all datasets that are provided as part of this project by the proposed system's data repository or the application data repository. Data stored in the USDOT-managed storage systems will include both public and restricted datasets; therefore, the data repository's storage will need to be partitioned into two storage groups—each with different levels of allowable user access.

Further details on these storage systems are discussed in the following sections.

### 4.4.1. Storage Systems

**Table 6** lists details associated with each data storage system. These include the type of system, the datasets that are anticipated to be stored in each data storage system, the initial storage date (based on the current proposed approach envisioned in Phase 1 of this project), the estimated update frequency, and the archiving and preservation period.

**Table 6. Data storage system details.**

| Data Storage System Type | Dataset Title(s) | Initial Storage Date | Frequency of Update | Archiving and Preservation Period |
|---|---|---|---|---|
| **Proposed System Data Repository Project Team—Public System** | • Sidewalk Data<br>• Graphed Sidewalk Data<br>• Fixed-Route Transit Service Data<br>• On-Demand Transit Service Data<br>• Transit Station Pathway and Attribute Data<br>• Validated Fixed-Route Transit Service Data<br>• Validated On-Demand Transit Service Data<br>• Validated Transit Station Pathway and Attribute Data | Storage begins after system acceptance testing. | Continuously, as data are submitted. | Through Phase 3 at a minimum. |
| **Proposed System Data Repository Project Team—Restricted System** | • Pathway Request Data<br>• Filtered Pathway Request Data<br>• Pathway Response Data<br>• Filtered Pathway Response Data | Storage begins after system acceptance testing. | Continuously, as data are submitted. | Through Phase 3. |
| **Application Data Repository Third-Party—Restricted System** | • Other Data Requests and Response | Storage begins after system acceptance testing. | Continuously, as data are submitted. | Through Phase 3. |

| Data Storage System Type | Dataset Title(s) | Initial Storage Date | Frequency of Update | Archiving and Preservation Period |
|---|---|---|---|---|
| **USDOT-managed—Public System** | • Sidewalk Data<br>• Graphed Sidewalk Data<br>• Fixed-Route Transit Service Data<br>• On-Demand Transit Service Data<br>• Transit Station Pathway and Attribute Data<br>• Validated Fixed-Route Transit Service Data<br>• Validated On-Demand Transit Service Data<br>• Validated Transit Station Pathway and Attribute Data | Storage begins one week after local storage occurs. | Semi-annually | 5 years |
| **USDOT-managed—Restricted System** | • Pathway Request Data<br>• Filtered Pathway Request Data<br>• Pathway Response Data<br>• Filtered Pathway Response Data<br>• Other Data Requests and Response | Storage begins one week after local storage occurs. | Semi-annually | 5 years |

## 4.4.2. Data Storage System Description

Details of the data storage system(s) will be determined as part of the Phase 2 development, as the structure, location, and back-up methods will depend on whether the system is hosted locally (e.g., at the UW) or supported by an interested teaming partner (e.g., a large data processing organization). It is anticipated that the data storage system will utilize back-up systems to provide full, incremental, or differential back-up capabilities to protect the data against drive failures. In this case, backup methods will meet USDOT requirements on data pickup and architecture. It is also anticipated that the data storage system will utilize access controls to provide access to only permitted users and administrators of the system, will limit access to datasets for certain user

classes (public and private), and will provide protection for sensitive private data. Cybersecurity policies will likely define access control, data encryption and physical security.

### 4.4.3. Cybersecurity Policies

Cybersecurity policies will be implemented to protect restricted datasets from unauthorized access. While this will be determined in greater detail as part of the Phase 2 effort, it is anticipated that cybersecurity policies will be system-specific, meaning the adopted approach will vary among the three types of data storage systems based on their owner organization. For certain data flows—namely, those that go to the proposed system's data repository—the applications that send restricted data may dictate some of the requirements, given that they are end-user-facing and that users are voluntary participants.

### 4.4.4. Data Security Policies and Procedures

Data security policies and procedures will be a necessary requirement for the proposed system. Many of the specific details will be determined as part of Phase 2 design work, but several high-level policies and procedures will need to be met:

1. **Confidentiality—**Access to the data repository will be allowed only for approved users, such as well-intentioned data contributors and application developers. Data contributors will be informed that no expectation of confidentiality exists with their data once submitted to the data repository, as those types of data provided (e.g., sidewalk and/or transit service/station information) will be generally for the public's information and will not include sensitive materials. For applications, the proposed system will only transmit generalized origin-destination information as a request, while user traveler preferences and any other personal information will be retained within the application itself. De-coupling the data requests and allowing the local application to do the analysis will aid in preserving user confidentiality.
2. **Availability—**Data contributors, having requested and received approval to contribute data, will expect that contributions will be uploaded in a timely fashion. Similarly, application developers that have requested and received approval to query for data will expect data to be provided in a timely fashion. By establishing approved groups that can work with the data repository, the system can be properly sized to handle the expected number of requests, provide responsive service, and not be susceptible to unauthorized data requests from external actors. Even though the approval process to utilize the data repository may constitute a simple request, similar to the use of certain Internet-based applications today, that will aid in preserving the system's ability to exchange data.
3. **Integrity—**The data repository will act as the approved record of data. The proposed system is designed so that most interactions with the data repository will be data requests, which will not impact the accuracy and consistency of data. However, to provide a robust set of data, data contributors will need to submit initial records and all associated updates to their pathway and transit networks. To prevent unauthorized, incorrect, or misunderstood modification of data, the proposed system will have validation and quality assurance checks in place. Validation will ensure that the data are in the correct data schema. Quality assurance—which will likely involve a person comparing data submissions with other known information—will maximize the likelihood of accurate data being placed into the proposed system.

4. **Authenticity and non-repudiation—**The proposed system will not involve any critical transactional data, involved with payment systems, or other security measures, so validating that data are authentic will be less of an issue than it would be for other systems. That said, the proposed system will rely on approved data validators that are trusted and competent in reviewing data submissions for schema conformance and accuracy.

### 4.4.5. Back-up and Recovery Policies and Procedures

Data should be backed up to provide a failsafe in the event of data loss or corruption. Backups will comply with the adopted USDOT policies for records retention at the time of deployment. The proposed system will utilize a form of data back-up that archives, at a minimum, each data contribution that is provided by a data contributor. For the research data that are collected as part of this effort, they will be backed up as well. Similar to business systems, data back-ups should be sent to an offsite location or a cloud service in the event of widespread damage to the proposed system's primary location.

Recovery of back-up data will need to occur in a timely fashion upon initiation of the restoration effort, although recovery times do not need to be near-instantaneous as found in back-up technology solutions like the Business Continuity and Disaster Recovery (BCDR) systems. With no part of the proposed system being responsible for mission-critical reporting to support public safety systems, BCDR-level data recovery will not be necessary. If data need to be recovered from the back-up location, a slower, more economical procedure may be utilized.

# 5. Data Standards

## 5.1. Data Standards

**Table 7** shows the data standards that are anticipated to be used as part of the proposed system. It provides descriptions of the following items:

- **Dataset Title:** The title of the dataset.
- **Data Standard(s):** The data standard anticipated to be used by the dataset.
- **Data Standard(s) Digital Object Identifiers (DOI):** The DOI(s) of the standard(s) for the data, if available (marked "N/A" if not available).
- **Open or Proprietary:** Whether the data standard is considered open or proprietary.
- **Data Standard(s) Rationale:** The rationale for using the chosen data standard for the dataset in question.

**Table 7. Data standards.**

| Dataset Title | Data Standard(s) | Data Standard(s) DOI(s) or URL(s) | Open or Proprietary | Data Standard(s) Rationale |
|---|---|---|---|---|
| **Sidewalk Data** | OpenSidewalks | https://wiki.openstreetmap.org/wiki/Proposed_features/sidewalk_schema | Open | OpenSidewalks provides graph-network attributes that objectively describe the pedestrian built-environment. |
| **Graphed Sidewalk Data** | OpenSidewalks | https://wiki.openstreetmap.org/wiki/Proposed_features/sidewalk_schema | Open | OpenSidewalks provides graph-network attributes that objectively describe the pedestrian built-environment. |
| **Fixed-Route Transit Service Data** | GTFS | https://gtfs.org/reference/static/ | Open | GTFS defines a common format for public transportation schedules and associated geographic information, and is widely adopted among transit agencies. |

| Dataset Title | Data Standard(s) | Data Standard(s) DOI(s) or URL(s) | Open or Proprietary | Data Standard(s) Rationale |
|---|---|---|---|---|
| **On-Demand Transit Service Data** | GTFS-Flex Extension | https://www.transitwiki.org/TransitWiki/index.php/GTFS-flex | Open | This proposed extension builds upon GTFS, which is widely adopted among transit agencies. |
| **Transit Station Pathway and Attribute Data** | GTFS-Pathways Extension and other affiliated extensions | https://gtfs.org/reference/static/<br><br>https://www.transitwiki.org/TransitWiki/index.php/GTFS-flex | Open | These proposed extensions build upon GTFS, which is widely adopted among transit agencies. |
| **Validated Fixed-Route Transit Service Data** | GTFS | https://gtfs.org/reference/static/ | Open | GTFS defines a common format for public transportation schedules and associated geographic information, and is widely adopted among transit agencies. |
| **Validated On-Demand Transit Service Data** | GTFS-Flex Extension | https://www.transitwiki.org/TransitWiki/index.php/GTFS-flex | Open | This proposed extension builds upon GTFS, which is widely adopted among transit agencies. |
| **Validated Transit Station Pathway and Attribute Data** | GTFS-Pathways Extension and other affiliated extensions | https://gtfs.org/reference/static/<br><br>https://www.transitwiki.org/TransitWiki/index.php/GTFS-flex | Open | These proposed extensions build upon GTFS, which is widely adopted among transit agencies. |
| **Pathway Request Data** | Hypertext Transfer Protocol Secure (HTTPS) | https://doi.org/10.17487/RFC2660 | Open | This communication protocol is widely used and provides secure communication for location-specific details. |

| Dataset Title | Data Standard(s) | Data Standard(s) DOI(s) or URL(s) | Open or Proprietary | Data Standard(s) Rationale |
|---|---|---|---|---|
| **Filtered Pathway Request Data** | HTTPS | https://doi.org/10.17487/RFC2660 | Open | This communication protocol is widely used and provides secure communication for location-specific details. |
| **Pathway Response Data** | OpenSidewalks; GTFS; GTFS-Flex Extension; GTFS-Pathways Extension and other affiliated extensions | https://gtfs.org/reference/static/<br><br>https://www.transitwiki.org/TransitWiki/index.php/GTFS-flex | Open | These data standards describe the sidewalk and transit service/station information available for a requested location. |
| **Filtered Pathway Response Data** | OpenSidewalks, GTFS, GTFS-Flex Extension, GTFS-Pathways Extension, and other affiliated extensions | https://gtfs.org/reference/static/<br><br>https://www.transitwiki.org/TransitWiki/index.php/GTFS-flex | Open | These data standards describe the sidewalk and transit service/station information available for a requested location. |

## 5.2. Versioning

The proposed system will initially utilize human data validators to ensure that new contributions to the data repository are accurate. Upon submission by a data contributor, the data validator will review and compare the contributed data with other known data sources, such as aerial images, photographs, or street-level imagery, to ensure that the new data are realistic. Upon completion of this check, the data validator will authorize the data to be entered into the data repository, where they will become (either in full or in part) a new version of the proposed graph network. Over time, some of these validation processes may become automated, potentially through a versioning schema, such as Semantic Versioning 2.0.0.[4] If any changes are made to any individual component of a working dataset and those changes are committed, a new version will be created. This new version will be documented and timestamped.

[4] Semantic Versioning 2.0.0. https://semver.org/.

The exact data repository version control procedures will be determined as part of development of Phase 2. However, version control is acknowledged to be a critical part of the proposed system, as identifying recent credible contributions will be extremely important for providing the latest information to users. Updates to data schema will result in notification being sent to approved API key holders—who would have provided contact information for use and include the data contributors and application developers—for their awareness of API changes as a result of schema changes.

## 5.3. Metadata

Metadata that are relevant to the datasets in this proposed system are outlined in **Table 8**. This information is anticipated to be worked out as part of Phase 2 development, and the proposed metadata listed herein will be updated as further information is garnered.

**Table 8. Anticipated metadata.**

| Dataset Title | Anticipated Metadata |
|---|---|
| • **Sidewalk Data**<br>• **Graphed Sidewalk Data** | Data collected through the OpenSidewalks data schema will follow the OSM data schema. OSM utilizes element, change set, and user metadata to describe information, such as the username making edits to the network, the number of edits, which editor was used, the commit message, and other data that support analysis of user metrics and insight into contributions. |
| • **Fixed-Route Transit Service Data**<br>• **On-Demand Transit Service Data**<br>• **Transit Station Pathway and Attribute Data**<br>• **Validated Fixed-Route Transit Service Data**<br>• **Validated On-Demand Transit Service Data**<br>• **Validated Transit Station Pathway and Attribute Data** | GTFS and its associated extensions will likely utilize existing GTFS metadata resources. The GTFS dataset specification includes various files, including a conditionally required “feed_info.txt”. This dataset file will include dataset metadata, such as publisher (name and webpage), version, language codes, and expiration information. Even though the extensions are still in the process of being adopted, it is likely that similar metadata will be adopted.<br>Metadata for validated datasets will match the metadata resources for nonvalidated GTFS data, but they also may include a tag to indicate whether they have been filtered/validated. |
| • **Pathway Request Data**<br>• **Filtered Pathway Request Data** | For data queries made by applications, the exact metadata will be determined as part of Phase 2. Metadata for that dataset are anticipated to likely include the contributing application, as well as the time when data were requested. |

| Dataset Title | Anticipated Metadata |
|---|---|
| • **Pathway Response Data**<br>• **Filtered Pathway Response Data** | Metadata for pathway response data will be determined as part of Phase 2. Pathway data sent to applications will likely include all metadata found in the OSM and GTFS standards, as well as additional tags that may include the receiving application and the time when data were sent. |
| • **Other Data Requests and Response** | Metadata for other data requests and responses will be determined as part of Phase 2, depending on the data resources that are utilized. It is very likely that metadata, at a minimum, will include metadata from the response application, as well as a timestamp of when the data exchanges occur. |

### 5.3.1. Metadata Types

This section outlines the different types of metadata that are anticipated to be included as part of the proposed system. Many of the details about what metadata will be available will be worked out as part of the Phase 2 development, but metadata are anticipated to be divided among descriptive, structural, and administrative types. The following is anticipated:

1. **Administrative Metadata—**Data that relate to the technical source of a digital asset, focusing on file type, how the asset was created, usage rights, ownership, use, or other insights. The National Information Standards Organization[5] defines three subtypes:
    a. Technical Metadata—Information relevant to data schema and any processing that was done, or data processing impact logs.
    b. Preservation Metadata—Information necessary for the long-term storage of digital assets, including owner information and authenticity.
    c. Rights Metadata—Information pertaining to usage rights and licensing.
2. **Descriptive Metadata—**Data that describe and identify information resources, including data contributor information, data requestor information for subset data requests, and attributes about the project:
    a. Bibliographic Attributes—Data contributor source.
    b. Query Attributes—Data requestor, or requesting application, source.
    c. Project Attributes—Static information to describe key parts of the project, such as geometry of the study area.

[5] National Information Standards Organization. http://www.niso.org/.

3. **Structural Metadata—**Information that facilitates navigation and presentation of electronic resources, such as the relationship between subset data and the contributed dataset:
    a. Relationship Attributes—Describes relationship among materials (e.g., data request utilized information supplied by a list of contributors).

## 5.3.2. Metadata Structure

The metadata structure for the proposed system will be known in Phase 2 after further design efforts have been undertaken. At a high level, a structured folder format will place data in a manner that is easily referenced by interested parties. An associated read me file will be placed in the master directory to aid in navigating the metadata.

Below is a description of a proposed metadata package that may be pursued as part of the Phase 2 design. The dataset numbering convention is intended to match the data flow callouts identified in **Table 2**. For example, “Dataset 1 Folder” corresponds to the dataset marked Sidewalk Data (Data Flow #1). In total, 12 folders will inventory all of the datasets collected as part of the project:

- **University of Washington ITS4US Deployment Project—Project Folder:**
    - Readme.txt—Discovery information for all datasets. Includes a link to the current DMP and the location of all datasets.
    - Project Attributes.csv—Static project data: geographic coverage area, participating data contributors, participating data aggregators.
    - Dataset 1 Folder (Sidewalk Data):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.
        - Datacontributor.csv—Table of submissions by data contributors.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.
    - Dataset 2 Folder (Graphed Sidewalk Data):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.
        - Dataprocessing.csv—List of dataset processes conducted on the data before delivery.
        - DataImpactLog.csv—Table of events that changed the dataset.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.
    - Dataset 3 Folder (Filtered Pathway Request Data—Sidewalks):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.

        - Dataprocessing.csv—List of dataset processes conducted on the data before delivery.
        - DataImpactLog.csv—Table of events that changed the dataset.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.
    - Dataset 4 Folder (Pathway Request Data—Sidewalks):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.
    - Dataset 5 Folder (Pathway Response Data—Sidewalks):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.
    - Dataset 6 Folder (Filtered Pathway Response Data—Sidewalks):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.
        - Dataprocessing.csv—List of dataset processes conducted on the data before delivery.
        - DataImpactLog.csv—Table of events that changed the dataset.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.
    - Dataset 7 Folder (Fixed-Route Transit Service, On-Demand Transit Service, and Transit Station Pathway and Attribute Data):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.
        - Datacontributor.csv—Table of submissions by data contributors.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.
    - Dataset 8 Folder (Validated Fixed-Route Transit Service, On-Demand Transit Service, and Transit Station Pathway and Attribute Data):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.
        - Dataprocessing.csv—List of dataset processes conducted on the data before delivery.

        - DataImpactLog.csv—Table of events that changed the dataset.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.
    - Dataset 9 Folder (Pathway Response Data—Transit Service and/or Stations):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.
    - Dataset 10 Folder (Pathway Request Data—Transit Service and/or Stations):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.
    - Dataset 11 Folder (Filtered Pathway Request Data—Transit Service and/or Stations):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.
        - Dataprocessing.csv—List of dataset processes conducted on the data before delivery.
        - DataImpactLog.csv—Table of events that changed the dataset.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.
    - Dataset 12 Folder (Filtered Pathway Response Data—Transit Service and/or Stations):
        - Readme.txt—Discovery information for this dataset.
        - Dataschema.csv—Table of all the fields in the dataset.
        - Dataprocessing.csv—List of dataset processes conducted on the data before delivery.
        - DataImpactLog.csv—Table of events that changed the dataset.
        - License.txt—License information for the dataset collected for the project. Depending on the data, this could have a unique license separate from that of the rest of the project.

### 5.3.3. Metadata Update Process

The goal for this project is to adopt a standard metadata set that will be used throughout the duration of the project to avoid any discrepancies in data that are reported. If a metadata update is necessary (adding, removing, or changing), the following process will be followed:

1. A memorandum will note the proposed changes to the metadata record with supporting reasons for making the change. This memorandum will be added to the project record.

2. Upon concurrence from project sponsors, the metadata record(s) will be adjusted in accordance with the proposed changes, with the noted date of change. This may be done in the data repository using an “updateMetadata” function. Approved API users—who would have had to submit contact information in order to get API access, namely data contributors and application developers—would be notified of changes.
3. Updated metadata will be stored in the data repository moving forward.
4. Metadata changes will be reflected in an updated version of the DMP.

# Appendix A. Acronyms and Glossary

| Acronym | Definition |
|---|---|
| API | Application Program Interface |
| BCDR | Business Continuity and Disaster Recovery |
| CBI | confidential business information |
| ConOps | Concept of Operations |
| CSV | Comma-separated values |
| DOI | Digital Object Identifier |
| DOT | Department of transportation |
| DMP | Data Management Plan |
| FHWA | Federal Highway Administration |
| FTA | Federal Transit Administration |
| GIS | Geographic Information System |
| GTFS | General Transit Feed Specification |
| GTFS-Flex | General Transit Feed Specification for demand-responsive or paratransit service |
| GTFS-Pathways | General Transit Feed Specification for pathways linking together locations within stations |
| HTTPS | Hypertext Transfer Protocol Secure |
| IE | Independent Evaluation |
| IES | Institute of Education Sciences |
| IRB | Institutional Review Board |
| ITS JPO | Intelligent Transportation Systems Joint Programs Office |
| JSON | JavaScript Object Notation |
| LiDAR | Light detection and ranging |
| N/A | Not Applicable/Not available |
| NCES | National Center for Education Statistics |
| ODbl | Open Data Commons Open Database License |
| OSM | OpenStreetMap |
| PII | Personally Identifiable Information |
| QGIS | Quantum geographic information system |
| Taskar Center or TCAT | Taskar Center for Accessible Technology at the University of Washington |
| TDEI | Transportation Data Equity Initiative |
| TRAC | Washington State Transportation Center at the University of Washington |
| USDOT | United State Department of Transportation |
| UW | University of Washington |
| XML | Extensible Markup Language |

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-21-869

U.S. Department of Transportation