SMART COLUMBUS

Data Management Plan

for the Smart Columbus Demonstration Program

FINAL REPORT | AUGUST 22, 2019

Produced by City of Columbus

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.

Acknowledgement of Support

This material is based upon work supported by the U.S. Department of Transportation under Agreement No. DTFH6116H00013.

Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the Author(s) and do not necessarily reflect the view of the U.S. Department of Transportation.

Acknowledgements

The Smart Columbus Program would like to thank the following members for their support of Smart Columbus and their work on the Technical Working Group Policy Team.

Ty Sonagere, CoverMyMeds	Dennis Hirsch, The Ohio State University
Jeff Hunsaker, HMB	Keir Lamont, The Ohio State University
David Landsbergen, The Ohio State University	Mehmet Munur, Tsibouris & Associates, LLC
Doug McCollough, City of Dublin	Dorene Stupski, Marriott International
Peter Voderberg, State of Ohio	Kirk Herath, Nationwide Insurance
David Daniel, Nationwide Insurance	Charles Campisano, City of Columbus
Amanda Girth, The Ohio State University	Tom Harris, HMB
John Sohner, HMB	Jeff Kanel, Centric
Nick Nigro, Atlas Policy	Brian Nutwell, Honda
Jim Perry, CAS	Jack Maher
Schlaine Hutchins, CoverMyMeds	

The Smart Columbus Program would also like to thank the authors, reviewers and contributors to this Data Management Plan.

Mandy Bishop, City of Columbus	Andrew Wolpert, City of Columbus
Jodie Bare, City of Columbus	Ryan Bollo, City of Columbus
Brian King, City of Columbus	Scott Millard, HNTB
Sherry Kish, HNTB	Ram Boyapati, Battelle
Tammy Chellis, Accenture	Warner Moore, GammaForce
Jeff Kupko, Michael Baker International	

Abstract

This Smart Columbus Data Management Plan provides operational information for the use of data within the Smart Columbus Operating System. This data will be the central data store for all relevant program data within the Smart Columbus demonstration and will provide users data through an open data portal web interface as well as Application Programming Interfaces. Smart Columbus will use this plan as a guide to perform the necessary operations for optimum program functionality in addition to properly securing, backing up, maintaining, and sharing the data. The Smart Columbus team will also utilize this plan to support proper privacy procedures and guidelines as defined in the Smart Columbus Data Privacy Plan.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1. PURPOSE OF THE PLAN

The purpose of the Data Management Plan (DMP) is to document how the data within the Operating System will be added, made accessible, and/or stored within the Operating System. The DMP also details how the data will be created, captured, transmitted, maintained, accessed, shared, secured, and archived. The DMP provides oversight for all eight Smart Columbus projects. Project-level data management development teams have used and will continue to use the guidance of this plan to resolve project-level designs while helping to fill out the full scope of the DMP.

Due to the dynamic nature of the Operating System (datasets can be added while operating), some documentation and processes may not be applicable to specific datasets. These will be reviewed and addressed during data curation, when a dataset is initially being evaluated for addition to the Operating System. Additionally, the Operating System may contain dataset metadata for datasets that are hosted on other publicly available systems. In this case, some information and procedures may not be applicable to those datasets.

The DMP is an operational guide for managing the data within the program. The DMP details how and where data will be shared, subject to applicable privacy, security, and other safeguards, and how the data will be made available to others to enable performance measurement and support independent evaluation.

Although the Operating System has a primary focus of not storing personally identifiable information (PII); in instances where the data includes PII or other restrictions, the DMP relies on the strict handling of PII detailed in **Section 5.4** band in the Data Privacy Plan (DPP).

This DMP satisfies the requirements of Cooperative Agreement No. DTFH6116H00013 between the USDOT and the City of Columbus. It follows the USDOT guidance for creating Data Management Plans provided at https://ntl.bts.gov/public-access/creating-data-management-plans-extramural-research.

As mentioned above, this DMP will be updated regularly as the Operating System continues to mature and offer more features, functionality, and datasets.

1.2. ORGANIZATION OF THE PLAN

This DMP is organized into the following chapters:

- **Chapter 1. Introduction**
- **Chapter 2. References**
- **Chapter 3. Data Description and Management**
- **Chapter 4. Standards Used**
- **Chapter 5. Sharing and Protecting Data**
- **Chapter 6. Re-Use, Redistribution, and Derivative Products Policies**
- **Chapter 7. Archiving and Preservation Plans**
- **Appendix A. Acronyms and Definitions**
- **Appendix B. Glossary**

1.3. PROGRAM DESCRIPTION

In 2016, the U.S. Department of Transportation (USDOT) awarded $40 million to the City of Columbus, Ohio, as the winner of the Smart City Challenge. With this funding, Columbus intends to address the most pressing community-centric transportation problems by integrating an ecosystem of advanced and innovative technologies, applications, and services to bridge the sociotechnical gap and meet the needs of residents of all ages and abilities.

With the award, the City established a strategic Smart Columbus[1] program with the following vision and mission:

- **Smart Columbus Vision:** Empower residents to live their best lives through responsive, innovative, and safe mobility solutions.
- **Smart Columbus Mission:** Demonstrate how Intelligent Transportation Systems (ITS) and equitable access to transportation can have positive impacts on every day challenges faced by cities.

To enable these new capabilities, the Smart Columbus program is organized into three focus areas addressing unique user needs; enabling technologies, emerging technologies, and enhanced human services. This portfolio of technical concepts was divided into eight projects as shown in **Figure 1: Smart Columbus Framework**.

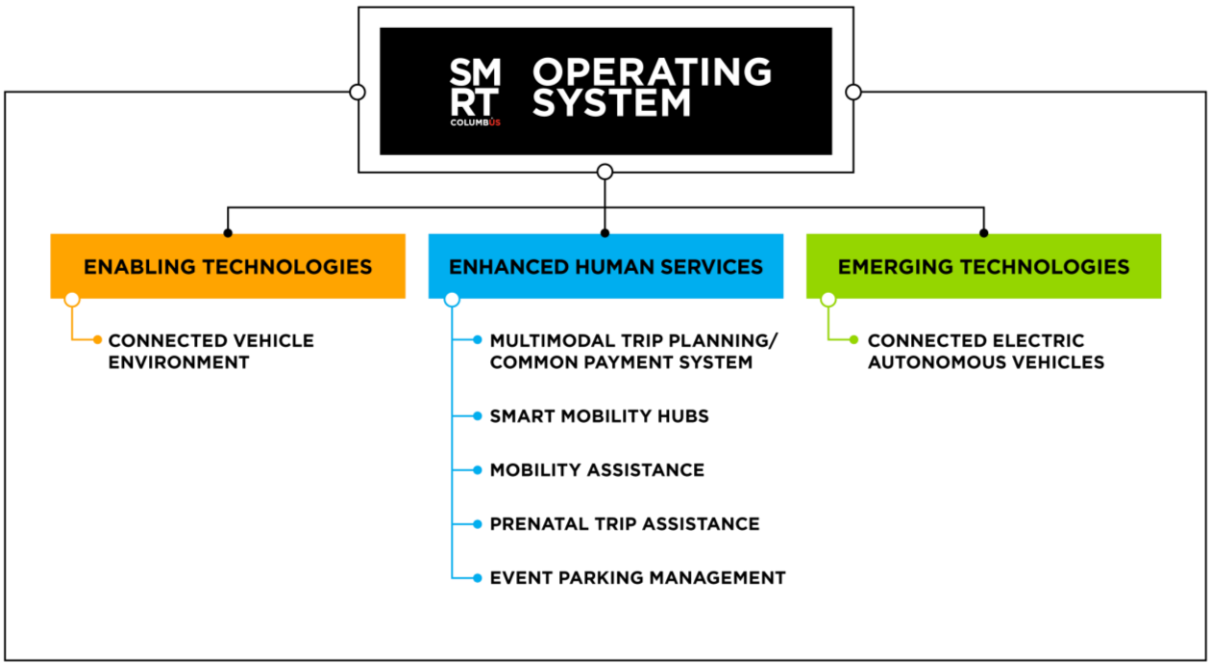

Figure 1: Smart Columbus Framework

Source: City of Columbus

[1] *While the City of Columbus Smart Columbus Program Office has grown to oversee many innovation initiatives, the scope of this document is any data that is in the Operating System or that is in any of the other USDOT-funded projects. The City of Columbus USDOT-funded Smart Columbus program will be known throughout this document as Smart Columbus.*

The Columbus Smart City Demonstration Projects are:

- **The Smart Columbus Operating System (Operating System)**

  The Operating System is the essence of Smart Columbus – it brings to life the innovation. The Operating System is being designed and built to collect data from a variety of inputs; including public, nonprofit, education-based, and private sector contributors. These inputs may come from other systems, devices, and people. All of which are a critical part of building this ecosystem of innovation. Data will be available for analytics and visualization as well as for artificial intelligence required by various smart city applications. The Operating System is a platform designed for big data, analytics, and complex data exchange. It will capture the data and provide a means for multitenant access to aggregate, fuse, and consume data.

  Datasets housed in the Operating System include the Smart Columbus demonstration projects, traditional transportation data, and data from other community partners, such as food pantries and medical services. The Operating System will be scalable and will demonstrate the potential for serving city and private sector needs well beyond the life of the Smart City Challenge award period.

- **Connected Vehicle Environment (CVE)**

  Cars, trucks, and buses will talk to the infrastructure and talk to one another to reduce traffic congestion and increase safety. The CVE will connect 1,500-1,800 vehicles and over 90 smart intersections across the region. Safety applications are intended to be installed on multiple vehicle types including transit buses, first responder vehicles, city and partner fleet vehicles, and private vehicles. Applications will be deployed to ensure emergency vehicles and the Central Ohio Transit Agency (COTA) Bus Rapid Transit (BRT) fleet can utilize signal prioritization when needed to ensure safety and efficiency.

- **Multimodal Trip Planning Application (MMTPA)/Common Payment System (CPS)**

  The MMTPA will provide a robust set of transit and alternative transportation options including routes, schedules, and dispatching possibilities. The application will allow travelers to request and view multiple trip itineraries and make reservations for shared-use transportation options such as bike-sharing, ride-hailing and car-sharing. Users will be able to compare travel options across modes, and plan and pay for their travel based upon current traffic conditions and availability of services.

  The CPS will serve as an account-based, back-office payment processor for the MMTPA and EPM applications. To facilitate integration with both applications, the CPS will provide landing pages and Application Programming Interfaces (APIs) allowing travelers to manage CPS accounts and issue payment requests for transportation and parking services. Requests for payment will flow through a payment broker microservice, which will be responsible for directing payment requests to the CPS back office, communicating payment status to the applications, and for capturing anonymous trip and payment data for use in analytics and performance measurement. The CPS back office will be compliant with Payment Card Industry (PCI) Data Security Standards (DSS), ensuring the security and confidentiality of PII.

- **Smart Mobility Hubs (SMH)**

  Smart Mobility Hubs will be deployed to serve travelers' needs more effectively by expanding transportation resources and offering access to comprehensive trip planning tools at designated locations. SMH sites are primarily located adjacent to existing COTA CMAX and transit center facilities and will help bridge the first-mile/last-mile gap between transit and destination by providing physical space for the consolidation of services such as bike/scooter share, car share, and ride-hailing. Interactive kiosks and public Wi-Fi will be made available to the traveler to view real-time travel information and to book multi-modal trip plans via the MMTPA/CPS.

- **Mobility Assistance for People with Cognitive Disabilities (MAPCD)**

  The City will deploy an innovative smartphone application for people with cognitive disabilities to transition off costly paratransit services and travel independently on the fixed-route bus system. The application will be piloted for between 15 to 30 individuals in the Columbus region in partnership with the Central Ohio Transit Authority (COTA) and The Ohio State University (OSU). The application will include a highly accurate, turn-by-turn navigator designed to be sufficiently intuitive such that older adults and groups with disabilities including the cognitively disabled can travel independently.

- **Prenatal Trip Assistance (PTA)**

  The City will develop a system for providing flexible, reliable, two-way transportation to expectant mothers using Medicaid Managed Care Organization brokered non-emergency medical transportation services.

- **Event Parking Management (EPM)**

  The City will integrate parking information from multiple parking facilities into a single availability and reservation services solution. This will allow travelers to search for and reserve parking in advance or on the go. More direct routing of travelers during large events is expected to reduce congestion during those times.

- **Connected Electric Autonomous Vehicles (CEAVs)**

  CEAVs that operate in a mixed-traffic environment interacting with other vehicles, bicyclists, and pedestrians will be deployed. The project provides an accessible and easily expandable first-mile/last-mile transportation solution to the region by deploying a fleet of multi-passenger CEAVs that will leverage the enhanced connectivity provided by the CVE and the citywide travel planning solution.

1.4. CORE FUNCTIONS OF THE OPERATING SYSTEM

Figure 2: Core Functions of the Smart Columbus Operating System depicts high-level system elements of the Operating System.

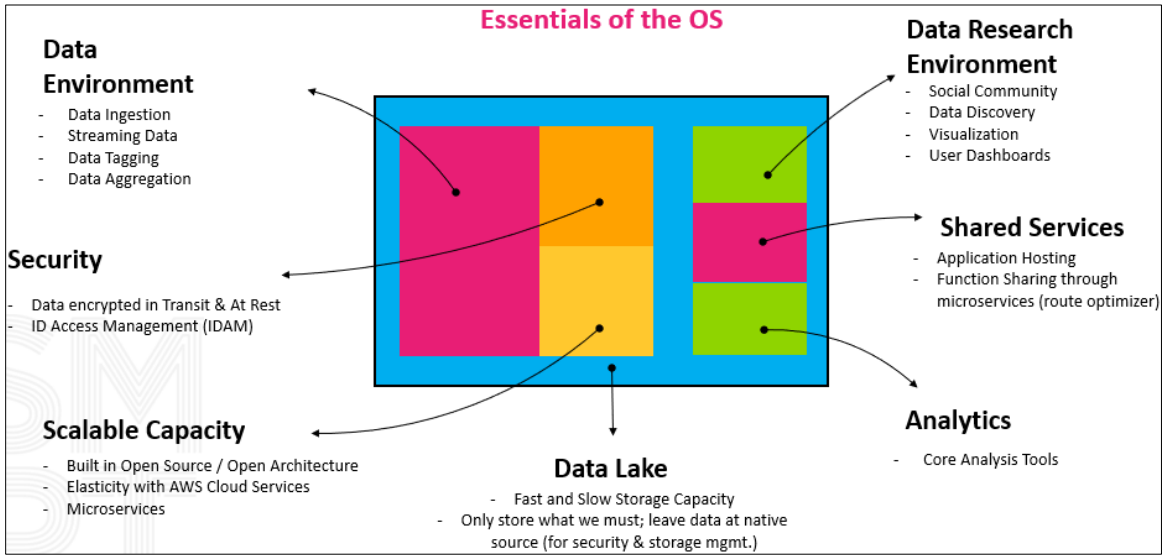

Figure 2: Core Functions of the Smart Columbus Operating System

Source: City of Columbus

The Operating System is a platform for smart cities development and operation. It consists of several core functions, which can be leveraged across the Smart Columbus program, as well as other functions (defined below) that will specifically enhance and support the "Smart Applications." The Operating System has the following core functions:

- **Data Environment:** The orderly ingestion, aggregation, and tagging of many forms of data from real-time, to slow-moving, or manually uploaded data.
- **Data Lake:** A storage repository that holds a massive amount of raw data in a secure way and makes it available to all the other supported operations in the system.
- **Security:** To ensure trust, it is imperative that the Operating System is exceptional at managing the users and systems that have access to it.
- **Scalable Capacity:** The Operating System is "scalable" and "elastic" which means that it can grow and shrink to meet the demand of the system at any given time.
- **Shared Services Environment:** Application components can be housed and made available to any number of applications connected to the Operating System.
- **Data Research Environment:** In a data-rich environment, Columbus and its residents, businesses, nonprofits, and visitors will be increasingly able to share, use, and leverage previously unavailable datasets to address complex problems and improve current operations and capabilities.
- **Business Analytics:** Analytics will also be used to predict future conditions and the potential benefits of implementing different operational strategies, control plans, and response plans coordinated among agencies with mobility providers.

1.5. SYSTEM OF SYSTEMS OVERVIEW

The Smart Columbus program has many interrelated systems that work together to provide a System of Systems (SoS). Information from these systems is shared in the Operating System. Both real-time and archived data is maintained in the Operating System for use by other Smart Columbus projects and future applications. The SoS provides Smart Applications, Smart Vehicles, and Smart Infrastructure to travelers in the Columbus area. The Operating System enables the SoS to share data with many other external systems to provide the framework for the services provided.

Figure 3: System of Systems External Context Diagramshows the relationship of the SoS to the external travelers and systems.

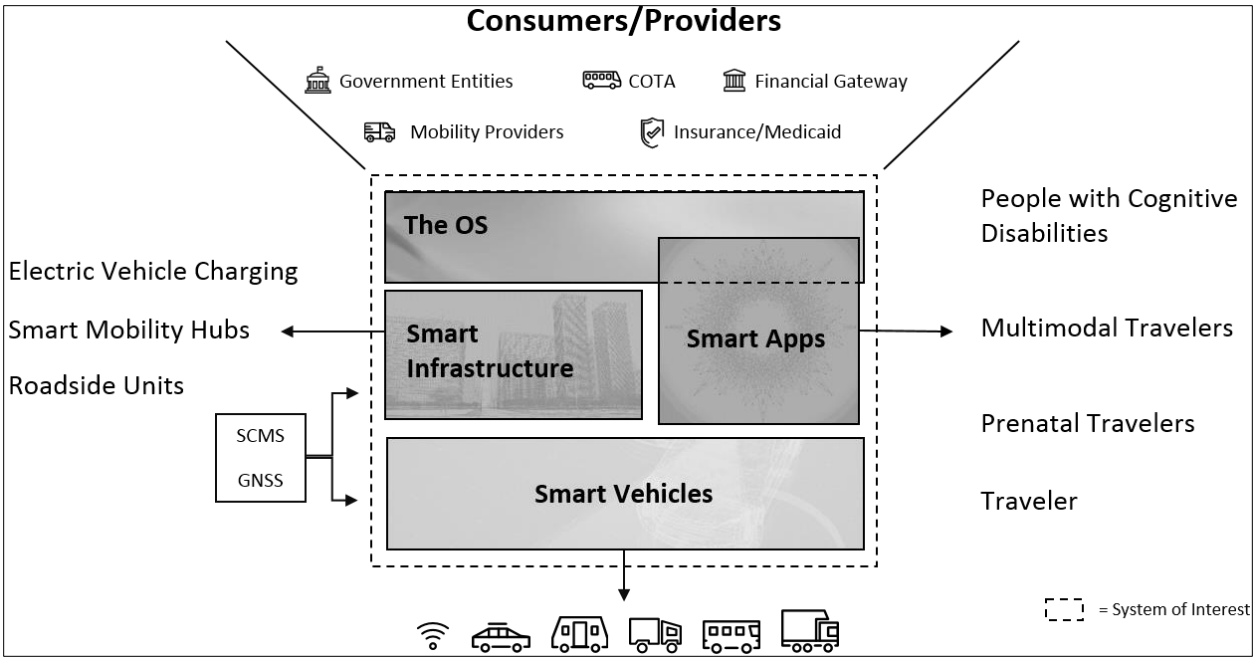

Figure 3: System of Systems External Context Diagram

Source: City of Columbus

The Smart Infrastructure element contains the roadside units, hubs, and corresponding network that enable interactions between these items and the Operating System. Smart Vehicles include the onboard units that will be installed in vehicles and include various vehicle types. Smart Applications include the software-oriented solutions that will deliver other Smart Columbus project capabilities such as multimodal trip planning, common payment, and prenatal trip assistance. The Operating System is the repository for all Smart Infrastructure and Smart Vehicles performance data as well as the shared services platform; allowing the Smart Applications to be directly integrated.

Chapter 2. References

Table 1: References lists documents and literature referenced during development of this document.

Table 1: References

Document Number	Title	Revision	Publication Date(s)
N/A	Creating Data Management Plans for Extramural Research, National Transportation Library https://ntl.bts.gov/public-access/creating-data-management-plans-extramural-research	N/A	3/22/2018
N/A	National Association of City Transportation Officials (NACTO) Policy 2018 https://nacto.org/wp-content/uploads/2018/02/NACTO-Policy-2018.pdf	N/A	March 2017
	Smart Columbus Data Privacy Plan, Internal Document, Drafts	Yes	4/19/2018, 9/28/2018, 12/31/2018,1/29/2019
	Smart Columbus Systems Engineering Management Plan	N/A	1/16/2018
N/A	Project Open Data Metadata Schema v1.1 https://project-open-data.cio.gov/v1.1/schema/	Yes	11/06/2014
N/A	Open Data Policy, Project Open Data https://project-open-data.cio.gov/	N/A	May 2018
N/A	M-13-13 Memorandum for the Heads of Executive Departments and Agencies: Open Data Policy – Managing Information as an Asset https://project-open-data.cio.gov/policy-memo/	N/A	May 2018
N/A	Ohio Revised Code § 149.43: Availability of public records for inspection and copying	N/A	12/19/2016
N/A	Ohio Revised Code § 1306.01: Definitions	N/A	9/14/2000
N/A	International Organization for Standardization ISO 8601	N/A	1988
N/A	Smart Columbus De-identification Policy	N/A	2019

Source: City of Columbus

Chapter 3. Data Description and Management

Data within the Operating System will be made available in different formats such as Comma-Separated Values (CSV), Extensible Markup Language (XML), JavaScript Object Notation (JSON), and many others. Datasets are added by the Operating System product team through a data curation process.

Due to the dynamic and ongoing nature of the data curation process, the number of datasets and resources managed and made available by the Operating System will constantly change. When adding data to the Operating System, datasets will need a description and other metadata from both the dataset and the elements that make up the data. Metadata adds important value to data to allow humans and machines to properly understand what information the dataset holds and to improve the usefulness of data. In addition to the metadata, a data dictionary provides column-level details about a dataset and provides guidance on their interpretation, accepted meanings, and representation. The metadata included in a data dictionary can assist in defining the context, scope, and characteristics of data elements, as well the rules for their usage and application; helping humans and machines better understand a given dataset.

Data will be categorized upon ingestion as either restricted or public data. Restricted data is any data that contains PII, is held for independent evaluation per the USDOT Cooperative Agreement or is held based on a private entity's contract with the Operating System. Public datasets are those datasets that are available for public consumption and contain no PII.

Smart Columbus is creating an inventory of the data that is or will be collected through the Smart Columbus program. The inventory will contain the following information about each dataset: name, type, source, responsible party for maintenance, collection approach, frequency and period of collection, expected users, value of the data, whether the data contains PII and where the data will be located. All projects are not live so this inventory will represent a snapshot in time and will be updated in one year when all projects are expected to be actively collecting data.

3.1. METADATA

The Operating System has been built to conform with the Federal Project Open Data Metadata Schema v 1.1.[2] While this metadata schema is widely adopted and based upon Data Catalog (DCAT)[3] standards, the Operating System has the capability of extending the metadata to contain additional fields that may be enforced on a per-dataset basis, on a dataset level, or on a system level. Metadata can be applied at the dataset, resource, and data element levels. A resource is part of a specific dataset that can be consumed and made available in multiple formats (CSV, JSON, XML, etc.). A resource is comprised of one or more data elements. This represents the minimum capability of the system; over the course of the program additional metadata standards may be supported.

Metadata related to the dataset shall, at a minimum, include the following fields in addition to other system-level metadata (ID numbers, etc.):

- Dataset Project Open Data Fields
  - Title
  - Description

[2] https://project-open-data.cio.gov/v1.1/schema/

[3] https://www.w3.org/TR/vocab-dcat/

  - Tags (keywords)
  - Modified date
  - Publishing organization
  - Point of contact name and email address
  - Public access level
  - License
  - Rights – detailed information if access level is not “public”
  - Spatial applicability of the dataset
  - Temporal range of the dataset (start and end dates)
  - Distributions (details of the resources)
  - Frequency – how often the dataset is published
  - Data standard – Uniform Resource Locator (URL) to a format the dataset would conform to
  - Data dictionary – URL to a data dictionary
  - Data dictionary type – file format of the data dictionary
  - Collection – if the dataset is part of a collection
  - Release date – date of formal issuance
  - Language – language of the dataset
  - Homepage URL – URL of a page related to the dataset
  - Related documents – URLs to documents that are related
  - Category
- Resource Metadata
  - Date last updated
  - Date metadata updated
  - Created date
  - Format (CSV, JSON, XML, etc.)
  - License

The Operating System project team is responsible for populating the metadata fields for datasets and every effort will be made to ensure that they are populated and correct during the data curation process. To comply with the Project Open Data metadata standards, the Operating System will provide a JSON formatted catalog of all datasets and resources in the *data.json* file located on the root directory of the Operating System website: https://www.smartcolumbusos.com/data.json.

3.2. DATA DICTIONARY

A data dictionary provides machine-readable detailed information for a dataset and its columns. For data shared with the Operating System, it is preferred that a data dictionary be provided by the source data provider in a machine-readable form such as CSV. The Operating System will house many datasets from many diverse sources, each will have its own data dictionary, for which the URL is provided in the metadata of the resource. If a data dictionary exists for a dataset, the URL will be in the "describedBy" field of the dataset metadata if accessed through the API and displayed in the "Data Dictionary" field if viewing through the Operating System website (as prescribed by the Federal Project Open Data Metadata Schema).

During the initial data curation process for datasets in which a data dictionary is unable to be provided, a data curator will work with the source data provider to develop a data dictionary. This will entail methods around creating data dictionaries such as defining business terms and relating them to column names, detailing the values of any codes, and providing detailed information on what each column represents. For example, if columns from different tables are *address_1, one_address, first_line_address, user_address*, and *company_address*, and if they all are used the same way with the exact same definition, the business term might be "Company Address." Using the same column names, if the data constituted a person instead of a company, "Personal Address" would be a business term for *user_address* and the rest would remain the under the Company Address classification.

To ensure consistency in data dictionary fields, mapping to a standard such as the National Information Exchange Model[4] will be performed upon ingestion if there is an appropriate standard that would apply to the data. This will help support the data communities' development within the Operating System.

3.3. SIZE AND SCALE OF DATA

Because the number of datasets located in the Operating System will continually grow as more Smart Columbus projects are brought online and additional datasets are identified, determining the size and scale of the data will be an ongoing effort. The Operating System project team will evaluate current usage statistics monthly and make determinations regarding whether cloud data storage space needs to be scaled up. The cloud-based architecture of the Operating System allows storage space to remain flexible as necessary.

Additionally, when a new project is ready to come online and provide data into the Operating System, an initial analysis will be performed by the Operating System project team that will estimate the amount of data by estimating the number of data points, the data frequency, and individual entry sizes. The Operating System product team will work closely with the other project teams to ensure all data sizing and velocity needs are met so that the Operating System remains stable and can ingest the amount of data necessary to meet program goals.

Data that is continuously streamed into the Operating System will be added to the data storage for that project and will continually be monitored for total space usage. During ingestion, a strategy will be defined that controls the size of the data. For example, instead of saving all events, it may be decided to only keep the latest event. Also, established policies will decide whether to purge everything older than a set number of days. As the Operating System will include a data tracking and grading system, datasets that go unused or receive low performance scores will be considered for deletion. This decision will be made on a dataset by dataset basis depending upon the use cases supported.

[4] *National Information Exchange Model, or NIEM, https://www.niem.gov/*

3.4. DATA ACQUISITION/CREATION

There are three primary methods of data acquisition within the Operating System:

1. A manual request submittal through the Operating System website at https://www.smartcolumbusos.com/share-your-data, or
2. The Operating System project team and data curators identify datasets that would be beneficial to the program, or
3. Other Smart Columbus project team(s) identify data needs or data to share and bring to the attention of the Operating System project team.

After a data-sharing request has been made, a formal data curation process (see **Figure 4: Data Ingestion Workflow**) will be conducted by the Operating System project team during the current configuration of the Operating System. This process involves working with the source provider to determine what data can be shared, validating if it contains any PII, what formats would be best suitable, what metadata is available, and identifying any additional resources such as data dictionaries or web pages that would help users better understand the data. Special consideration will be given to the risk that a dataset may represent as a method of re-identifying persons through the combining of the ingested dataset with other datasets in the system.

The Operating System project team will then determine and develop the method of ingestion for the data and metadata. Data will initially be processed within an isolated environment to be validated and approved. Once validated, the process continues, and data will be moved to the production website where the data will be made available to users of the Operating System.

Data-sharing requests may be prioritized based on various factors during the data curation process:

- Value and relativity of data to Smart Columbus projects
- Current availability of data
- Complexity of ingestion into the Operating System, including de-identifying any PII data
- Size of data

Figure 4: Data Ingestion Workflow

Source: City of Columbus

3.5. PERFORMANCE MEASUREMENT DATA

Smart Columbus has a Performance Measurement Plan (PfMP) that details out how the Operating System and each project will measure their performance. Specifically, the Operating System will conduct user surveys to identify the use and usability of the data.

The Operating System will be used to house, publish, and distribute Smart Columbus program performance measure data and further categorize the data as public or restricted. Independent Evaluators (IEs) and other authorized users require access to data collected specifically pursuant to the PfMP. This data will be provided through the Operating System user interface and through an API to query the data. Data that is not ingested by the Operating System will be uploaded directly to the Secure Data Commons (SDC) for IE usage. Unauthorized users will be denied access to the restricted data through the user interface or through an API.

3.6. EXTERNAL INGESTION INTERFACES

External interfaces provide a method for other projects within the Smart Columbus demonstration to interface with the Operating System to send data to the Operating System. The Operating System contains multiple external interfaces for data ingestion: the API, secure file transfer protocol (SFTP) and batch upload.

- **API:** The Operating System ingests data through PUSH and PULL API services.
  - **PULL APIs:** An external vendor creates an API and feeds the project data into it. The Operating System will PULL the project data from this API.
  - **PUSH APIs:** The Operating System creates an API. External vendor PUSHES the project data into this API.
- **SFTP:** An external vendor creates a Secure File Transfer Protocol and provides credentials (Username, password, IP address, and File Location) to the Operating System. The Operating System sets up a Cron job and pulls data from this site.
- **Batch upload:** An external vendor sends batches of data through an email, FTP or other data sharing mechanism. The Operating System manually ingests this data. (This is not used for project data but for acquiring key performance indicators.)

3.7. FREQUENCY OF DATA COLLECTION

During the data curation and discovery process, the Operating System data curator will work with the source data provider to determine how often the data is collected, updated, and published. This information will determine what methods are used to get the data and metadata into the Operating System. This can be defined down to near real-time. A metadata field, frequency, describes how often the data is updated. The metadata fields identified in **Table 2: Data Collection Frequency Metadata Fields** provide detailed information about the frequency of data collection.

Table 2: Data Collection Frequency Metadata Fields

Field Name	Description
Frequency	How often the dataset is published
Last Updated	Date/Time that the data was last updated. ISO-8601[5] repeating duration
Temporal	Start and end date for data
Release Date	Date of issuance

Source: City of Columbus

3.8. RELATIONSHIP OF NEW DATA TO EXISTING DATA

The Operating System is configured such that data updating strategies can be defined by the data curator. Common strategies include file replacement, upsert, dedupe, and merge and rolling sync. The action taken on a dataset may be recorded in the metadata.

3.9. CONCEPT OF ORGANIZATIONS

The Operating System uses the concept of Organizations as a key concept for grouping of datasets to an entity that provides the dataset. This enables users of the Operating System to use the source entity's name as a filter when querying for datasets. The Operating System project team can add new Organizations through the data curation process and group datasets together from that source entity so that related entity metadata will automatically be provided with the data. This allows the data, although similar, to be managed separately when coming from various sources. The Operating System and API would allow users to query similar data across multiple Organizations to create a "mash-up" of diversely sourced data. While the concept of Organizations within the Operating System will stay constant, the implementation and definition of it may change as future requirements are defined and the projects mature.

Examples of Organizations within the Operating System include:

- Ohio Geographically Referenced Information Program (OGRIP)
- Mid America Association of State Transportation Officials (MAASTO)
- City of Columbus
- Central Ohio Transit Authority (COTA)
- Ohio Department of Transportation (ODOT)

3.10. DATA USERS

Users of the Operating System are diverse; users include community members, technology developers, researchers, companies, government entities, and others. The Operating System will be designed to allow all types of users easier and better access to data for use and analysis. After users have access to the data and APIs, they can use the data to create web applications, or import into many different data analysis and/or visualization tools. The usage pattern will vary from data source and metadata discovery to flat file downloads and on-demand API requests.

[5] *International Organization for Standardization*

It is expected that each type of user will have different habits and use cases with the Operating System. For example:

- **Community members:** General interest and browsing available datasets for short periods of time using online visualization tools
- **Technology developers:** Targeted interest in specific datasets, continued or repetitive use of the same datasets through APIs or download links
- **Researchers and project evaluators:** Downloading datasets or querying through the provided API for either local analysis or analysis in the online analytics environment

Data will be available to the IE through the SDC and the Operating System website.

3.11. VALUE OF THE DATA

The Operating System is focused on providing datasets that are valuable to users. While Smart Columbus will survey the users on their personal value of the data and also track the usage numbers for each dataset, Smart Columbus understands that value is a subjective measure for each user of the Operating System. Therefore, Smart Columbus focuses on creating valuable data by focusing on ensuring that the data is as usable as possible.

During the data curation process, the data curator works with the source entity to provide as much information as possible along with the dataset so that it has value to users.

3.12. ROLES RESPONSIBLE FOR DATA MANAGEMENT

The Operating System has five roles that are responsible for the management of the data. During project development and operation, all roles except data provider will be performed by members of the Operating System project team. Data providers are members of the entities providing the data to the Operating System. Below are some of the roles that will be responsible for data management.

- **System Administrators:** Responsible for the upkeep, configuration, and reliable operation of the Operating System to maintain the integrity and availability of the data.
- **Data Curators:** Involved with the design and integration between the Operating System and entities that contribute data. Ongoing efforts to validate data and usage and improve datasets and relationships with providers.
- **Data/System Architects:** Responsible for the design and integration of all system back-end components used to ingest data.
- **Data Providers:** Responsible for working with the Operating System to ensure their entities' data is valid and compliant.
- **Privacy and Security Board:** Advise Smart Columbus on privacy and security issues.

The Organizations concept within the Operating System allows authorized users of an entity to act as the data provider for the data the entity provides. Authorized users can manage datasets, metadata, and the data hosted within the Operating System.

3.13. MANAGEMENT AND AUDIT CONTROLS

Data management and audit controls are important aspects of the Operating System. Data curators and administrators are primarily responsible for the management of datasets. The governance of the inbound content of remotely sourced datasets will be done by contract or signed agreement. The Operating System logs each action to configure datasets or load data. The Operating System will store logs in a separate read-only data store to control the integrity and validity of the logs. These factors provide integrity controls for the data and metadata.

For datasets on the OS that are remotely linked to the data provider source, the provider of the data carries the sole responsibility of ensuring compliance with data privacy and de-identification. For the purposes of this document, remotely linked datasets are referred to as datasets that are not hosted on the OS but are linked to the source location of the data.

When a dataset underperforms, the dataset statistics maintained by the OS can be used to initiate further investigation. The data provider will be contacted and asked to remediate any anomalies detected. In the event of persistent underperformance or irresponsible inclusion of private data, datasets may be removed from access or deleted.

Chapter 4. Standards Used

4.1. DATA FORMATS

The Operating System will only use platform-independent and nonproprietary formats to focus on machine-readability of the data. To accomplish this, it is preferred that any data source that is in a non-machine-readable format will be converted to a different format during the data ingestion design process. A sample, non-comprehensive, list of machine-readable and non-machine-readable formats is provided in **Table 3: Sample File Types**.

Table 3: Sample File Types

Machine-Readable	Non-Machine-Readable
JSON	PDF
XML	JPG
CSV	TIFF
RDF	MP4
	WAV

Source: City of Columbus

In addition to the data provided, metadata must be provided that complies with the metadata standards defined in **Section 4.3**.

Datasets that are available in other publicly available systems can be read into the Operating System if the system publishes a metadata catalog that is compliant with one of the following formats: Comprehensive Knowledge Archive Network (CKAN), DCAT Vocabulary, DCAT Resource Description Framework (RDF), or Project Open Data.

Due to the variety of projects that will be deployed under the demonstration, there may be many data standards and formats that are ingested into the Operating System. Correspondingly, multiple formats will be used for data ingestion and the workflow of data ingestion will be tailored to the required data type and elements. During ingestion, prioritization may be given to automated processes, specific data sources, and/or validated files. Since data from various sources and formats are anticipated, coordination across projects is recommended. Data providers will register with the Operating System and after approval and agreement to terms of service (or other required documentation), submit prepared datasets. No data will be accepted anonymously. Ingestion may be manual or API, which will be decided by the Operating System project team during the design phase of the data ingestion process.

4.2. COMMUNICATING ABOUT THE DATA

The main method of describing datasets will be through the dataset's associated data dictionary and related documents. The two fields in the metadata that describe these are "Data Dictionary" and "Related Documents." These two metadata fields can contain hyperlinks to resources that help define and describe the data and any other useful resources that are available to users. These resources will be discovered or defined during the data curation process.

In addition to the metadata fields, data stories can be published on the Operating System website that contain information about a dataset, a potential problem or need that might be solved, and information on how to consume the data using the Operating System API. It is the intention that these data stories will help spark discovery of data services within the Operating System to support applications to help respond to specific community needs.

4.3. METADATA SCHEMA, STORAGE, AND MANAGEMENT

Datasets within the Operating System will comply with the Project Open Data Metadata Schema v1.1. This schema is a standard defined and used by the U.S. Government and is extensible to include other necessary fields. The Operating System complies with all Project Open Data requirements for its catalog and datasets. The Operating System extends the metadata to include other common metadata fields that are populated when the dataset is first scheduled for ingestion.

4.4. DATA CONSUMPTION METHODS

Users of the Operating System will consume data in multiple ways:

- Downloading a file that contains the dataset (various formats will be available depending on the dataset)
- Viewing structured data with an in-browser previewer/viewer
- Consuming the API to query the data
- Analyzing/visualizing data through analysis tools like JupyterHub

Because all data is desired to be in a non-proprietary format, once retrieved, data will be able to be used within many different tools as needed. Users do not need to register a user account to interact with the data or API.

At this time, there is not a limit on the amount of data a given user can request through the website, but API requests will have request throttling set based on originating Internet Protocol (IP) address to prevent overloading the system with requests. The Operating System will set a public use standard that everyone complies with unless under separate contractual agreement with the City. The Operating System will be managing by exception with contracts.

Metadata consumption is subject to the same controls as the content of the dataset.

4.5. QUALITY CONTROL MEASURES

All data providers, including project and Operating System data providers, are responsible for their own quality controls which, when applicable, are expected to conform with relevant industry standards, such as the American Society for Testing and Materials' (ASTM) Quality Control Standards.[6]

Specifically, submitted project and Operating System data should possess or undergo the following:

- An identified, authenticated submitting entity
- All available metadata about the dataset
- A provenance plan based upon SLAs

[6] https://www.astm.org/Standards/quality-control-standards.html

- Initial ingestion process that will identify any high-risk data, such as PII, PCI, or Personal Health Information (PHI)

The Operating System will determine freshness, completeness and validity of the data initially and periodically as defined in the associated SLA. This will provide a score that will be appended to the data page. Usage statistics will also be generated automatically. Validation rules may be used to flag unusual entries. Should outliers, missing or otherwise anomalous entries be found, the data providers may be contacted for verification. Flagged data may be changed to show null value or may be rejected in whole.

Chapter 5. Sharing and Protecting Data

5.1. SHARING DATA

Public data that has been vetted through the Operating System project team and undergone the data curation, design, and ingestion processes will be shared publicly and available to all users. Once data is identified that needs access control and authorization mechanisms, the appropriate controls will be put into place – specifically when other Smart Columbus programs are ready to begin sharing data through the Operating System. The Operating System has a concept of restricted datasets with built in authentication which can be used. To access restricted datasets, authentication, and authorization will be required for the user interface and the API.

5.2. USER AUTHENTICATION

Data within the Operating System is categorized as restricted or public. These concepts are further defined in the DPP.

If a dataset is marked as restricted, a user must authenticate and have privileges to gain access to read the data.

The classes of users within the Operating System define what datasets they have access to as described below. As additional datasets are brought into the Operating System, additional user classes may be defined to meet requirements.

- **Unauthenticated User**: Access to all public datasets with API limits
- **Authenticated User**: Access to all public datasets, enhanced datasets, and restricted datasets to which organization administrators have provided access through role-based access control

5.3. CONCERNS WITH SHARING

The Operating System is built with a foundation of sharing – sharing data within the system – from sending links to datasets, to building proof of concept applications, to analyzing data for a research paper or article. This open nature of the Operating System means that the risk of sharing the data is minimized because of the thorough process during data curation to limit confidential information from being ingested. This does not eliminate the risk of private person re-identification as the complexity of assessing all datasets against one another for the presence of a re-identification risk is difficult to nearly impossible given the systems and processes currently available. The operation will rely upon the curation process to maintain an active oversight of such risks.

5.4. DE-IDENTIFYING DATA

During data curation of datasets, the data will be evaluated to see whether it contains PII. If a dataset is found to contain PII, the dataset will be rejected and sent back to the data provider for de-identification based on the Smart Columbus De-identification Policy.

Chapter 6. Re-Use, Redistribution, and Derivative Products Policies

6.1. PERMISSIONS TO MANAGE DATA

The data within the Operating System is generally collected outside of the Operating System before it is ingested into the Operating System. Therefore, in most cases, the right to manage the data belongs to the source entity contributing the data. During data curation and with the acknowledgement and review of the data provider, the Operating System project team can manipulate the data only if it does not change the meaning of the data – for example to remove duplicate entries, to change file formats, to de-identify, or add proper geotags to the data. The Operating System project team functions purely in an administrative role and they can move, archive, index, or perform other maintenance on the data to promote accessibility, performance, security, or reliability of the data.

6.2. INTELLECTUAL PROPERTY OWNER OF DATA

The Operating System data complies with Ohio Revised Code section 149. In general, the Operating System and the City of Columbus serve as data providers for the Operating System data. Any entity submitting public data must agree to release any claim of ownership or fees when providing data to the Operating System. Private entities may retain ownership of the data by contractual agreement.

In Ohio, public records are governed by Ohio Revised Code section 149.011 (G):

> *"…any document, device, or item, regardless of physical form or characteristic, including an electronic record as defined in section 1306.01[7] of the Revised Code, created or received by or coming under the jurisdiction of any public office of the state or its political subdivisions, which serves to document the organization, functions, policies, decisions, procedures, operations, or other activities of the office."*

Data provided from governments, businesses, and agencies may not be records that serve "to document the organization, functions, policies, decisions, procedures, operations, or other activities of the" City of Columbus. Therefore, that data will not be held to the same public records disclosure and retention standards by the City of Columbus.

For data contained in the Operating System provided by the City of Columbus, the City department who shares the information will serve as the owner of that public record and will therefore have all duties under Ohio law to maintain, retain, and disclose those documents per Ohio law. Users of the Operating System will be able to access the owning department and contact information in the dataset's metadata under the "Publisher," "Contact Name," and "Contact Email" fields.

6.3. COPYRIGHTS TO DATA

Providers of public data will certify that all data provided is free of any copyright or other obligation by indicating the data can be published under an open license. All public data within the Operating System will remain free of copyright. Private entities may retain copyrights of the data by contractual agreement.

[7] http://codes.ohio.gov/orc/1306.01v1

6.4. TRANSFER OF RIGHTS

Data rights will not be transferrable.

6.5. DATA LICENSING AND REDISTRIBUTION

The Operating System will be able to retain public and restricted data. Public data housed within the Operating System will be required to be under an "Open License" which, in general, requires the following conditions:

- **Re-use**: Allow for reproductions, modifications, and derivative works
- **Redistribution**: No restriction on selling or giving away
- **No Discrimination:** Must not discriminate against any person or group of persons

Examples of open licenses that can be used on the Operating System are:

- Creative Commons Attribution[8]
- Creative Commons Attribution Share-Alike[9]
- Creative Commons CCZero[10]
- GNU Free Documentation[11]

The license that applies to a specific dataset is within that dataset's metadata under the "license" field, which contains a URL to the description and terms of the license. Private entities may choose to license their data in any fashion agreed upon by contract.

[8] http://opendatacommons.org/licenses/by/1.0/

[9] https://creativecommons.org/licenses/by-sa/4.0/

[10] https://creativecommons.org/publicdomain/zero/1.0/

[11] http://www.gnu.org/licenses/fdl-1.3.en.html

Chapter 7. Archiving and Preservation Plans

7.1. ARCHIVING STRATEGY

The Operating System has a data archive component. The Operating System uses readily available and redundant services for all its data storage and archives. The cloud infrastructure of the Operating System provides multiple redundant copies of both static files and structured data through their respective services. The cloud environment services provide managed backup services, which are customizable and configurable. High-value datasets will be enabled to take advantage of this feature. In the event of data corruption, a valid copy of the data may be retrieved from a backup copy and applied to the corrupt copy to correct any corruption.

Digital preservation practices, including auto-recovery, integrity monitoring, and redundancy are incorporated to support data integrity.

All long-term preservation efforts will endeavor to comply with the ISO Technical Report (ISO/TR) 18492:2005 standard (long-term preservation of electronic-document-based information). The success of any long-term preservation efforts will be dependent on the entity that owns the project post-grant. For data that does not have a contractual archival requirement, the decision to retain a dataset will primarily be at the discretion of the data curator. As the Operating System matures, the data usage statistics and user ratings will determine the desire for the Data Curator to retain a given dataset.

7.2. TIME BETWEEN COLLECTION AND SUBMISSION TO ARCHIVE

The timeliness of availability of data modifications differ slightly between two cloud-based products that store data, but they both operate on the same principle. The Operating System cloud environment will keep multiple copies of any data within a single region. The cloud environment will rely upon back-end processes that asynchronously copy changes to replicas. Although the time to consistency is not readily available and can vary, it is typically close to real-time.

7.3. BACKUP AND DISASTER RECOVERY

The Operating System will rely on the geographically redundant and distributed nature of the cloud-based data storage repository. Whenever a change or update occurs, a copy of the original data will move asynchronously to a data archive assuming this remains feasible given storage costs. The Operating System administrators will work with the defined cloud-based data storage provider to retrieve an archive should core data become corrupt or require a restore. As the Operating System matures, the Operating System will store its data in multiple cloud computing regions for a higher level of data redundancy.

7.4. PROTECTION FROM MODIFICATION OR DELETION

Operations that require write access require user registration and an access key. Any party submitting data must register, gain submission approval, and remain in good standing. Team members with administrative privileges in the cloud-based host web interface must be an authorized user and use multifactor authentication. The Operating System will have an integrated Identity and Access Management system that utilizes single sign-on and multifactor authentication.

7.5. DATA RETENTION

As the volume of the data that the Operating System houses increases over time, the Data Curator will evaluate applying expiration policies to datasets or data within a dataset. This may include the moving of infrequently accessed data to other, less expensive storage or to make a recommendation to purge it in accordance to Ohio public records law requirements.

Appendix A. Acronyms and Definitions

Table 4: Acronym List contains program level acronyms used throughout this document.

Table 4: Acronym List

Acronym/Abbreviation	Definition
API	Application Programming Interface
ASTM	American Society for Testing and Materials
BRT	Bus Rapid Transit
CEAV	Connected Electric Autonomous Vehicle
CKAN	Comprehensive Knowledge Archive Network
CMAX	COTA's brand name for its BRT line
COTA	Central Ohio Transit Authority
CPS	Common Payment System
CSV	Comma Separated Value
CV	Connected Vehicle
CVE	Connected Vehicle Environment
DCAT	Data Catalog
DMP	Data Management Plan
DPP	Data Privacy Plan
DSS	Data Security Standards
EPM	Event Parking Management
GNSS	Global Network Satellite System
IE	Independent Evaluator
IP	Internet Protocol
ISO	International Organization for Standardization
ISO/TR	International Organization for Standardization Technical Report
ITS	Intelligent Transportation Systems
JSON	JavaScript Object Notation
MAASTO	Mid America Association of State Transportation Officials
MAPCD	Mobility Assistance for People with Cognitive Disabilities
MMTPA	Multimodal Trip Planning Application
ODOT	Ohio Department of Transportation
OGRIP	Ohio Geographically Referenced Information Program

Acronym/Abbreviation	Definition
OS	Operating System
OSU	The Ohio State University
PCI	Payment Card Industry
PDF	Portable Document Format
PfMP	Performance Measurement Plan
PHI	Personal Health Information
PII	Personally Identifiable Information
PTA	Prenatal Trip Assistance
RDF	Resource Description Framework
SAE	Society of Automotive Engineers
SLA	Service Level Agreement
SMH	Smart Mobility Hub
SoS	System of Systems
SSN	Social Security Number
TNC	Transportation Network Companies
URL	Uniform Resource Locator
XML	Extensible Markup Language
USDOT	U.S. Department of Transportation

Source: City of Columbus

Appendix B. Glossary

Table 5: Glossary contains project specific terms used throughout this document.

Table 5: Glossary

Term	Definition
Authentication	The testing or reconciliation of evidence of a user's identity. It establishes and verifies that a user is who they say they are.
Authorization	The rights and privileges granted to a person or process.
Data	Data is raw (unorganized and unprocessed) digital messages sent between components. From Society of Automotive Engineers (SAE) Standard J2735: Representations of static or dynamic entities in a formalized manner suitable for communication, interpretation, or processing by humans or by machines.
Dataset	A collection of related sets of resources to include metadata that defines dataset and its contents.
Data Element	It is a component of a dataset that makes up its resources. This could be a file, a record, row, cell, or column.
Data Ingestion	Obtaining and importing data for use or storage.
Dedupe	Process to eliminate duplicate data.
Information	Processed data that is organized, structured, or presented in a given context to make it useful.
PII	The information that can be used to distinguish or trace an individual's identity, such as their name, Social Security Number (SSN), biometric records, etc., alone, or when combined with other personal or identifying information, which is linked or linkable to a specific individual, such as date, place of birth, and mother's birth name. The definition of PII is not anchored to any single category of information or technology. Rather, it requires a case-by-case assessment of the specific risk that an individual can be identified by examining the context of use and combination of data elements. Non-PII can become PII whenever additional information is made publicly available. This applies to any medium and any source that, when combined with other available information, could be used to identify an individual.
Privacy	Defined as control over the extent, timing, and circumstances of sharing oneself (physically, behaviorally, or intellectually) with others.
Pull Interface	Defined as the interface from which the data will be pulled from the Operating System.
Push/Self-Service Interface	Defined as the interface through which the data will be pushed into the Operating System.
Requirements	Set of information necessary to accomplish one action.

Term	Definition
Resource	A resource is part of a specific dataset that can be consumed and made available in multiple formats (CSV, JSON, XML, etc.).
Transmit	Sharing data directed to a specific receiver. In the case of transmission between systems, all transmitted data is signed and encrypted where required based on SAE J2945/1.
Upsert	Method of adding/updating data. If the specific entry exists the data is updated, if it does not exist it is added.

Source: City of Columbus

THE CITY OF
COLUMBUS
ANDREW J. GINTHER, MAYOR