Smart Columbus

Data Management Plan for the Smart Columbus Demonstration Program

www.its.dot.gov/index.htm

Final Report – September 28, 2018

Source: City of Columbus – November 2015
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.

Acknowledgment 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.
---|---|---|

4. Title and Subtitle
Data Management Plan Smart Columbus Demonstration Program

5. Report Date
DRAFT September 28, 2018

6. Performing Organization Code

7. Author(s)
Jeff Hunsaker (HMB), Ty Sonagere (CoverMyMeds), Jodie Bare (City of Columbus), Scott Millard (HNTB), Sherry Kish (HNTB)


9. Performing Organization Name and Address
City of Columbus
90 West Broad Street
Columbus, OH 43215-9004

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
DTFH61-16H00013

12. Sponsoring Agency Name and Address
U.S. Department of Transportation (USDOT)
Federal Highway Administration (FHWA)
Office of Acquisition and Grants Management
1200 New Jersey Avenue, SE
Mail Drop: E62-204
Washington, DC 20590

13. Type of Report and Period Covered
Data Management Plan


15. Supplementary Notes

16. Abstract
This Smart Columbus Data Management Plan provides operational information for the use of data within the Smart Columbus Operating System platform. 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 an Application Programming Interface. This plan will be used as a guide by the team to ensure the necessary operations are performed to ensure optimum program functionality in addition to properly securing, backing up, maintaining, and sharing the data. It will also be used to support proper privacy procedure and guidelines as defined in the Data Privacy Plan.

17. Keywords
Smart City Challenge, Smart Columbus, Data Management Plan, Smart Columbus Operating System

18. Distribution Statement

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
41

22. Price

Form DOT F 1700.7 (8-72)  Reproduction of completed page authorized
# Table of Contents

Table of Contents.................................................................................................................. i

Chapter 1. Introduction ........................................................................................................ 1
  1.1 Project Description ........................................................................................................ 1
  1.2 Elements of the Operating System ............................................................................... 5
  1.3 System of Systems Overview ....................................................................................... 6
  1.4 Purpose of the Plan ....................................................................................................... 7
  1.5 Organization of the Plan ............................................................................................. 8

Chapter 2. References ............................................................................................................ 9

Chapter 3. Data Description ................................................................................................ 11
  3.1 Metadata ..................................................................................................................... 11
  3.2 Data Dictionary ........................................................................................................... 13
  3.3 Size and Scale of Data ............................................................................................... 13
  3.4 Data Acquisition/Creation ......................................................................................... 14
  3.5 Performance Measurement Data ............................................................................... 15
  3.6 External Interfaces ..................................................................................................... 15
  3.7 Frequency of Data Collection .................................................................................... 16
  3.8 Relationship of New Data to Existing Data ............................................................... 16
  3.9 Organizations ............................................................................................................ 16
  3.10 Data Consumers ....................................................................................................... 17
  3.11 Value of the Data ..................................................................................................... 17
  3.12 Party Responsible for Data Management ............................................................... 18
  3.13 Management and Audit Controls ............................................................................. 18

Chapter 4. Standards Used ................................................................................................ 21
  4.1 Data Formats .............................................................................................................. 21
  4.2 Versioning Strategy .................................................................................................... 22
  4.3 Communicating About the Data ................................................................................ 22
  4.4 Metadata Schema, Storage and Management ......................................................... 22
  4.5 Data Consumption Methods ..................................................................................... 22
  4.6 Quality Control Measures ....................................................................................... 23

Chapter 5. Justification for and Nature of Changes ............................................................ 25
  5.1 Sharing Data .............................................................................................................. 25
5.2 Private or Confidential Information ................................................................. 25
5.3 Concerns with Sharing ..................................................................................... 26
5.4 Deidentifying Data .......................................................................................... 26
5.5 Anonymizing Data .......................................................................................... 26

Chapter 6. Re-Use, Redistribution and Derivative Products Policies .................. 27
6.1 Permissions to Manage Data ........................................................................... 27
6.2 Intellectual Property Owner of Data ............................................................... 27
6.3 Copyrights to Data .......................................................................................... 28
6.4 Transfer of Rights ........................................................................................... 28
6.5 Data Licensing and Redistribution ................................................................. 28

Chapter 7. Archiving and Preservation Plans ...................................................... 29
7.1 Archiving Strategy .......................................................................................... 29
7.2 Time Between Collection and Submission to Archive ..................................... 29
7.3 Backup and Disaster Recovery ....................................................................... 29
7.4 Protection from Modification or Deletion ....................................................... 30
7.5 Data Retention ............................................................................................... 30

Appendix A. Acronyms and Definitions ............................................................. 31
Appendix B. Glossary .......................................................................................... 33

List of Tables
Table 1: References ............................................................................................. 9
Table 2: Data Collection Frequency Metadata Fields ........................................... 16
Table 3: Sample File Types .................................................................................. 21
Table 4: Acronym List .......................................................................................... 31
Table 5: Glossary of Terms .................................................................................. 33

List of Figures
Figure 1: Smart Columbus Framework .................................................................. 4
Figure 2: Elements of the Smart Columbus Operating System ......................... 5
Figure 3: System of Systems External Context Diagram ..................................... 7
Figure 4: Data Ingestion Workflow ...................................................................... 15
Chapter 1. Introduction

1.1 Project 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 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 of 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 nine individual projects.

The Columbus Smart City Demonstration Projects, depicted in Figure 1: Smart Columbus Framework, 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. (see Figure 1). Data will be available for analytics and visualization as well as for artificial intelligence required by various smart city applications. The operating system is how most of the Smart Columbus solutions can integrate with each other. The Operating System is a platform designed for Big Data, Machine Learning and Artificial Intelligence, Analytics, and complex data exchange. It will capture the data and provide a means for multi-tenant access to aggregate, fuse and consume data.

  The Data Curator will have a platform that will ingest and filter Personally Identifiable Information (PII). This will forward the filtered and transformed dataset to the Operating System where access to data services from multiple sources and tenants exists. Datasets housed in the OS 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 and increase safety. The CVE will connect 1,800 vehicles and 113 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. The data created by the system will be anonymized, de-identified, aggregated and stored by the Operating System for historical analysis and visualization.

- **Multimodal Trip Planning Application/Common Payment System (MMTPA/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, Transportation Network Companies (TNCs) and car-sharing. Using MMTPA/CPS, users can compare travel options across modes, and plan and pay for their travel based upon current traffic conditions and availability of services. The data created by the system will be anonymized, de-identified, aggregated and stored by the Operating System for historical analysis and visualization. The Route Optimization function will reside within the Operating System platform and be supported by the near real-time data-handling in the Operating System.

- **Smart Mobility Hubs**
  COTA bus stops along the BRT CMAX corridor and transit centers will be transformed into smart mobility hubs, where someone getting on or off the bus can easily access the next leg of their trip. Public Wi-Fi will be a key enabler for the hub and its points of connection (Wi-Fi is also present in COTA's stations, CMAX, and buses). The city plans to outfit the hubs with kiosks to assist in travel planning and expanded transportation options via other modes, such as bike- and car-sharing. The smart mobility hubs will be linked with COTA systems to provide transit information with real-time arrival and departure times to the passengers waiting at the hubs.

- **Mobility Assistance for People with Cognitive Disabilities**
  The city will develop and deploy an application that will allow persons with cognitive disabilities to independently traverse the city via COTA's fixed bus route system. The mobile 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 and visually disabled can travel independently. The data created by the system will be anonymized, de-identified, aggregated and stored by the Operating System for historical analysis and visualization.

- **Prenatal Trip Assistance**
  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. The data created by the system will be anonymized, de-identified, aggregated and stored by the Operating System for historical analysis and visualization.
• **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. The data created by the system will be anonymized, de-identified, aggregated and stored by the Operating System for historical analysis and visualization.

• **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. The data created by the system will be anonymized, de-identified, aggregated and stored by the Operating System for historical analysis and visualization.

• **Truck Platooning**
Freight signal prioritization on connected vehicle (CV) enabled trucks will be deployed to reduce freight-induced congestion and queuing. In addition, multiple two-vehicle CV-enabled truck platoons will be deployed from Columbus to the eastern Ohio area. Wireless communications will be added to existing vehicle technologies to allow trucks to reduce their headways when traveling on freeways. On arterials, these vehicles will receive platoon intent signal priority enabling two trucks to traverse an intersection during the same signal phase cycle. Platooning is also expected to save fuel and reduce vehicle emissions. This project is anticipated to increase the efficiency and stewardship of logistics companies by improving freight mobility and reducing emissions. The data created by the system will be anonymized, de-identified, aggregated and stored by the Operating System for historical analysis and visualization.
Figure 1: Smart Columbus Framework

Source: City of Columbus
1.2 Elements of the Operating System

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

### OPERATING SYSTEM DETAILS

#### ESSENTIALS OF THE OS

**Data Environment**
- Data Ingestion
- Streaming Data
- Data Tagging

**Data Lake**
- Fast and Slow Storage Capacity
- Only store what we must; leave data at native source (for security & storage mgmt.)

**Security**
- Data encrypted in Transit & At Rest
- ID Access Management (IAM)

**Scalable Capacity**
- Built with Open Source/Open Architecture
- Elasticity with AWS Cloud Services
- Microservices

**Data Research Environment**
- Social Community
- Data Discovery
- Analytics
- Visualization

**Shared Services**
- Application Hosting
- Function Sharing (ML, Route Optimizer)

**Business Analytics**
- Users of SCDS can analyze data to draw conclusions and solve for community problems

*Source: City of Columbus*

**Figure 2: Elements of the Smart Columbus Operating System**

The Operating System is a platform for Smart Cities development. It consists of several core functions, which can be leveraged across the Smart Columbus program, as well as other functions that will specifically enhance and support “Smart Applications” (*Figure 3: System of Systems External Context Diagram*). The core functions in the Operating System are described below:

- **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.
Chapter 1. Introduction

- **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.

- **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 and Mobility Providers.

1.3 **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 are shared in the Smart Columbus 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 Diagram** shows the relationship of the SoS to the external travelers and systems.

The Smart Infrastructure element contains the roadside units (RSUs), hubs, and corresponding network that enable interactions between these items and the Operating System. Smart Vehicles include the onboard units (OBUs) 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.
1.4 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 platform. The DMP also details how the data will be created, captured, transmitted, maintained, accessed, shared, secured and archived. The DMP provides oversight for all nine Smart Columbus projects that are undergoing lower-level, more detailed development, each at its own appropriate pace. This initial plan sets forth the system essentials while allowing for project-level details to be added quarterly as the nine Smart Columbus projects progress. The approach is iterative, bringing this plan forward in manageable steps as the projects that it guides inform it. Project-level data management development teams will 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 only be serving as a centralized metadata repository for remote datasets. In this case, some information and procedures may not be applicable to those datasets.

The DMP is an operational guide for managing the data. The DMP explains in detail how and where the 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 sensitive data or PII; in instances where the data includes PII or other restrictions, the DMP relies on the strict handling of PII detailed 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 platform continues to mature and offer more features, functionality and datasets.

1.5 Organization of the Plan

This DMP is organized into the following chapters:

- Chapter 1. Introduction
- Chapter 2. References
- Chapter 3. Data Description
- Chapter 4. Standards Used
- Chapter 5. Justification for and Nature of Changes
- Chapter 6. Re-Use, Redistribution and Derivative Products Policies
- Chapter 7. Archiving and Preservation Plans
- Appendix A. Acronyms and Definitions
- Appendix B. Glossary
## Chapter 2. References

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

<table>
<thead>
<tr>
<th>Document Number</th>
<th>Title</th>
<th>Revision</th>
<th>Publication Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Smart Columbus DPP, Internal Document, Draft</td>
<td>N/A</td>
<td>4/19/2018</td>
</tr>
<tr>
<td></td>
<td>Smart Columbus Systems Engineering Management Plan</td>
<td>N/A</td>
<td>1/16/2018</td>
</tr>
<tr>
<td>N/A</td>
<td>Open Data Policy, Project Open Data <a href="https://project-open-data.cio.gov/">https://project-open-data.cio.gov/</a></td>
<td>N/A</td>
<td>May 2018</td>
</tr>
<tr>
<td>N/A</td>
<td>M-13-13 Memorandum for the Heads of Executive Departments and Agencies: Open Data Policy – Managing Information as an Asset <a href="https://project-open-data.cio.gov/policy-memo/">https://project-open-data.cio.gov/policy-memo/</a></td>
<td>N/A</td>
<td>May 2018</td>
</tr>
<tr>
<td>N/A</td>
<td>Ohio Revised Code § 149.43: Availability of public records for inspection and copying</td>
<td>N/A</td>
<td>12/19/2016</td>
</tr>
<tr>
<td>N/A</td>
<td>Ohio Revised Code § 1306.01: Definitions</td>
<td>N/A</td>
<td>9/14/2000</td>
</tr>
</tbody>
</table>

*Source: City of Columbus*
Chapter 3. Data Description

Data within the Operating System will be made available in different formats such as Comma Separated Value (CSV), Extensible Markup Language (XML), JavaScript Object Notation (JSON) and many others. Datasets are added locally 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. The Operating System platform primarily contains two types of datasets: local and remote. Local datasets are added to the Operating System platform tagged and hosted within itself. Remote datasets are hosted on other remote open-data portals and added to the Operating System primarily for the metadata to be discovered by the Operating System users. Remote datasets are also periodically re-harvested for their metadata, a process by which the metadata is updated within the Operating System to the current available data.

When adding data to the Operating System, datasets will need a description and other metadata from both the dataset and the individual 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 improves 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 scope and characteristics of data elements, as well the rules for their usage and application; helping humans and machines better understand a given dataset.

3.1 Metadata

The Operating System has been built to conform with the Federal Project Open Data Metadata Schema v 1.1. While this metadata schema is widely adopted and based off Data Catalog (DCAT) 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

---

1 https://project-open-data.cio.gov/v1.1/schema/
2 https://www.w3.org/TR/vocab-dcat/
The Operating System project team is responsible for populating the metadata fields for local datasets and every effort will be made to ensure they are populated and correct during the data curation process. Remote datasets have their metadata read from what is exposed from the remote website or service and therefore the content of a remote dataset’s metadata is the responsibility of the remote source. Remote datasets must provide equivalent metadata fields that can be mapped to the appropriate fields to meet the minimum metadata requirements defined by the Project Open Data standards.
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 being shared for use within the Operating System, the data dictionary is preferably provided by the source data steward 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 Application Programming Interface (API) and displayed in the “Data Dictionary” field if viewing through the Operating System website.

During the initial data curation process for local datasets in which a data dictionary is unable to be provided, a data curator will work with the source entity 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 a column from different tables are address_1, one_address, first_line_address, user_address, and company_address, 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 was a person instead of a company, “Personal Address” would be a business term for user_address and the rest would remain under the Company Address classification. To ensure consistency in data dictionary fields, mapping to a standard such as NIEM\(^3\) 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 on a monthly basis and make determinations regarding whether data storage space needs to be scaled up or down within the cloud-based architecture of the Operating System platform, allowing 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.

---

\(^3\) National Information Exchange Model, https://www.niem.gov/
Chapter 3. Data Description

Data that is continuously streamed into the Operating System will be added to the data storage for that program 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 to latest event. Also, we may decide 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 of use cases we need to support.

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,

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 or SPII, what formats would be best suitable, what metadata is available, and identifying any additional resources such as data dictionaries or webpages 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 instance of the Operating System platform and once validated and approved it will be deployed to testing where it can be shared with the data provider to perform testing and additional validation. Once validated, the process 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 and SPII/PII data
- Size of data
3.5 Performance Measurement Data

Data to measure the performance of the Smart Columbus programs will be added to the Operating System as CSV files as public or private datasets within designated organizations. Independent Evaluators and other authorized users will be provided access to the data through the user interface and through the API to query the data. Unauthorized users will be denied access to the private data through the user interface or through the API.

3.6 External 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
Chapter 3. Data Description

contains one external interface for data ingestion: the API. The API provides a push interface for the ingestion of data from external resources outside the Smart Cities program.

Additional interfaces will be added when requirements for additional interfaces are discovered as other Smart Columbus projects progress.

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 platform. For local datasets, this can be defined down to near real-time. For remote datasets, this is typically a scheduled, automated reoccurring process. A metadata field, Frequency, describes how often the data is updated. The metadata fields identified in Table 2: Data Collection Frequency Metadata Fields provides detailed information about the frequency of data collection.

### Table 2: Data Collection Frequency Metadata Fields

<table>
<thead>
<tr>
<th>Field Description</th>
<th>Metadata Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Accrual Periodicity</td>
<td>How often the dataset is published</td>
</tr>
<tr>
<td>Last Updated</td>
<td>Modified</td>
<td>Date/Time that the data was last updated. ISO-8601 repeating duration.</td>
</tr>
<tr>
<td>Temporal</td>
<td>Temporal</td>
<td>Start and end date for data</td>
</tr>
<tr>
<td>Issued</td>
<td>Release Date</td>
<td>Date of issuance</td>
</tr>
</tbody>
</table>

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.

3.9 Organizations

The Operating System platform has a concept of Organizations. The Operating System will use 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 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 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 platform and API would allow users to query similar data across multiple Organizations to create a “mash-up” of data. While the concept of organizations within the Operating System...
System will stay constant, the implementation and definition of it may change as future requirements are defined and the Columbus Smart City projects mature.

Examples of organizations within the Operating System include:

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

### 3.10 Data Consumers

Users of the Operating System platform will be diverse; the Operating System will be designed to allow all types of users easier and better access to data for use and analysis. In addition to supporting the other Smart Columbus projects, users are expected to be community members, technology developers, researchers, companies, government entities and others. After users have access to the data and APIs, they can use the data to create websites, 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.

It is expected that each type of consumer 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

Many datasets within the Operating System will not require authentication and will be publicly accessible. A dataset can be marked private, which then means it can only be seen or queried by users that have the appropriate access permissions.

Data will also be shared with the Independent Evaluator (IE) through the Operating System website or by other processes set up by the Operating System project team and the IE.

### 3.11 Value of the Data

The Operating System is focused on providing high-quality datasets to its users. There are two methods used to accomplish this. First, the formal data curation process is used – during this process the data
curator works with the source entity to provide as much information as possible with the dataset so that it has value to users. Additionally, datasets will be re-evaluated for quality and value on a regular basis.

The second method of providing high quality datasets is allowing users to provide usability ratings on individual datasets which help identify both to other Operating System users and the Operating System project team which datasets are or are not providing value.

### 3.12 Party Responsible for Data Management

The Operating System platform will have many roles that are responsible for the management of the data. Below are some of the roles that will be responsible for data management. During project development and operation all roles except Data Steward will be performed by members of the Operating System project team. Data Stewards are members of the organizations providing the data to the Operating System.

- **System Administrators** – Responsible for 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 Architects** – Responsible for the design and integration of all system back-end components.
- **Data Stewards** – Responsible for working with the Operating System to provide valuable organizational data that is validated and compliant.

The organizations concept within the Operating System allows authorized users of an organization to act as the Data Steward for the data the organization provides. Authorized users can manage datasets, their 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. For remote datasets, the responsible party for auditing and management is the source system. This is due to the Operating System simply harvesting the remote source data and metadata and putting additional value-added services around them such as unified API access. For this reason, the Operating System team has no control over the data or metadata within remote datasets.

For local datasets, data curators and administrators are primarily responsible for the management of datasets. Each action within the Operating System platform is logged, either in the cloud environment or within the individual software components. Additionally, the data library keeps a public “activity stream” of updates to datasets. These factors will help ensure integrity of the data and metadata. The Operating System will create machine statistics relative to the accuracy and completeness of the dataset as it is ingested. This will provide a score that will be appended to the data page. As the social component is added to the system, each dataset will have the ability to take and track user ratings as to the usefulness and value of each dataset. The individual will be able to discuss the dataset amongst other users. When a dataset underperforms, the ratings can be used to initiate further investigation into the data. The data
steward 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 temporarily removed from public access or deleted from the system.

The Operating System will store logs in a separate read-only data store that ensures integrity and validity of the logs.
Chapter 4. Standards Used

4.1 Data Formats

The Operating System platform will only use platform-independent and nonproprietary formats to focus on machine-readability of the data. To accomplish this, it will be encouraged that any data source that is in a format that is not machine-readable attempt to 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

<table>
<thead>
<tr>
<th>Machine-Readable</th>
<th>Non-Machine-Readable</th>
</tr>
</thead>
<tbody>
<tr>
<td>JSON</td>
<td>PDF</td>
</tr>
<tr>
<td>XML</td>
<td>JPG</td>
</tr>
<tr>
<td>CSV</td>
<td>TIFF</td>
</tr>
<tr>
<td>RDF</td>
<td>MP4</td>
</tr>
<tr>
<td>XML</td>
<td>WAV</td>
</tr>
</tbody>
</table>

Source: City of Columbus

In addition to the data provided, metadata must be provided that complies with the metadata standards defined in Section 4.4 Metadata Schema, Storage and Management.

For remote datasets, the source system must publish a metadata catalog that is compliant with one of the following formats: Comprehensive Knowledge Archive Network (CKAN), DCAT, 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 platform. 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 is anticipated, coordination across projects and flexibility in staffing is recommended. Data depositors will register with the Operating System and after approval and agreement to terms of service (or memorandum of understanding), 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 Versioning Strategy

The Operating System will support versioning data on only high value datasets. In most cases, datasets can be rebuilt from the original source. When it is deemed necessary, the following versioning strategies will be used:

1. Datasets will be partitioned to create a historical log that can be used to restore data
2. An entire virtual datastore can be set to version control

4.3 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 URLs 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 will 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.

4.4 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 platform extends the metadata to include other common metadata fields that are populated when the dataset is first scheduled for ingestion.

When metadata is changed through the API or through a web browser by a user, an audit log of that action is recorded and stored in the dataset’s activity stream for others to see. This enables users to view a history of the dataset metadata.

4.5 Data Consumption Methods

Users of the Operating System platform will consume the 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 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.

There is no limit on how much data a user can request through the web site, but API requests will have request throttling set based on originating 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.

4.6 Quality Control Measures

All data providers are responsible for their own quality controls which, when applicable, are expected to conform with common industry standards, such as the ASTM’s Quality Control Standards⁴. When data is ingested into the system, the dataset is configured, and the monitoring controls are configured such that machine scores can be taken upon ingest for a baseline and then periodically in an automated fashion. Freshness, completeness and accuracy of the data will be determined initially and monitored periodically as defined in the associated SLAs. Along with SLA data, usage statistics will be taken automatically. Finally, subjective metrics can be provided by the community as defined below. These may include value and validity of the data. The number of the contributors for the respective datasets will be tracked. Statistical summaries may also be used to flag unusual entries. Should outliers, missing or otherwise anomalous entries be found, the data providers will be contacted for verification. Resolution of the flagged data may range from entry of a null value to rejection of the dataset, depending on the established Service Level Agreement (SLA) or Quality Control (QC) Plan.

Specifically, submitted data should possess or undergo the following:

• An identified, authenticated submitting entity
• All available metadata about the data set
• A provenance plan to include the SLAs
• Initial ingestion process that will identify any obvious PII, Sensitive Personal Information (SPI) or Personal Health Information (PHI)

The Operating System will create machine statistics relative to the accuracy and completeness of the dataset as it is ingested. This will provide a score that will be appended to the data page. As the social component is added to the system, each dataset will have the ability to take and track user ratings as to the usefulness and value of each dataset. The individual will be able to discuss the dataset amongst other users. When a dataset underperforms, the ratings can be used to initiate further investigation into the data. The data steward will be contacted and asked to remediate any anomalies detected. In the event of

⁴ https://www.astm.org/Standards/quality-control-standards.html
persistent underperformance or irresponsible inclusion of private data, datasets may be frozen or deleted from the Operating System.
Chapter 5. Justification for and Nature of Changes

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 for when other Smart Columbus programs are ready to begin sharing data through the Operating System. The Operating System has a concept of private datasets with built in authentication which can be used. To access private datasets, authentication and authorization will be required for the user interface and the API.

5.2 Private or Confidential Information

Data within the Operating System is classified into one of three categories: Non-PII, PII, and Sensitive Personally Identifiable Information (SPII). The exact definition and handling of sensitive information is detailed in the Smart Columbus DPP.

During the data curation process of a dataset there is a heavy focus on privacy and confidentiality – if a dataset contains PII or SPII, an evaluation will be performed to determine if the sensitive data is necessary. If the sensitive data is not necessary, the data provider will be asked to remove the data before transferring it to the Operating System. If the data steward cannot remove the sensitive data, it will be anonymized, redacted, or removed during the ingestion process so that it will contain no confidential data.

If a dataset must contain PII or SPII, it will be marked and managed as private within the Operating System.

When a dataset is marked as private, a user must authenticate and have privileges to gain access to read the data. All accounts within the Operating System will be on a per-user basis, there will be no accounts that are shared by multiple people.

The Operating System has the capability of allowing users to combine datasets to create completely new datasets. This represents a risk of re-identification. Users will be warned to avoid developing datasets that may reidentify individuals and merged datasets may be audited before being allowed to be made public.

The classes of users within the Operating System define what dataset 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.
Chapter 5. Justification for and Nature of Changes

- **Unauthenticated User** – access to all public datasets with API limits
- **Authenticated User** – access to all public datasets and private datasets to which organization administrators have provided access

### 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 auditing 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 auditing process to maintain an active oversight of such risks.

### 5.4 Deidentifying Data

During data curation of local datasets, the data will be evaluated to see whether it contains SPII/PII information. If found to contain SPII/PII information, the confidential information will be evaluated for complete removal before being sent to the Operating System. If it cannot be removed, it will be masked or redacted during the data ingestion design process based on the technical controls defined in the Smart Columbus Data Privacy Plan.

Remote datasets will not go through the de-identifying process as the data is presumably cleaned prior to being shared on the source. The Operating System project team will use all reasonable efforts to find personal data on remote datasets and if personal data is found, a dataset can be removed or marked private and the source notified of the discovery.

Periodically throughout the program and part of the peer review processes for updates, changes to the system will be reviewed to ensure that no SPII/PII data is stored in databases, logs, files or anywhere that it should not be stored.

### 5.5 Anonymizing Data

During data curation, the data and source will be evaluated to see if it contains identifiable data that can be used to determine the specific identity, location and movement of a private person thereby violating personal privacy. If route or residence information is found to clearly identify the specific whereabouts of an individual, the data will be obfuscated using a randomizing method that randomly changes the endpoint with enough offset that the true endpoint cannot be clearly determined. If, performing this process renders the data unusable for research purposes, then the dataset will be made private or will not be ingested into the system. If ingested as personal data, it will be masked during the data ingestion process. If data must be masked, procedures defined in the Smart Columbus DPP will be utilized. Datasets that are anonymized will contain metadata indicating that they have been anonymized.
Chapter 6. Re-Use, Redistribution and Derivative Products Policies

6.1 Permissions to Manage Data

The data within the Operating System is generally collected and utilized outside of the Operating System before it is ingested into the Operating System. Therefore, in most cases, the right to manage the data is on the source entity contributing the data. During data curation and with the acknowledgement of the source 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 data management professional 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 stewards 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 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

\[http://codes.ohio.gov/orc/1306.01v1\]
will be able to access the owning department and contact information in the datasets 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.

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 private data. Public data housed within the Operating System will be required to be under an “Open License” which in general requires the following conditions:

- **Reuse** – 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 platform are:

- Creative Commons Attribution
- Creative Commons Attribution Share-Alike
- Creative Commons CCZero
- GNU Free Documentation

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 contractual agreement.

---

6 [http://opendatacommons.org/licenses/by/1.0/](http://opendatacommons.org/licenses/by/1.0/)
7 [https://creativecommons.org/licenses/by-sa/4.0/](https://creativecommons.org/licenses/by-sa/4.0/)
8 [https://creativecommons.org/publicdomain/zero/1.0/](https://creativecommons.org/publicdomain/zero/1.0/)
9 [http://www.gnu.org/licenses/fdl-1.3.en.html](http://www.gnu.org/licenses/fdl-1.3.en.html)
Chapter 7. Archiving and Preservation Plans

7.1 Archiving Strategy

The Operating system is primarily a data archive. 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.

Data integrity over time will be ensured by building digital preservation practices into the system. These include functions such as auto-recovery, integrity monitoring, and redundancy.

All long-term preservation efforts will endeavor to comply with the International Organization for Standardization 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-contract. For data that does not have 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 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 cloud-based data storage provider to retrieve an archive should core data become corrupt or require a restore. If it is determined in the future as a requirement, it is
possible for the Operating System to store its data in multiple regions to ensure higher redundancy of the data.

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. Production administrators accessing the cloud-based host web interface must be an authorized user and use multi-factor authentication. The Operating System will have an integrated Identity and Access Management (IDAM) system that has single sign-on multi factor authentication capabilities.

7.5 Data Retention

As the volume of the data that the Operating System platform houses increases over time, data administrators 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 Act requirements.
Table 4: Acronym List contains program level acronyms used throughout this document.

<table>
<thead>
<tr>
<th>Acronym/Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWS</td>
<td>Amazon Web Services</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>BRT</td>
<td>Bus Rapid Transit</td>
</tr>
<tr>
<td>CEAV</td>
<td>Connected Electric Autonomous Vehicle</td>
</tr>
<tr>
<td>CKAN</td>
<td>Comprehensive Knowledge Archive Network</td>
</tr>
<tr>
<td>CMAX</td>
<td>COTA’s brand name for its first bus rapid transit line</td>
</tr>
<tr>
<td>COTA</td>
<td>Central Ohio Transit Authority</td>
</tr>
<tr>
<td>CPS</td>
<td>Common Payment System</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma Separated Value</td>
</tr>
<tr>
<td>CV</td>
<td>Connected Vehicle</td>
</tr>
<tr>
<td>CEAV</td>
<td>Connected Electric Autonomous Vehicle</td>
</tr>
<tr>
<td>CVE</td>
<td>Connected Vehicle Environment</td>
</tr>
<tr>
<td>DCAT</td>
<td>Data Catalog</td>
</tr>
<tr>
<td>DMP</td>
<td>Data Management Plan</td>
</tr>
<tr>
<td>DPP</td>
<td>Data Privacy Plan</td>
</tr>
<tr>
<td>IDAM</td>
<td>Identity and Access Management</td>
</tr>
<tr>
<td>IE</td>
<td>Independent Evaluator</td>
</tr>
<tr>
<td>ISO</td>
<td>International Organization for Standardization</td>
</tr>
<tr>
<td>JSON</td>
<td>JavaScript Object Notation</td>
</tr>
<tr>
<td>MMTPA</td>
<td>Multimodal Trip Planning Application</td>
</tr>
<tr>
<td>NIEM</td>
<td>National Information Exchange Model</td>
</tr>
</tbody>
</table>
## Acronyms and Definitions

<table>
<thead>
<tr>
<th>Acronym/Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>OBU</td>
<td>Onboard Unit</td>
</tr>
<tr>
<td>PDF</td>
<td>Portable Document Format</td>
</tr>
<tr>
<td>PHI</td>
<td>Personal Health Information</td>
</tr>
<tr>
<td>PII</td>
<td>Personally Identifiable Information</td>
</tr>
<tr>
<td>QC</td>
<td>Quality Control</td>
</tr>
<tr>
<td>RDF</td>
<td>Resource Description Framework</td>
</tr>
<tr>
<td>RDS</td>
<td>Relational Database Service</td>
</tr>
<tr>
<td>RSU</td>
<td>Roadside Units</td>
</tr>
<tr>
<td>SAE</td>
<td>Society of Automotive Engineers</td>
</tr>
<tr>
<td>SLA</td>
<td>Service Level Agreement</td>
</tr>
<tr>
<td>SPI</td>
<td>Sensitive Personal Information</td>
</tr>
<tr>
<td>SPII</td>
<td>Sensitive Personally Identifiable Information</td>
</tr>
<tr>
<td>SSN</td>
<td>Social Security Number</td>
</tr>
<tr>
<td>TNC</td>
<td>Transportation Network Company</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resource Locator</td>
</tr>
<tr>
<td>VIN</td>
<td>Vehicle Identification Number</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
<tr>
<td>USDOT</td>
<td>U.S. Department of Transportation</td>
</tr>
</tbody>
</table>

Source: City of Columbus
Appendix B. Glossary

Table 5: Glossary of Terms contains program level terms used throughout this document.

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authentication</td>
<td>The testing or reconciliation of evidence of a user’s identity. It establishes and verifies that a user is who they say they are.</td>
</tr>
<tr>
<td>Authorization</td>
<td>The rights and privileges granted to a person or process.</td>
</tr>
<tr>
<td>Data</td>
<td>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.</td>
</tr>
<tr>
<td>Dataset</td>
<td>A collection of related sets of resources to include metadata that defines dataset and its contents.</td>
</tr>
<tr>
<td>Data Element</td>
<td>It is a component of a dataset that makes up its resources. This could be a file, a record, row, cell, or column.</td>
</tr>
<tr>
<td>Data Ingestion</td>
<td>Obtaining and importing data for use or storage.</td>
</tr>
<tr>
<td>Information</td>
<td>Processed data that is organized, structured or presented in a given context to make it useful</td>
</tr>
<tr>
<td>Non-sensitive Data as PII</td>
<td>Some information may be non-sensitive or anonymous by itself, but when coupled with other available or discoverable data, can become PII and even SPII. For example, two recent decisions by the U.S. Court of Appeals for the First Circuit (In re Hulu Privacy Litig, 2014) and (Yershov v. Gannett), not only throw into question how PII may be understood, but also threaten to create a circuit split should any other circuit court tackle whether the definition of PII includes anonymous identifiers, geolocation data and elements of data that are sometimes passed from a streaming service to third parties, such as analytics providers.</td>
</tr>
<tr>
<td>PII</td>
<td>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 and place of birth, Mother's maiden name. The definition of PII is not anchored to any single category of</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>---------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td>Privacy</td>
<td>Defined as control over the extent, timing, and circumstances of sharing oneself (physically, behaviorally, or intellectually) with others.</td>
</tr>
<tr>
<td>Pull interface</td>
<td>Defined as the interface from which the data will be pulled from the Operating System</td>
</tr>
<tr>
<td>Push / Self-Service interface</td>
<td>Defined as the interface through which, the data will be pushed into the Operating System</td>
</tr>
<tr>
<td>Requirements</td>
<td>Set of information necessary to accomplish one action.</td>
</tr>
<tr>
<td>Resource</td>
<td>A resource is part of a specific dataset that can be consumed and made available in multiple formats (CSV, JSON, XML etc.).</td>
</tr>
</tbody>
</table>
| SPII                      | It is a subset of PII, which if lost, compromised or disclosed without authorization, could result in substantial harm, embarrassment, inconvenience, or unfairness to an individual. Sensitive PII requires stricter handling guidelines because of the increased risk to an individual if the data are compromised. The following PII is always (de facto) sensitive, with or without any associated personal information:  
  • SSN  
  • Passport number  
  • Driver’s license number  
  • Vehicle Identification Number (VIN)  
  • Biometrics, such as finger or iris print  
  • Financial account number such as credit card or bank account number  
  • The combination of any individual identifier and date of birth, or mother’s maiden name, or last four digits of an individual’s SSN  
In addition to de facto Sensitive PII, some non-sensitive data may be deemed sensitive based on context, as discussed next. |
| 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.                                                                                   |