WisDOT Data Inventory/Catalog

Andrew Graettinger, Ph.D.
Yang Li, Ph.D.
Xiao Qin, Ph.D., PE
Mark Gottlieb, PE

Institute for Physical Infrastructure and Transportation (IPIT)
University of Wisconsin-Milwaukee

WisDOT Project ID 0092-21-65

March 2022

RESEARCH & LIBRARY Unit

WISCONSIN DOT
PUTTING RESEARCH TO WORK

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
WisDOT Data Inventory/Catalog

5. Report Date
March 2022

6. Performing Organization Code

7. Author(s)
Andrew Graettinger, Yang Li, Xiao Qin, Mark Gottlieb

8. Performing Organization Report No.

9. Performing Organization Name and Address
Institute for Physical Infrastructure and Transportation (IPIT)
University of Wisconsin-Milwaukee
Milwaukee, WI 53201-0784

10. Work Unit No.

11. Contract or Grant No.
0092-21-65

12. Sponsoring Agency Name and Address
Wisconsin Department of Transportation
Research & Library Unit
4822 Madison Yards Way Room 911
Madison, WI 53705

13. Type of Report and Period Covered
Final Report
January 2021-March 2022

14. Sponsoring Agency Code

15. Supplementary Notes
If applicable, enter information not included elsewhere, such as translation of (or by), report supersedes, old edition number, alternate title (e.g. project name), or hypertext links to documents or related information.

16. Abstract

As WisDOT evolves to a data-driven decision-making organization, centralized and consistent information about datasets throughout the entire enterprise becomes increasingly important. The WisDOT Data Inventory\Catalog research project completed by the Institute for Physical Infrastructure and Transportation (IPIT) at the University of Wisconsin-Milwaukee (UWM) focused on finding digital datasets and pertinent information about those datasets throughout WisDOT. Much of the data at WisDOT have been collected, stored, and analyzed to meet specific goals or requirements. While this approach works, it creates siloed data, individualized schema, ad hoc metadata, and could be a security risk if the data is not stored or used appropriately. Alternatively, the benefits to an organization with robust data governance and data cataloging are: multi-dataset analytics, new insights into data, new connections between data, the ability for data discovery, and improved data quality and data security.

The research team employed a Qualtrics survey and the responses were aggregated and are presented in this report. From the analysis and variety of survey results it is clear that WisDOT needs to establish enterprise-wide data governance and data cataloging which will harmonize data sources, properly control access, document ownership, and create both technical and descriptive information about the who, what, where, when, and why of enterprise data.

17. Key Words
Data Inventory, Data Catalog, Data Governance, Data Discovery, Organizational Structure, Survey

18. Distribution Statement
No restrictions. This document is available through the National Technical Information Service.
5285 Port Royal Road
Springfield, VA 22161

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
55

22. Price

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

DISCLAIMER

This research was funded by the Wisconsin Department of Transportation (WisDOT) and the Federal Highway Administration (FHWA) under project 0092-21-65. The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views of the Wisconsin Department of Transportation (WisDOT) or the Federal Highway Administration (FHWA) at the time of publication.

This document is disseminated under the sponsorship of the Wisconsin Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof. This report does not constitute a standard, specification or regulation.

The United States Government does not endorse products or manufacturers. Trade and manufacturers' names appear in this report only because they are considered essential to the object of the document.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

As WisDOT evolves to a data-driven decision-making organization, centralized and consistent information about datasets throughout the entire enterprise becomes increasingly important. The WisDOT Data Inventory\Catalog research project completed by the Institute for Physical Infrastructure and Transportation (IPIT) at the University of Wisconsin-Milwaukee (UWM) focused on finding digital datasets and pertinent information about those datasets throughout WisDOT. Much of the data at WisDOT have been collected, stored, and analyzed to meet specific goals or requirements. While this approach works, it creates siloed data, individualized schema, ad hoc metadata, and could be a security risk if the data is not stored or used appropriately. Alternatively, the benefits to an organization with robust data governance and data cataloging are: multi-dataset analytics, new insights into data, new connections between data, the ability for data discovery, and improved data quality and data security.

The research team employed a Qualtrics survey comprised of 12 questions that was distributed to 580 people to identify datasets and information about those datasets. Through this process, 230 datasets associated with 106 people were identified. The responses from the survey were aggregated and are presented in this report. From the analysis and variety of survey results it is clear that WisDOT needs to establish enterprise-wide data governance and data cataloging which will harmonize data sources, properly control access, document ownership, and create both technical and descriptive information about the who, what, where, when, and why of enterprise data.

While data governance and data cataloging are large and on-going undertakings, the value added to a data-driven organization offsets the cost. An enterprise-wide initiative like this will require resources including at least two full time positions: a Data Governance Director, and a Data Catalog Administrator, as well as a data catalog software. The new positions, in collaboration with existing data stewards, will be responsible for creating, managing, and implementing data governance rules as well as collecting, storing, analyzing, and leveraging the metadata information within the data catalog to discover new relationships that advance the mission of WisDOT.

1 BACKGROUND

Problem Statement

Wisconsin Department of Transportation's (WisDOT's) Enterprise Data Services (EDS) Section is responsible for coordinating data management and governance activities across the agency. The bulk of WisDOT data is maintained in EDS-administered databases. However, significant data resources are also managed by vendors or WisDOT business areas in other siloed database and file systems in-house, at vendor sites, or in the cloud. While there are advantages to WisDOT's decentralized approach to data management and access, without a robust data catalog and implemented data governance this decentralized approach can hinder data discovery, understanding, and integration, as well as consistent data driven decision-making.

This project is the first step in efforts to identify all existing data resources (regardless of who manages them or where they reside) to facilitate development of a WisDOT data catalog.

Objectives

The objectives of this project are to explore and provide an implementation-ready WisDOT Data Inventory that documents the structure, quality, definitions, and usage of WisDOT data. The ultimate goal of the results of this project is to help move the agency toward a single source for current metadata, reduce data silos, improve analytics, and improve discovery.

Organization of Report

The remainder of this report is organized as follows: Section 2 presents a literature review on the data governance, data inventory/catalog, and best practices in data management proposed or done by state DOTs and agencies. Section 3 presents the methods of the data discovery survey design and distribution strategy. Section 4 presents detailed survey data results and analyses. Section 5 presents the review of data catalog vendors and applications with associated evaluation criteria. Section 6 concludes this report with a summary and a discussion of the directions for future work.

2 LITERATURE REVIEW

This section provides general introductions to two major concepts within the data management ecosystem, which are: 1) data governance and 2) data inventory/catalog. Additionally, this literature review focuses on current best practices for developing an organizational data inventory and catalog. Typical information included in an organizational data inventory and catalog include key parameters, such as: lineage, classification, description, need, application, users, etc. This section is organized as follows: 1) concepts of data governance; 2) concepts of data inventory/catalog; and 3) best practices in data management identified in the literature.

2.1 Data Governance

Data governance is a systematic process and set of rules for managing data at an enterprise level. Like any governance, this requires buy-in and support at all levels, from the leadership of an organization to the people "in the field" who collect and use the data. Therefore, robust data governance requires a data governance director as well as a data governance steering committee to oversee the creation and implementation of rules that maximize data value and minimize interruption or additional work for data owners.

While it is easy to understand and focus on a transportation asset, like a bridge or traffic light, it is more abstract to think about the information (the database describing the bridge for example) as a transportation asset in itself that requires the same rigor and thoughtfulness to describe as the asset class themselves. This can be seen by reviewing the NCHRP Research Report 956: Guidebook for Data and Information Systems for Transportation Asset Management (Spy Pond Partners, LLC & Atkins North America, Inc., 2021) which focuses on the collection, description, and storage of data that describes an asset. Or one could look at GIS in Transportation Data Governance & Data Management (Green & Anthony, 2018) which independently is looking at similar data management issues, but related to spatial location data. Both of these studies shed light on the critical data issues related to a specific topical area, but because data management is being studied and developed within isolated business units, metadata produced will most likely be disconnected from others in a DOT. A holistic agency-wide approach that meets the needs of individual groups and contributes to the greater mission of a DOT is needed. With enterprise data governance, individual groups do not need to learn and develop their own data practices, they simply need to adopt the enterprise data governance standards which saves time and resources and ensures metadata consistency.

Through a peer exchange among DOT Chief Data Officers from across the country, some lessons learned were advanced from their experiences with data governance implementation (Vandervalk et al., 2020).

Small and Simple:

- Keep Data Governance groups small
- Let Data Governance grow organically
- Keep Data Governance simple
- Perfectionism is the enemy of progress

- Let innovation happen

Roles and Responsibilities:

- It takes a champion to accomplish Data Governance
- Rely on workforce/hiring staff/transfer knowledge from experts retiring to new generation
- CDO responsibilities cannot be accomplished by part-time assignment
- Define a clear role for Information Technology (IT) and business owners

Organizational buy-in and cultural changes:

- Create executive engagement in every step
- Conduct internal advertisement and marketing to change culture, workforce, and procedures
- Focus on what is good for the business and avoid getting sidetracked
- Standards and Procedures
- Define data lifecycle
- Ensure alignment of organizational process and technology

Ohio DOT (ODOT) is several years into implementing a Data Governance strategy, which is outlined in (Albee et al., 2020). As part of that strategy, ODOT has developed a three-tiered Governance Framework shown in Figure 1. Data Governance Drivers is the top tier, with governance activities supporting the drivers and data life cycle management making up the day-to-day data management workflows. As seen along the left side of Figure 1, data governance buy-in is at all levels of the agency.

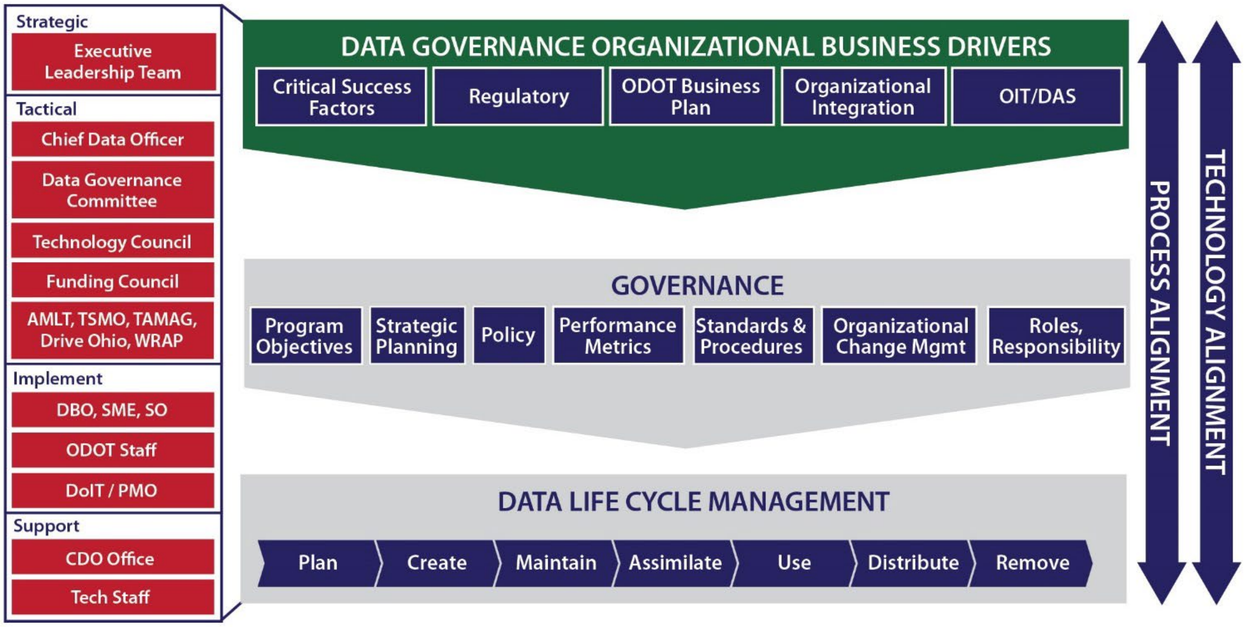


Figure 1. Data Governance Structure of ODOT

It should be noted that an independent Data Governance Office was created to oversee ODOT's Data Governance Committee and Data Governance Framework (Albee et al., 2020).

With an increase in data assets, more organizations are realizing there is value in understanding, standardizing, and connecting enterprise data. However, within the distributed and independent nature of many DOTs, there are significant barriers that inhibit the ability to leverage data resources. Data governance is becoming the de facto approach DOTs are implementing to maximize data value while minimizing disruption to mission critical activities.

2.2 Data Inventory and Data Catalog

It is important to understand the depth and breadth of enterprise data and the status of metadata describing these digital assets. A data inventory can help inform and guide an organization as it strives to improve data management and establish data governance.

Generally speaking, a data inventory could be considered as a complete record of the data information kept and maintained by an organization. It can also be treated as a process of adding tags derived from the data classification to index the contents of the stored data and make individual components more easily searchable. In other words, the data inventory is structured to provide a data dictionary and indexing for the user metadata annotations (Stillerman, 2016). Fundamentally, a data inventory is similar to building the backend of a search engine for data assets.

At a minimum, a data inventory – sometimes also called a data map – is a crucial step to knowing what data an organization has and improving the management of these data assets. The inventory will increase accountability within the organization, from end users who engage with the ultimate services, to engineers who put together analyses, to the data stewards who track data quality, to the chief information officers who make decisions, and finally to data security personnel who keep the data safe and secure. Additionally, a good data inventory can also lead to better overall reporting, decision-making and operational performance. One example for illustrating such benefits can be found in a project by The Texas Department of Transportation (TxDOT), the Project 5-9053-01, "Enhancing Road Weather Management during Wildfires and Flash Floods through New Data Collection, Sharing, and Public Dissemination Technologies", in which they built up an online, interactive data inventory and catalog which aims at curating multidisciplinary data sources to achieve better road weather management (Bhat et al., 2019).

As the amount of data increases in an organization, the need to understand what the datasets are and their data management increases, in terms of data accessibility, quality, and timeliness of updates. Without a proper inventory and data catalog in place for data management, it is virtually impossible for an organization as large as a state DOT to handle the increasingly large amounts of data due to data accumulation.

Figure 2 shows an example of modern data architecture (Wells, 2021), and it could be seen that the data catalog functions as a crucial component to connect all data with data management for an organization. Generally, the population of a data catalog will be conducted after: 1) establishing a data governance structure, and 2) determining a thorough data inventory scope and plan. Simply put, a data catalog is a detailed and organized inventory of all data in an organization. A data catalog stores metadata, which combined with data management and search tools to help data users find the data they need and provide information to evaluate the fitness of data for achieving more values for an organization from their data assets. More specifically, the data catalog should include metadata about: 1) datasets; 2) processing; 3) searching; and 4) people – owners and users.

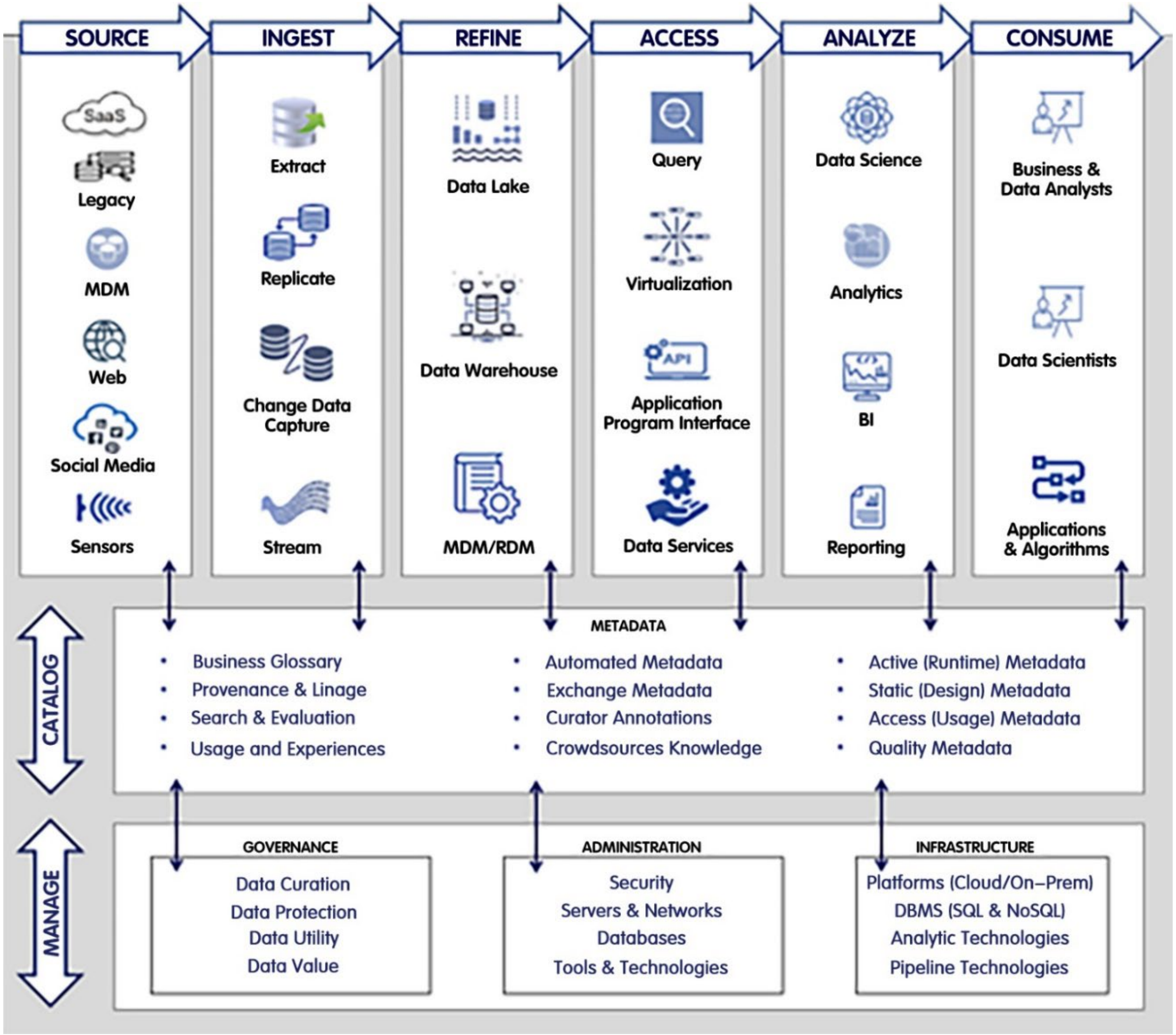


Figure 2. Modern Data Architecture (Wells, 2021)

The core for a good data catalog is the metadata. Commonly, a catalog is designed to gather and store information, or data, about the data inventory and also about processes, people, and platforms related to the data. Generally, the metadata tools in the past collected business, process, and technical metadata, and data catalogs continue such practice, but do much more.

They collect metadata about datasets, metadata about processing, metadata for searching, and metadata for and about people. Figure 3 shows a logical data model that represents typical metadata content of a data catalog (Wells, 2019).

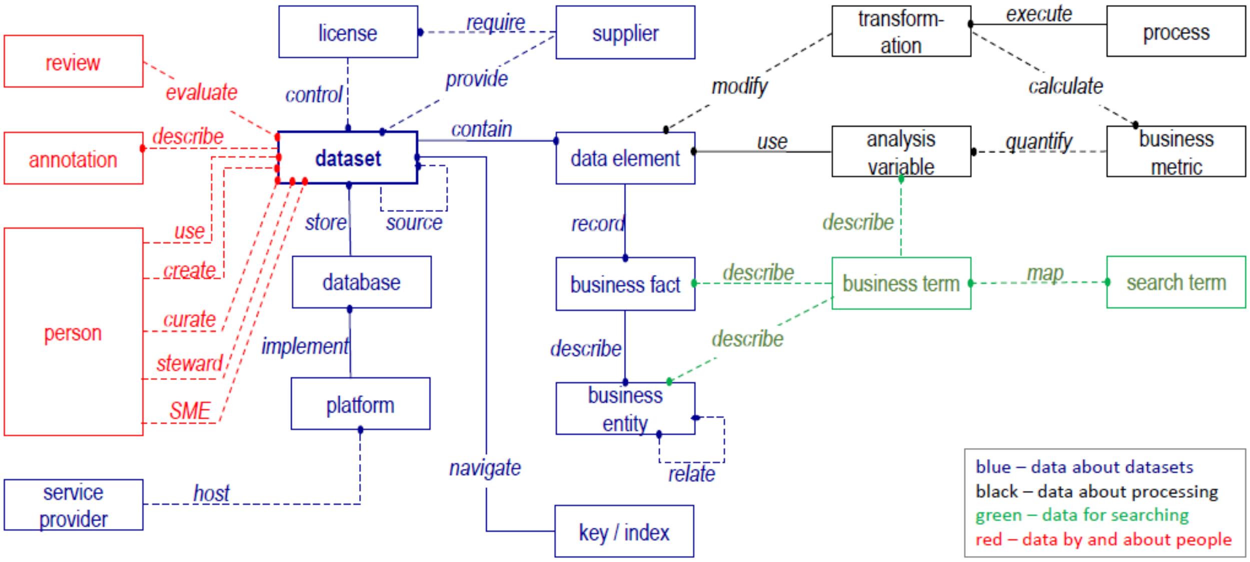


Figure 3. Example of Metadata in A Data Catalog (Wells, 2019)

As seen in Figure 3, blue boxes and lines are about data, red boxes and lines are about people, black boxes and lines are about processes and green boxes and lines are about searching. All of this information is decided upon through data governance and is stored in a data catalog.

From the literature, a general set of methods to move an organization towards robust data management is described. These methods could be adopted and applied by organizations such as state DOTs. The following is a general step-by-step process of building a data catalog (Varshney, S., n.d.):

1. Accessing Metadata for Datasets:
   - As mentioned previously, metadata is used by data catalogs to identify the data tables, files, and databases.
2. Building a Data Dictionary:
   - A data dictionary contains the description and Wiki of every table or file and all metadata entities.
3. Profiling to See the Data Statistics:
   - Profiles of the data are informative summaries which explain the data to help data users have a quick understanding of the data, discover data quality issues, risks, and overall trends.
4. Marking Relationship Amongst Data:
   - This is a critical process to help data users discover related data across multiple databases, which could be done in the following ways:

    1. *Through human knowledge*
    2. *Through advanced algorithms*
    3. *Through query logs*
- Building Lineage:
    - This step aims at helping track data from its origin to its destination and the lineages built explain the different processes in the data flow.
- Organizing Data:
    - This step helps arrange the data in a technical format, which will make the most sense to data users. The following shows some techniques for this step:
        1. *Tagging*
        2. *Organizing by an amount of usage*
        3. *Organizing by specific users' usage*
        4. *Through automation*

Additionally, the catalog should be easily accessible and the security of the information in the data catalog should be ensured.

2.3 DOTs Data Management Best Practices

This section focuses on a literature review on the best practices of data management either proposed or launched by state DOTs and agencies. Four state DOTs, one local government agency, and one NCHRP report have been identified as leaders within State DOTs in this field with published results. The following subsections provide detailed information for each case.

2.3.1 FDOT ROADS Project

In the fall of 2014, the Florida Department of Transportation (FDOT) began to develop an enterprise-wide information technology strategy. Five problems were recognized that needed to be emphasized: 1) It was difficult to know which data were available; 2) Data were difficult to access; 3) Lack of standardized approach to data management; 4) No enterprise-level view of data; and 5) Teams wanted a “one-stop shop”.

In response to the abovementioned issues, FDOT then developed an initiative project called Reliable, Organized, Accurate Data Sharing (ROADS). The goal of the ROADS initiative is to improve data reliability, simplify data sharing across FDOT, and have readily available and accurate data to make informed decisions.

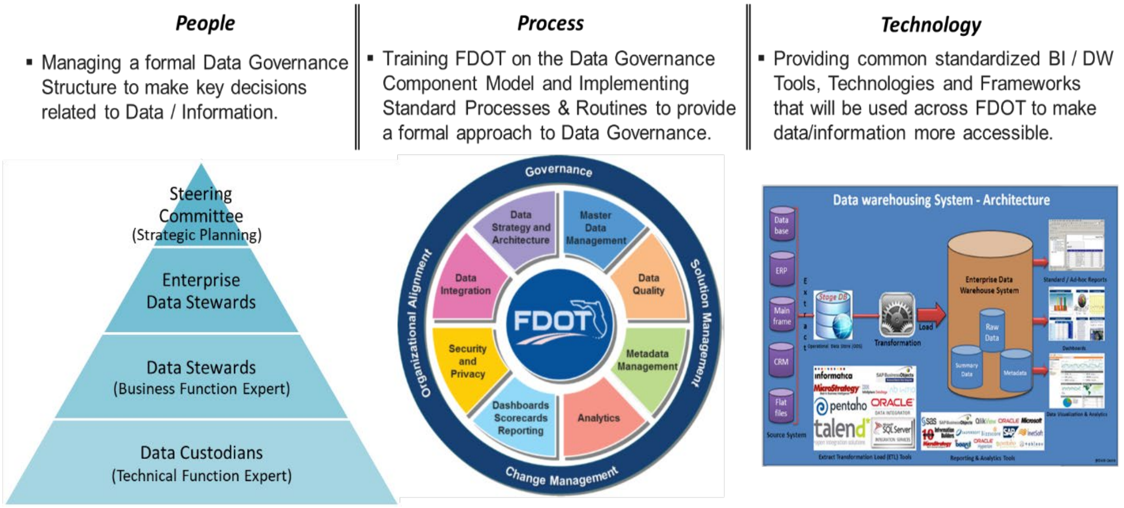
In order to assess FDOT’s needs, surveys among employees and interviews throughout the central office and 7 districts were conducted. From these surveys, 63 distinct information gaps were identified from the results.

With the data management issues identified, FDOT then focused on the key elements to improve data management (Figure 4):

- People, who is going to take charge of the data management,

- Process, which represents the Standard Processes & Routines that will be applied during the data management implementation,

- Technology, which are any BI / DW Tools, Technologies and Frameworks provided to support the data management.



People

- Managing a formal Data Governance Structure to make key decisions related to Data / Information.

Process

- Training FDOT on the Data Governance Component Model and Implementing Standard Processes & Routines to provide a formal approach to Data Governance.

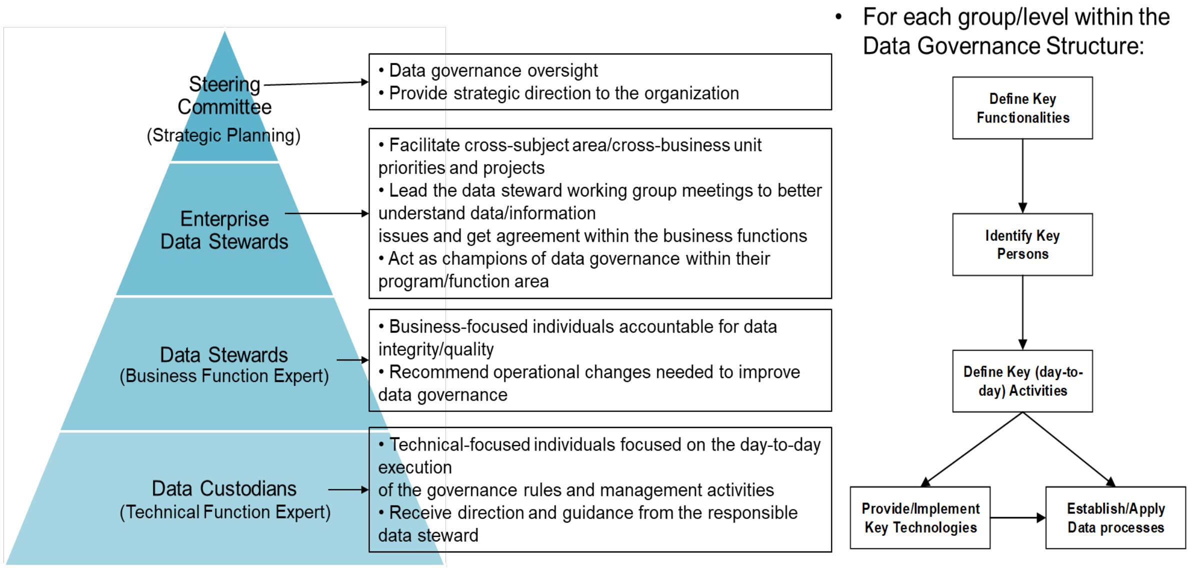
Technology

- Providing common standardized BI / DW Tools, Technologies and Frameworks that will be used across FDOT to make data/information more accessible.

Steering Committee (Strategic Planning)
Enterprise Data Stewards
Data Stewards (Business Function Expert)
Data Custodians (Technical Function Expert)
Governance
Solution Management
Change Management
Organizational Alignment
FDOT
Data Strategy and Architecture
Master Data Management
Data Quality
Metadata Management
Analytics
Dashboards Scorecards Reporting
Security and Privacy
Data Integration
Data warehousing System - Architecture
Data base
ERP
Main frame
CRM
Flat files
Source System
Extract
Stage DB
Operational Data Store (ODS)
Transformation
Load
Enterprise Data Warehouse System
Raw Data
Summary Data
Metadata
Standard / Ad-hoc Reports
Dashboards
Data Visualization & Analytics
Extract Transformation Load (ETL) Tools
Reporting & Analytics Tools

Figure 4. FDOT ROADS Key Elements

On the left of Figure 4, the "People" also represent the data governance structure to be implemented in the project. As displayed in Figure 5, for each group/level within the Data Governance Structure the major functionality of the group has been defined, and then the key activities for this group of personnel. Next, corresponding to the activities, standardized technologies, solutions will be provided to assist the completion of their data management processes.



Steering Committee (Strategic Planning)

- Data governance oversight
- Provide strategic direction to the organization

Enterprise Data Stewards

- Facilitate cross-subject area/cross-business unit priorities and projects
- Lead the data steward working group meetings to better understand data/information issues and get agreement within the business functions
- Act as champions of data governance within their program/function area

Data Stewards (Business Function Expert)

- Business-focused individuals accountable for data integrity/quality
- Recommend operational changes needed to improve data governance

Data Custodians (Technical Function Expert)

- Technical-focused individuals focused on the day-to-day execution of the governance rules and management activities
- Receive direction and guidance from the responsible data steward

- For each group/level within the Data Governance Structure:

Define Key Functionalities
Identify Key Persons
Define Key (day-to-day) Activities
Provide/Implement Key Technologies
Establish/Apply Data processes

Figure 5. FDOT Data Governance Structure

In addition, as shown in Figure 6, this project also provides a formal structure for Data Governance, Solution Management, Change Management, and Organizational Alignment. The structure guides decisions related to information, standardized processes and routines to formalize Data Governance implementation, a set of resources for training FDOT staff on Data Governance, and common, standardized approaches to acquiring, managing, and disposing of business intelligence and data warehousing tools that will be used across FDOT to make information more accessible. As an initial step in implementing Data Governance as part of this framework, FDOT has developed and released a Transportation Data Portal for visualizing, questioning, analyzing, and interpreting available data. The Transportation Data Portal is a platform for locating data related to FDOT's core mission and will be enhanced and maintained in a manner that is consistent with the Data Governance structure.

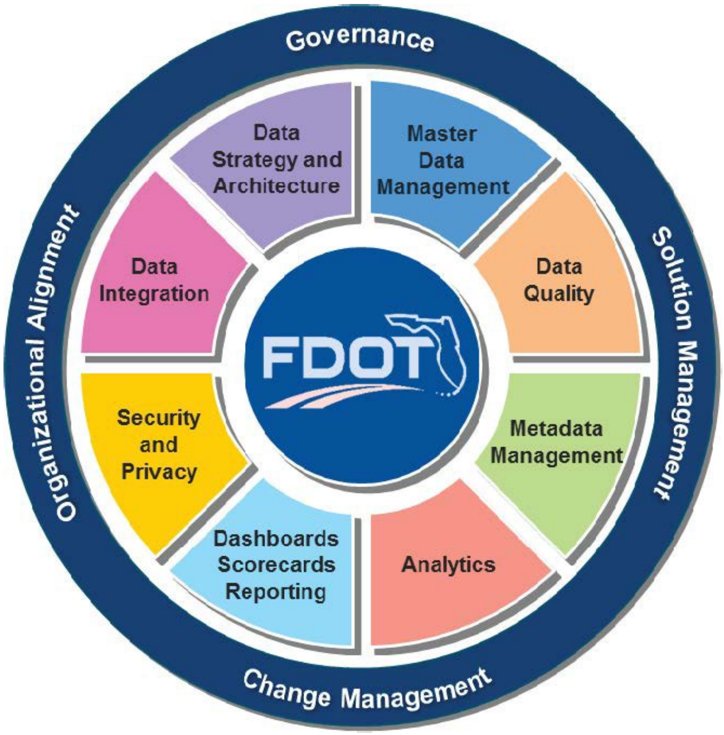


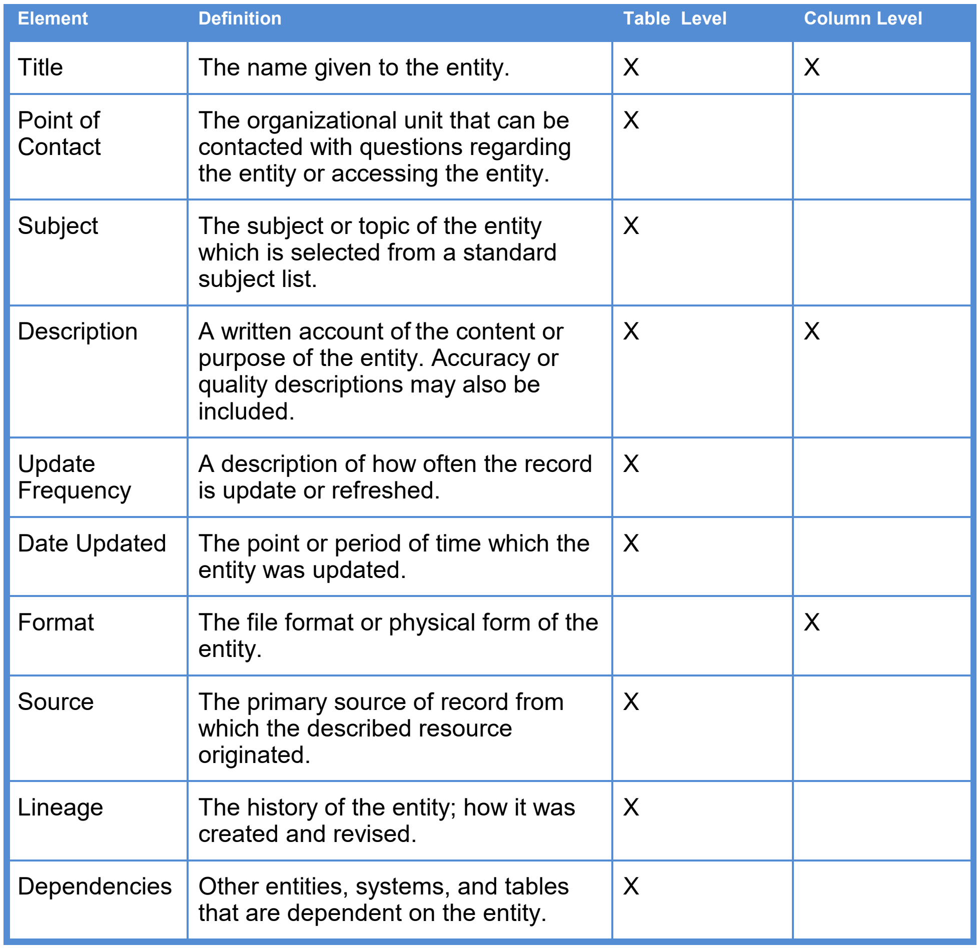
Figure 6. FDOT's Data Governance Component Model

2.3.2 MnDOT Data Governance

In 2008, Minnesota Department of Transportation (MnDOT) initiated a process to develop a business plan for data. The purpose of the plan was to strengthen the alignment between data program investments and the business needs of the department.

The data business planning process provided a framework to respond to growing transportation data and information gaps and requirements. In addition, it provided a platform for considering how stronger data management practices can: 1) increase transparency and accountability; 2) expand the reliability and utility of data to meet business decision making needs; 3) create efficiencies in accessing, sharing and using data and information; 4) standardize processes and systems that reduce redundancy and promote consistency of data; and 5) optimize new information management and spatial data tools and methods.

In regard to the data catalog/inventory, first MnDOT conducted a survey on three key business emphasis areas with a focus on 5 evaluation criteria (i.e., accessible, accurate, complete, credible, and timely), while setting up seven principles for data management: 1) data shall be managed as a state asset; 2) data quality fits its purpose; 3) data is accessible and shared as permitted; 4) data includes standard metadata; 5) data definitions are consistently used; 6) data management is everyone's responsibility; and 7) data shall not be duplicated. Then they established a standard for metadata (Figure 7) and collected them to build up their data catalog. It should be noted that the definitions provided in Figure 7 were developed by MnDOT and are not necessarily the terminology or definitions used at WisDOT.



Element	Definition	Table Level	Column Level
Title	The name given to the entity.	X	X
Point of Contact	The organizational unit that can be contacted with questions regarding the entity or accessing the entity.	X	
Subject	The subject or topic of the entity which is selected from a standard subject list.	X	
Description	A written account of the content or purpose of the entity. Accuracy or quality descriptions may also be included.	X	X
Update Frequency	A description of how often the record is update or refreshed.	X	
Date Updated	The point or period of time which the entity was updated.	X	
Format	The file format or physical form of the entity.		X
Source	The primary source of record from which the described resource originated.	X	
Lineage	The history of the entity; how it was created and revised.	X	
Dependencies	Other entities, systems, and tables that are dependent on the entity.	X	

Figure 7. MnDOT Metadata Element Standards

After successfully implementing the data business plan, MnDOT then defined the data domains, which consist of 9 with one single responsible data steward for each domain as shown in Table 1.

And then, within each domain, 5-20 subject areas were further identified, also with one single responsible data steward for each subject area, which can be seen in Table 2.

Table 1. Data Domains Used in Minnesota DOT Data Governance Model

Data Domain	Domain Description	No. of Subject Areas
Business stakeholder/ customer	Data on the interface with external stakeholders with whom MnDOT has business or customer relationships and data about internal and external communications	10
Financial	Data related to receiving, managing, and spending funds	14
Human resources	Data about individual employees	10
Infrastructure	Data on the basic facilities that make up or interface with the transportation system	13
Planning, programming, and projects	Data that provide direction for and management of projects	11
Recorded events	Data on time-based occurrences that take place on the transportation system or that affect the transportation system	19
Regulatory	Data on topics that are controlled or directed by legal requirements	20
Spatial	Data that define locations on earth or in space, including GIS, CAD, latitude/longitude, xyz coordinates, sections of roadway, or boundaries	5
Supporting assets	Data on all items that affect or support the transportation system (e.g., building and facility, fleet, communications towers)	12

Table 2. Data Subject Areas within Minnesota DOT Infrastructure Data Domain

Subject Area	Description
Airport data	Data on the publicly owned system of Minnesota airports.
Bicycle data	Data on bicycle facilities within Minnesota's transportation system, including existing/future data on state bikeways and U.S. bicycle routes, shared-use paths, protected bike lanes, bike lanes, shared lane markings and bicycle boulevards.
Bridge data	Data on the design, construction, and maintenance of bridges, including bridge condition and load ratings. Data can be contained within Pontis and structure information management system (SIMS).
Drainage structure data	Data on hydraulic features such as culverts, channels, storm tunnels, retention ponds, and drains.
Interchange, intersection, and section data	Data that describe the location of roadway intersections, the location of specific portions (sections) of roadway, and the location of places where two roadways cross (intersect) designed to permit traffic to move freely from one road to another without crossing another line of traffic.
Parking facility data	Data on the ABC distributor ramps and other facilities in Minneapolis.
Rail crossing data	Data on the highway rail grade crossings and characteristics where roadways and railroad tracks intersect.
Right of way and contaminated property data	Data on the acquisition (purchase, lease) and management of real estate/property in transportation corridors or as part of the state rail bank, which is owned by or up for purchase by MnDOT.
Roadway data	Data on location, jurisdiction, classification, surface type and width, reference points, cross sections, control sections, oversize/overweight/twin trailer routes, and project history for the statewide highway system.
Safety feature data	Data on the guardrails, median barriers, railings, crash cushions, roadway lighting, rest areas, and similar hardware or facilities that are used to improve safety on the road system.
Sidewalk data	Data on pedestrian accommodations within MnDOT's transportation system, including Americans with Disabilities Act (ADA) compliance data on sidewalks, curb walks, and pedestrian bridges.
Smooth road data	Data on the ride rating (smooth ride) of the roadways.
Traffic control device data	Data on all signs, signals, markings, and other devices used to regulate, warn, or guide traffic, placed on, over, or adjacent to state trunk highways. Data on all of the devices covered by the *Manual on Uniform Traffic Control Devices*.

2.3.3 MDOT Data Governance Plan

The Data Office in the Michigan Department of Transportation (MDOT), is in the Transportation Secretary's Office (TSO), and has begun a process to implement Data Governance in 2008. This allows MDOT to: 1) better understand and document the data assets and information systems used by their employees in support of daily activities and MDOT's broader mission; 2) define data management roles, responsibilities, and procedures; and 3) consistently manage data availability, usability, integrity, and security.

Similar to FDOT's ROADS, MDOT tried to identify issues/gaps among people, processes, and technologies. In relation to data, they believed that the catalogs should be reviewed and revised by data business owners within each stakeholder office to ensure that all data systems, data standards, roles, and responsibilities, etc., are correctly identified. They should also be reviewed at least on an annual basis, or monthly if changes occur that require updating the information listed in the catalog. At the same time, the information systems catalog, which serves as the data catalog, was also developed. Moreover, for both catalogs, the plan was to establish the attributes required to filter, sort, and understand data assets in place, and determine gaps in information and related staff capabilities (to collect, manage, analyze, and report data) to support business process improvements. Table 3 shows some sample attributes for both the data catalog and the information systems catalog.

Additionally, a data dictionary was developed, which is a descriptive list of data elements collected and maintained to ensure consistency of terminology. The dictionary was developed by data stewards and their IT counterparts in consultation with each other. The data dictionary was standardized across the transportation business units and was built manually or using software applications. The following elements were included in a data dictionary: 1) a listing of data objects including names and definitions; 2) properties of data elements (data type, size, indexes, etc.); 3) reference data (classification and descriptive domains); 4) missing data and quality-indicator codes; and 5) business rules, such as validation rules for schema or data quality. Furthermore, guidelines adopted from American Health Information Management Association (AHIMA) were used for developing the data dictionary.

Table 3. Sample Attributes for Both Data Catalog and Information Systems Catalog

Data Catalog: Sample Attributes	Information Systems Catalog: Sample Attributes
▪ The name of the data asset ▪ A brief description of the function of the data asset ▪ Core business processes at MDOT that are supported by the data asset ▪ An indication of which information systems rely on the data comprising that data asset ▪ List of data business owners, with their contact information. Data business owners may be associated with an office who manages the data and metadata for information systems within their area of responsibility for a business unit, maintain the data dictionaries for the data assets within their office, and establish business requirements for the use of the data ▪ List of data stewards responsible for the data, with their contact information. Data stewards ensure data is managed according to MDOT policies ▪ Instructions for accessing data standards and definitions associated with the collection and use of the data	▪ The name of the information system ▪ A brief description of the information system ▪ Core business processes at MDOT that are supported by the information system ▪ An indication of whether the information system is a system of record, and, if so, for what ▪ List of associated data assets that feed into or are generated by the information system ▪ System developer ▪ Operating System ▪ Software Language ▪ Version ▪ Instructions for accessing standards and definitions associated with the use of the information system ▪ List of IT staff who are responsible for managing and maintaining the information system, with their contact information. The information system owners may be associated with an office who manages the information systems within their area of responsibility for a business unit, and establish business requirements that are supported by the information system ▪ Permissions and level of access granted to classes of system users

2.3.4 SCDOT Asset Data Collection Assessment Project

In 2016, the South Carolina Department of Transportation (SCDOT) launched an asset data collection assessment project to ensure that the future SCDOT database specifications and data collection efforts support the MAP-21 requirements for data-driven performance-based management of transportation facilities, as well as meet the needs of SCDOT in a cost-effective manner. In order to achieve the goal of the project, a questionnaire was produced and distributed to understand: the types of data, sources of data, data format, data storage/access/sharing, and applications. Additional information requests asked for: data dictionaries, data collection manuals and procedures, data management documentation, data verification procedures, meta-data, and most up-to-date cost information for maintaining data. This data survey was sent to the data stewards. Additionally, SCDOT established several measures of effectiveness (MOEs) for federal data reporting requirements, state or local data reporting requirements, data collection resource requirements, data collection frequency, availability of resources, importance for traffic operational improvements, importance for safety improvements, importance for maintenance, importance for risk management. SCDOT also reviewed all available data to identify the data gaps and issues associated with their data catalog/inventory project.

2.3.5 NCHRP Research Report 952: Guidebook for Managing Data from Emerging Technologies for Transportation

Published in 2020, the NCHRP Research Report 952: Guidebook for Managing Data from Emerging Technologies for Transportation, will assist state DOTs to better manage their data assets, especially with the dramatic growth in data volume. This research report is in the form of a guidebook and provides a framework for managing big data from emerging technologies, including data from connected and automated vehicles and data linked to new mobility initiatives (e.g., smart city programs). The guidebook also outlines a process for applying that framework to incorporate these data into the decision-making process.

One of the supporting tools provided in this report is the “Data Sources Catalog Tool”. The report recommends that the following information about data should be periodically assessed by the state DOTs: 1) what data sources are in use; and 2) what data sources are available to be used. Such assessments could prevent the DOTs from overlooking data sources that could be vital to current or future projects and provide a better understanding of how data sets are connected to support the creation of a metadata catalog, planning for storage, development of new data pipelines, and better organization of the data lake structure. Maintaining a detailed catalog of data sources is one of the first and best ways to understand the nature of an agency’s data and guide the development of the data analytics processes that can be built.

Tables 4 through 6 provide examples of the way to construct a data sources catalog to summarize the specificities of each data source. To best review and assess the needs of available data sources, each data source is represented by its own row, with columns briefly describing the various facets of the data source.

Table 4. Information Gathering Form

Data Name	**Live Traffic Feed**	
Data Location	Z:/DataLake/LiveFeeds/Traffic_XML/	
Data Description	XML data pulled from roadside sensors every 10 seconds	
Data Sensitivity	No sensitive information or PII	
	Data Governance Roles	
Name of Role	**Description of Role**	**Personnel Filling Role**
Data Owner	Exercises administrative control over the data. Concerned with risk management and determining appropriate access to data. This role is typically filled by the most senior executive within the division that controls, creates, or most often uses the data.	
Data Steward	Ensures the quality and fitness of the data. Concerned with the meaning and correct use of data. This role is typically filled by a division SME with domain knowledge relevant to the data or by a member of the data team.	
Data Custodian	Exercises technical control over the data. Concerned with implementing safeguards, managing access, and logging information. This role is typically filled by IT personnel, such as system or database administrators.	
Data Curator	Manages the inventory of data sets. This includes cataloging the data, maintaining descriptions for the data, and recording the data usage. This role is typically filled by senior IT personnel or by a member of the data team.	
Data Coach	Collaborates with business data users to improve skills and promote data usage. This role is typically filled by a member of the data team or by a data SME within a division.	

Table 5. Information Cataloging Form

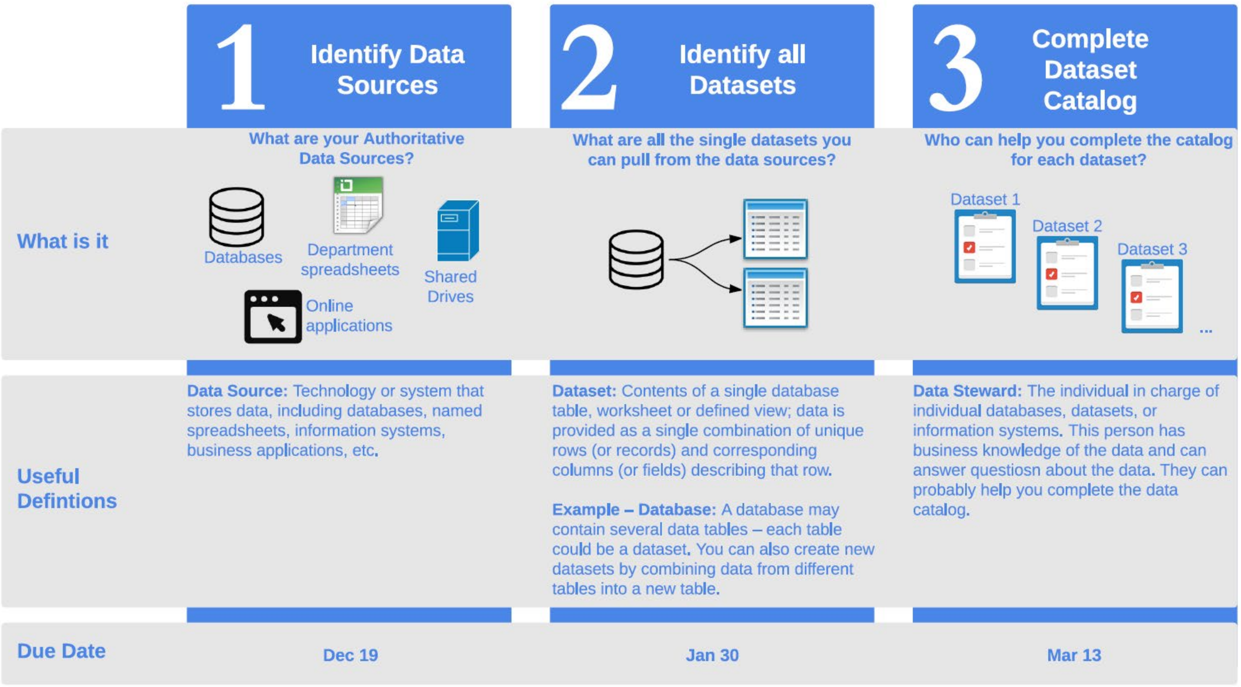
Data Name	Location	Description	Sensitivity	Owner	Steward	Custodian	Curator	Coach
Live Traffic Feed	Z:/DataLake/LiveFeeds/Traffic_XML/	XML data pulled from roadside sensors every 10 seconds	No sensitive information or PII					
Traffic Incident Performance Measures	Z:/DataLake/Historical/TIMPMs/	Traffic Incident data reported by responders	PII: License Plates					

Table 6. Data Source Assessment Example

Data Source	Description	Ownership	Format	Size	Cost	Security Level	Granularity	Restrictions	Update Frequency	Projects	Last Reviewed
Waze Incidents	Traffic speeds based on global positioning systems probe data	Internal	XML	2.1 TB total	$70,000 /year	Proprietary	Predefined roadway segments	Cannot share without permission	1 minute	Work Zones, Signal Timing	03/12/2019
Snowplow AVL	Probe data from snowplows	Internal	REST API	4 TB total	$4 /truck	No PII	0.01-mile point	None	1 minute	DOTPJ, Work Zones	01/15/2019
CoCoRahs	Certified crowdsourced weather reports	CoCoRahs Network	XML	380 MB total	Free	No PII	Interpolated from number of reports	None	24 hours	SNIC, Possibly DOTPJ	04/03/2019
Incident Reports	Individual incident reports collected from participating local agencies	Internal	CSV	500 MB total	$15 /month	Sensitive	1 row = 1 incident	None	Monthly batch upload	A-110, possible use in A-123	02/22/2019

2.3.6 City and County of San Francisco's Open Data Program – DataSF

Initiated in 2009, the City and County of San Francisco launched its open data portal, the DataSF, to: 1) stimulate new ideas and services; 2) increase internal data sharing; 3) simplify Sunshine Requests; 4) improve Data Quality; 5) reduce unwanted web traffic; and 6) change how data is used. Because the Data Catalog was the most important and time-consuming part of the project, the City and County of San Francisco developed a three-step method associated with the critical timeline and key tasks for creating the data catalog, which is illustrated in Figure 8.



	1 Identify Data Sources	2 Identify all Datasets	3 Complete Dataset Catalog
What is it	What are your Authoritative Data Sources? Databases Department spreadsheets Shared Drives Online applications	What are all the single datasets you can pull from the data sources?	Who can help you complete the catalog for each dataset? Dataset 1 Dataset 2 Dataset 3 ...
Useful Defintions	**Data Source:** Technology or system that stores data, including databases, named spreadsheets, information systems, business applications, etc.	**Dataset:** Contents of a single database table, worksheet or defined view; data is provided as a single combination of unique rows (or records) and corresponding columns (or fields) describing that row. **Example – Database:** A database may contain several data tables – each table could be a dataset. You can also create new datasets by combining data from different tables into a new table.	**Data Steward:** The individual in charge of individual databases, datasets, or information systems. This person has business knowledge of the data and can answer questiosn about the data. They can probably help you complete the data catalog.
Due Date	Dec 19	Jan 30	Mar 13

Figure 8. One Page Summary of Data Inventory Process for DataSF

These tasks had been distributed to each department for brainstorming lists of data sources, datasets, and metadata. During brainstorming, the following questions were provided to help identify and discover data sources:

- What information systems does your department use?
- What databases does your department use?
- What applications capture information or are used in your business processes?
- Are some data resources kept in spreadsheets (on shared or individual drives)?
- What information are we already publishing and where did that information come from?

In addition, for each of the data sources, the data steward provided at least the name and the brief description of the data sources and information that captures any technical details and point of contacts.

2.4 Conclusions

Based on the goals of this project and the information gathered during the literature review, the project team, in consultation with WisDOT, decided to perform a data inventory for WisDOT. It was determined that the information needed for the data inventory would be best collected through a survey sent to key WisDOT personnel and consultants. The survey questions were modeled after similar exercises described in the literature. It was also determined through the literature review, specifically State DOT best practices, that WisDOT is trailing some states with respect to enterprise level data management.

3 METHODOLOGY

This section will describe the methodology employed for performing the agency-wide data inventory, including the following sections: 1) the overview of the methodology (i.e., the application of a survey for data discovery within WisDOT); 2) survey design (i.e., tool and question design); 3) focus group and follow up strategy.

3.1 Overview

Despite the fact that the bulk of WisDOT data is stored and maintained in EDS-administered databases, there exist many datasets that are managed by individual business units at WisDOT or by outside vendors. These datasets are somewhat invisible to enterprise systems and often are in siloed databases and file systems, at vendor sites, or in the cloud. In order to gather information on these datasets, a data discovery survey was designed, tested, and distributed.

3.2 Survey Design

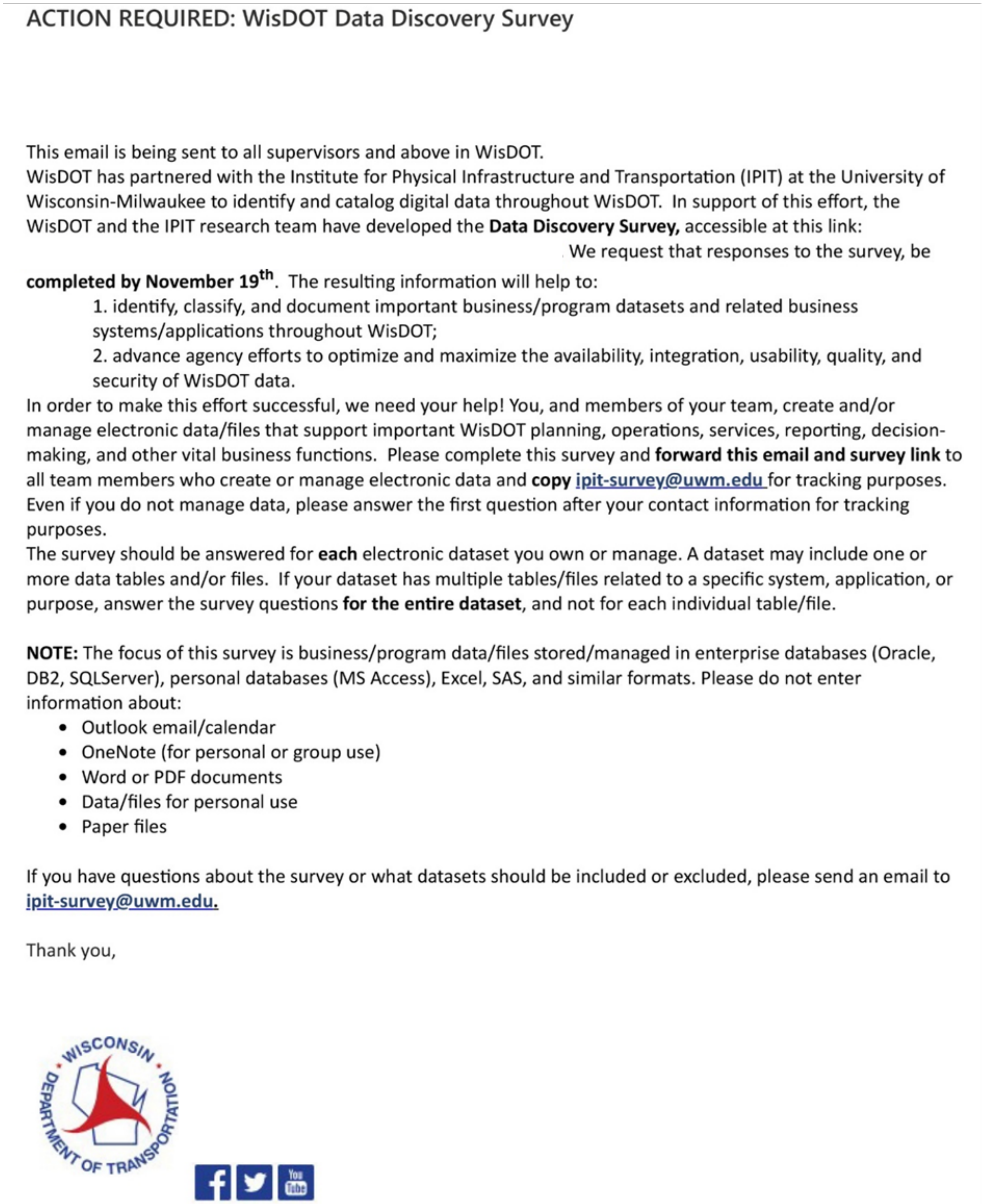
The data discovery survey was designed and administered through the UWM Qualtrics Survey tool. The questions were approved by WisDOT and included the following categories of questions listed below:

1. Contact information
    a. Name
    b. Email
2. Affiliation information
    a. Internal
        i. Organizational information (Executive Office/Division; Office/Bureau; Section; Unit)
    b. External
3. Data ownership
4. Detailed data information
    a. Data subject/domain
    b. Platform/format
    c. Location/storage
    d. Business metadata (including completeness of the metadata)
    e. Classification

Detailed survey information can be found in the Appendix A. Before distributing the survey to focus groups, the survey was tested by both the research team and a small pilot group within WisDOT.

3.3 Focus Group and Follow up Strategy

WisDOT's EDS Section within BITS and IPIT worked together to identify a list of people who might potentially own/manage data. The initial email for releasing the data discovery survey was sent to all supervisors and above in WisDOT (see email in Figure 9 below).



ACTION REQUIRED: WisDOT Data Discovery Survey

This email is being sent to all supervisors and above in WisDOT.
WisDOT has partnered with the Institute for Physical Infrastructure and Transportation (IPIT) at the University of Wisconsin-Milwaukee to identify and catalog digital data throughout WisDOT. In support of this effort, the WisDOT and the IPIT research team have developed the **Data Discovery Survey,** accessible at this link: We request that responses to the survey, be **completed by November 19th**. The resulting information will help to:

1. identify, classify, and document important business/program datasets and related business systems/applications throughout WisDOT;
2. advance agency efforts to optimize and maximize the availability, integration, usability, quality, and security of WisDOT data.

In order to make this effort successful, we need your help! You, and members of your team, create and/or manage electronic data/files that support important WisDOT planning, operations, services, reporting, decision-making, and other vital business functions. Please complete this survey and **forward this email and survey link** to all team members who create or manage electronic data and **copy** ipit-survey@uwm.edu for tracking purposes. Even if you do not manage data, please answer the first question after your contact information for tracking purposes.
The survey should be answered for **each** electronic dataset you own or manage. A dataset may include one or more data tables and/or files. If your dataset has multiple tables/files related to a specific system, application, or purpose, answer the survey questions **for the entire dataset**, and not for each individual table/file.

NOTE: The focus of this survey is business/program data/files stored/managed in enterprise databases (Oracle, DB2, SQLServer), personal databases (MS Access), Excel, SAS, and similar formats. Please do not enter information about:

- Outlook email/calendar
- OneNote (for personal or group use)
- Word or PDF documents
- Data/files for personal use
- Paper files

If you have questions about the survey or what datasets should be included or excluded, please send an email to ipit-survey@uwm.edu.

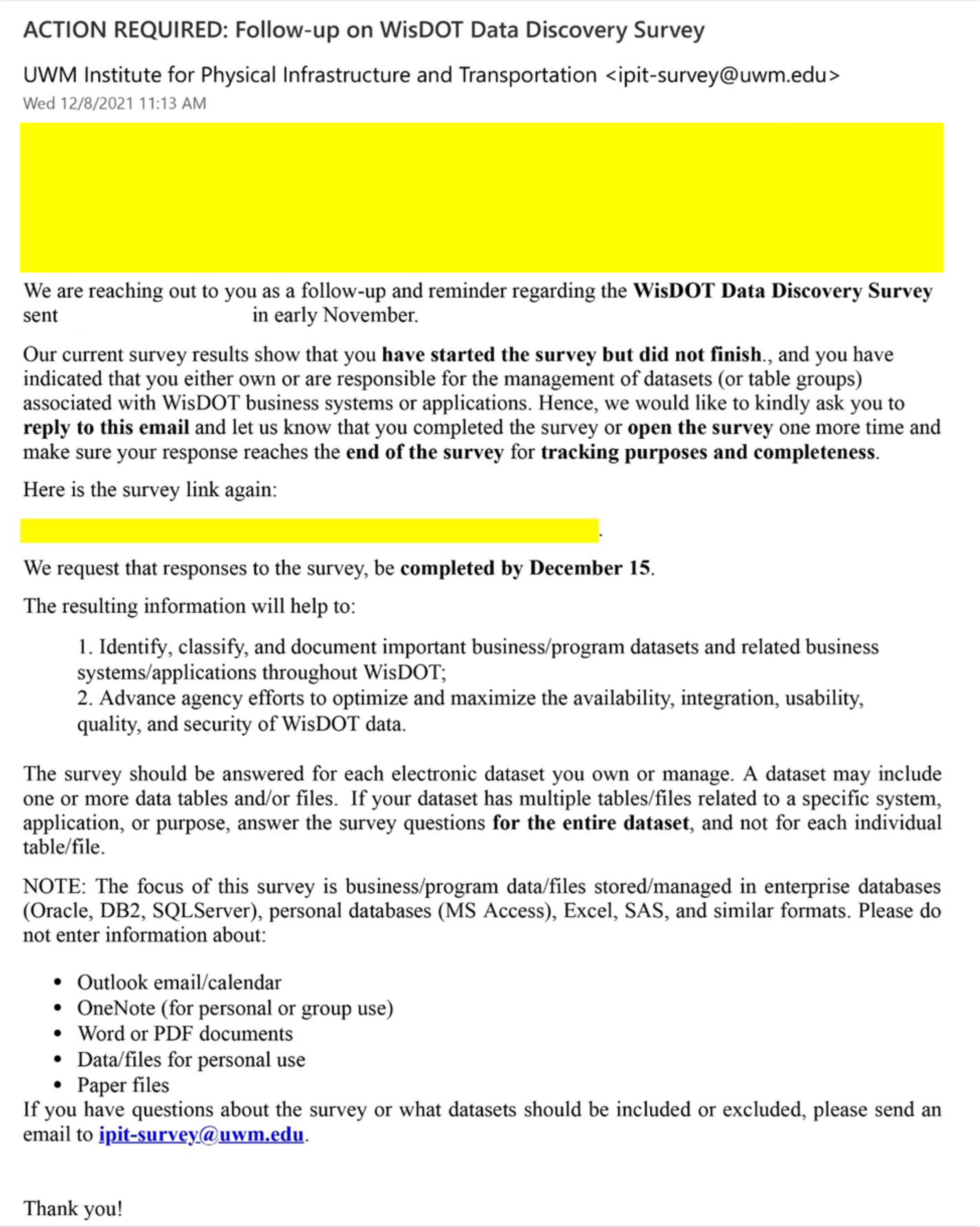
Thank you,

Figure 9. Initial Email for Releasing the Data Discovery Survey to WisDOT's Manager/Supervisor Level

After the initial email requesting participation in the data discovery survey, it was noted that not all recipients of the email responded to the survey within the given period of time. In order to increase participation and gather additional survey responses, the research team decided to conduct several follow-ups based on the preliminary analysis of the survey results, and the following selected groups were identified:

1. Supervisors who started the survey but have not finished
2. Supervisors who have not responded to the survey
3. Data stewards who have not participated in the initial survey

Corresponding emails have been designed and sent accordingly, which can be seen in the Figures 10-11.



ACTION REQUIRED: Follow-up on WisDOT Data Discovery Survey

UWM Institute for Physical Infrastructure and Transportation <ipit-survey@uwm.edu>
Wed 12/8/2021 11:13 AM

[redacted]

We are reaching out to you as a follow-up and reminder regarding the **WisDOT Data Discovery Survey** sent in early November.

Our current survey results show that you **have started the survey but did not finish**., and you have indicated that you either own or are responsible for the management of datasets (or table groups) associated with WisDOT business systems or applications. Hence, we would like to kindly ask you to **reply to this email** and let us know that you completed the survey or **open the survey** one more time and make sure your response reaches the **end of the survey** for **tracking purposes and completeness**.

Here is the survey link again:

[redacted].

We request that responses to the survey, be **completed by December 15**.

The resulting information will help to:

1. Identify, classify, and document important business/program datasets and related business systems/applications throughout WisDOT;
2. Advance agency efforts to optimize and maximize the availability, integration, usability, quality, and security of WisDOT data.

The survey should be answered for each electronic dataset you own or manage. A dataset may include one or more data tables and/or files. If your dataset has multiple tables/files related to a specific system, application, or purpose, answer the survey questions **for the entire dataset**, and not for each individual table/file.

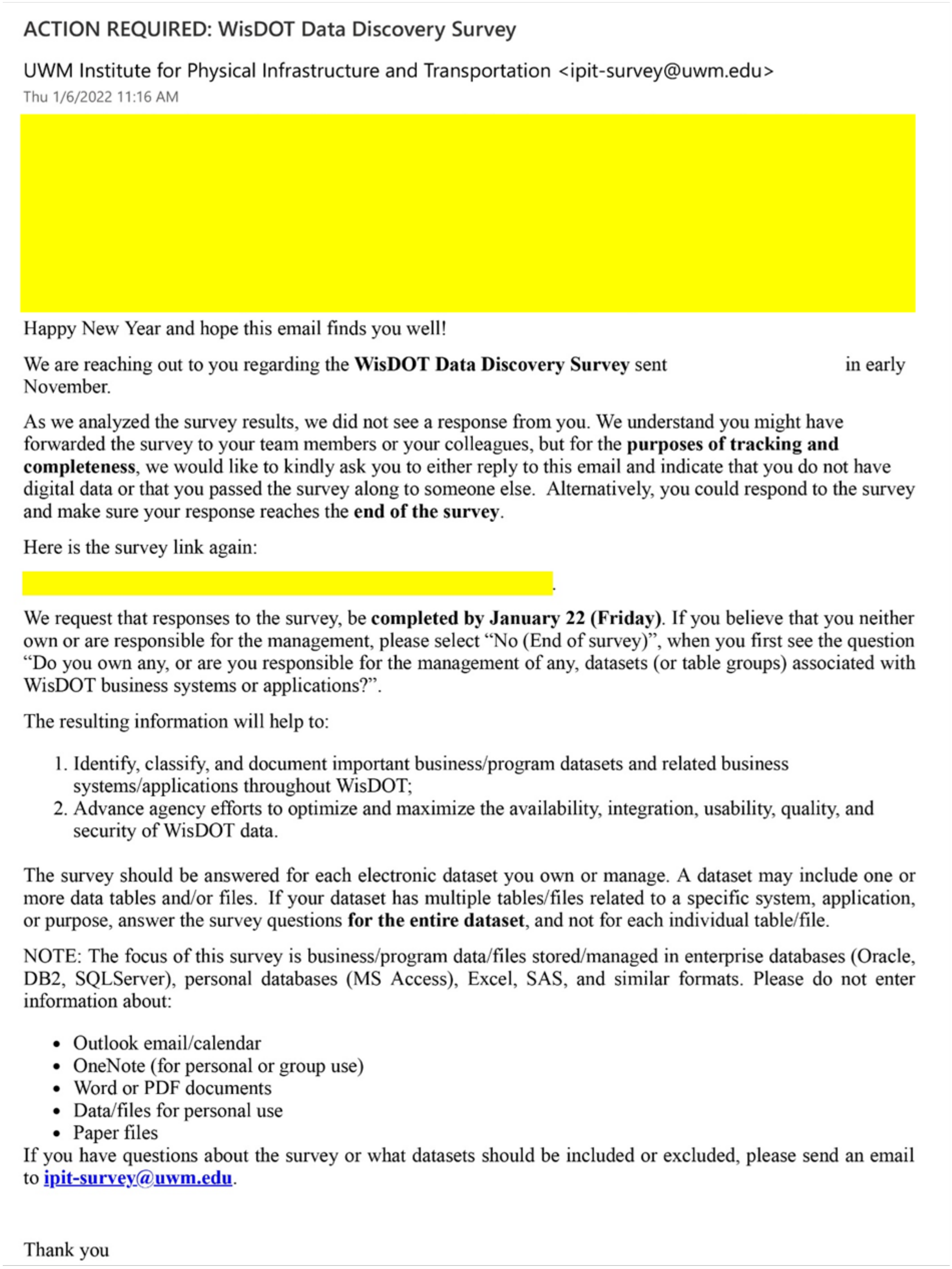
NOTE: The focus of this survey is business/program data/files stored/managed in enterprise databases (Oracle, DB2, SQLServer), personal databases (MS Access), Excel, SAS, and similar formats. Please do not enter information about:

- Outlook email/calendar
- OneNote (for personal or group use)
- Word or PDF documents
- Data/files for personal use
- Paper files

If you have questions about the survey or what datasets should be included or excluded, please send an email to **ipit-survey@uwm.edu**.

Thank you!

Figure 10. Follow-up Email for Supervisors Who Started but Haven't Finished the Survey



ACTION REQUIRED: WisDOT Data Discovery Survey

UWM Institute for Physical Infrastructure and Transportation <ipit-survey@uwm.edu>

Thu 1/6/2022 11:16 AM

[redacted]

Happy New Year and hope this email finds you well!

We are reaching out to you regarding the **WisDOT Data Discovery Survey** sent in early November.

As we analyzed the survey results, we did not see a response from you. We understand you might have forwarded the survey to your team members or your colleagues, but for the **purposes of tracking and completeness**, we would like to kindly ask you to either reply to this email and indicate that you do not have digital data or that you passed the survey along to someone else. Alternatively, you could respond to the survey and make sure your response reaches the **end of the survey**.

Here is the survey link again:

[redacted].

We request that responses to the survey, be **completed by January 22 (Friday)**. If you believe that you neither own or are responsible for the management, please select "No (End of survey)", when you first see the question "Do you own any, or are you responsible for the management of any, datasets (or table groups) associated with WisDOT business systems or applications?".

The resulting information will help to:

1. Identify, classify, and document important business/program datasets and related business systems/applications throughout WisDOT;
2. Advance agency efforts to optimize and maximize the availability, integration, usability, quality, and security of WisDOT data.

The survey should be answered for each electronic dataset you own or manage. A dataset may include one or more data tables and/or files. If your dataset has multiple tables/files related to a specific system, application, or purpose, answer the survey questions **for the entire dataset**, and not for each individual table/file.

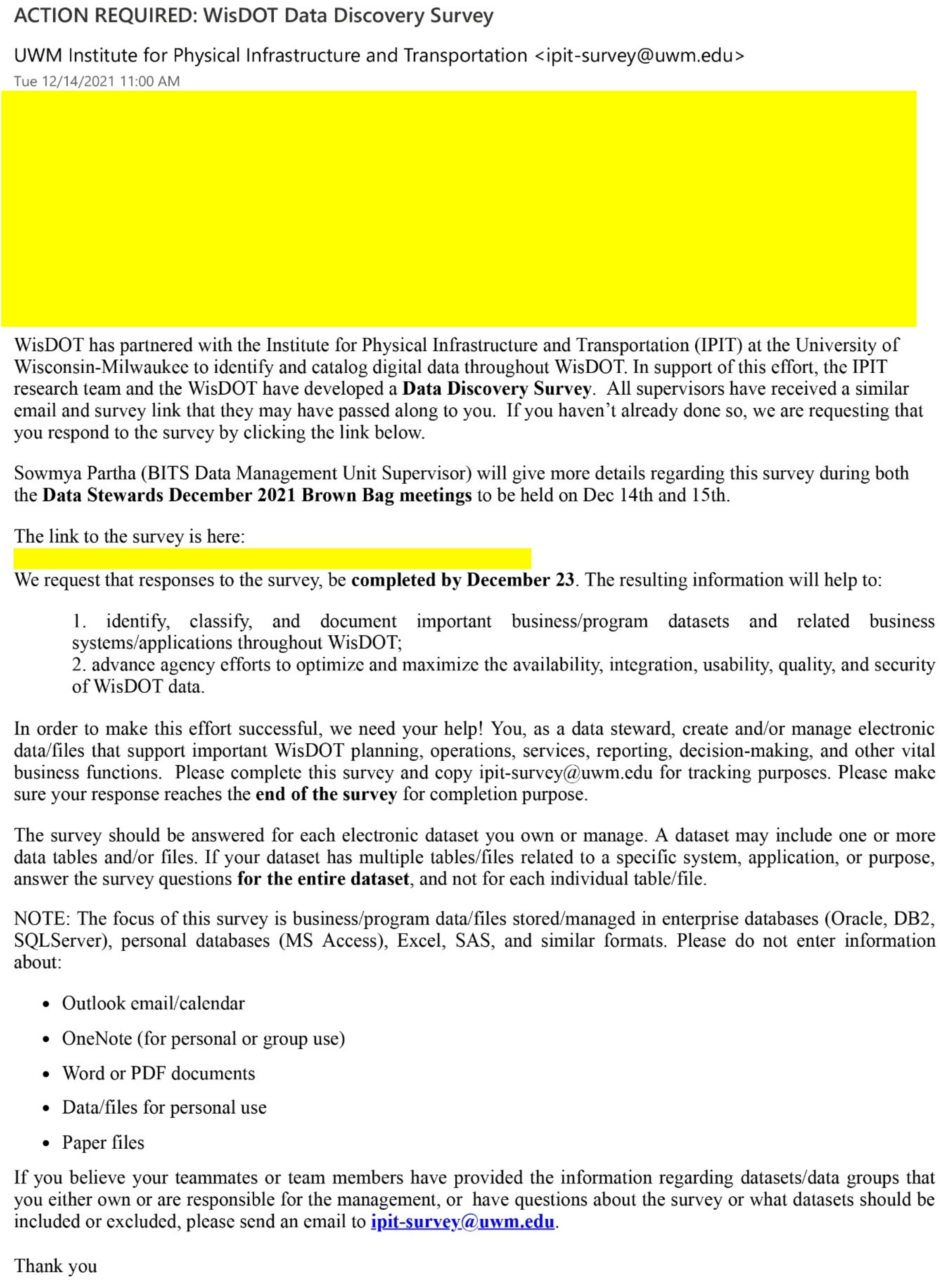
NOTE: The focus of this survey is business/program data/files stored/managed in enterprise databases (Oracle, DB2, SQLServer), personal databases (MS Access), Excel, SAS, and similar formats. Please do not enter information about:

- Outlook email/calendar
- OneNote (for personal or group use)
- Word or PDF documents
- Data/files for personal use
- Paper files

If you have questions about the survey or what datasets should be included or excluded, please send an email to **ipit-survey@uwm.edu**.

Thank you

Figure 11. Follow-up Email for Supervisors Who Haven't Responded to the Survey



ACTION REQUIRED: WisDOT Data Discovery Survey

UWM Institute for Physical Infrastructure and Transportation <ipit-survey@uwm.edu>

Tue 12/14/2021 11:00 AM

[redacted]

WisDOT has partnered with the Institute for Physical Infrastructure and Transportation (IPIT) at the University of Wisconsin-Milwaukee to identify and catalog digital data throughout WisDOT. In support of this effort, the IPIT research team and the WisDOT have developed a **Data Discovery Survey**. All supervisors have received a similar email and survey link that they may have passed along to you. If you haven't already done so, we are requesting that you respond to the survey by clicking the link below.

Sowmya Partha (BITS Data Management Unit Supervisor) will give more details regarding this survey during both the **Data Stewards December 2021 Brown Bag meetings** to be held on Dec 14th and 15th.

The link to the survey is here:
[redacted]

We request that responses to the survey, be **completed by December 23**. The resulting information will help to:

1. identify, classify, and document important business/program datasets and related business systems/applications throughout WisDOT;
2. advance agency efforts to optimize and maximize the availability, integration, usability, quality, and security of WisDOT data.

In order to make this effort successful, we need your help! You, as a data steward, create and/or manage electronic data/files that support important WisDOT planning, operations, services, reporting, decision-making, and other vital business functions. Please complete this survey and copy ipit-survey@uwm.edu for tracking purposes. Please make sure your response reaches the **end of the survey** for completion purpose.

The survey should be answered for each electronic dataset you own or manage. A dataset may include one or more data tables and/or files. If your dataset has multiple tables/files related to a specific system, application, or purpose, answer the survey questions **for the entire dataset**, and not for each individual table/file.

NOTE: The focus of this survey is business/program data/files stored/managed in enterprise databases (Oracle, DB2, SQLServer), personal databases (MS Access), Excel, SAS, and similar formats. Please do not enter information about:

- Outlook email/calendar
- OneNote (for personal or group use)
- Word or PDF documents
- Data/files for personal use
- Paper files

If you believe your teammates or team members have provided the information regarding datasets/data groups that you either own or are responsible for the management, or have questions about the survey or what datasets should be included or excluded, please send an email to **ipit-survey@uwm.edu**.

Thank you

Figure 12. Follow-up Email for Data Stewards Who Haven't Responded to the Survey

4 RESULT ANALYSES

This section presents the analysis of survey results and is organized as follows: 1) respondent profiles (i.e., at organizational level); 2) platforms and locations of datasets; 4) business metadata; 3) data security/classifications; 5) data subjects/domains.

4.1 Results of Respondent Profiles

The survey was sent to 580 people. At the time the survey was closed 422 people responded to the survey. Figure 13 shows how the responses were categorized. After removing blank surveys, i.e., no dataset managed by that person, and merged responses from the same person, we received survey responses from 297 people. Among all these 297 people, there are 280 finished responses and 17 unfinished (i.e., provided at least contact information). As denoted earlier, the survey is designed to gather information of datasets/data groups throughout WisDOT, so we further filtered both groups (finished vs. unfinished responses) based on respondents' input in terms of their data ownership. As a result, among 17 unfinished responses, 10 indicated they either own or manage datasets/data groups, but only 2 provided information for at least 1 dataset/data group. As for the 280 finished groups, 104 indicated they either own or manage datasets/data groups and the other 176 respondents answered "no" for data ownership. Finally, information for 230 datasets/data groups are provided by both groups, in which 3 were from the unfinished responses and 227 were from the finished responses.

The organizational level information of the respondents that participated in the survey is shown in Figure 14. As would be expected, 292 respondents are WisDOT employees and very few, only 5 respondents, are from outside WisDOT: Synergy, University of Wisconsin, Symphony Corporation, Michael Baker International, OES Consultant. As seen in Figure 14, for WisDOT employees, most respondents are from the Division of Transportation System Development (DTSD) with a total number of 157 respondents, 2 are from DTSD administration; 56 are from the Division of State Patrol (DSP); 26 are from the Division of Motor Vehicle (DMV); 22 are from Division of Transportation Investment Management (DTIM); 21 are from Division of Business Management (DBM); 3 are from Executive Offices; and 2 are from Division of Budget and Strategic Initiatives (DBSI).

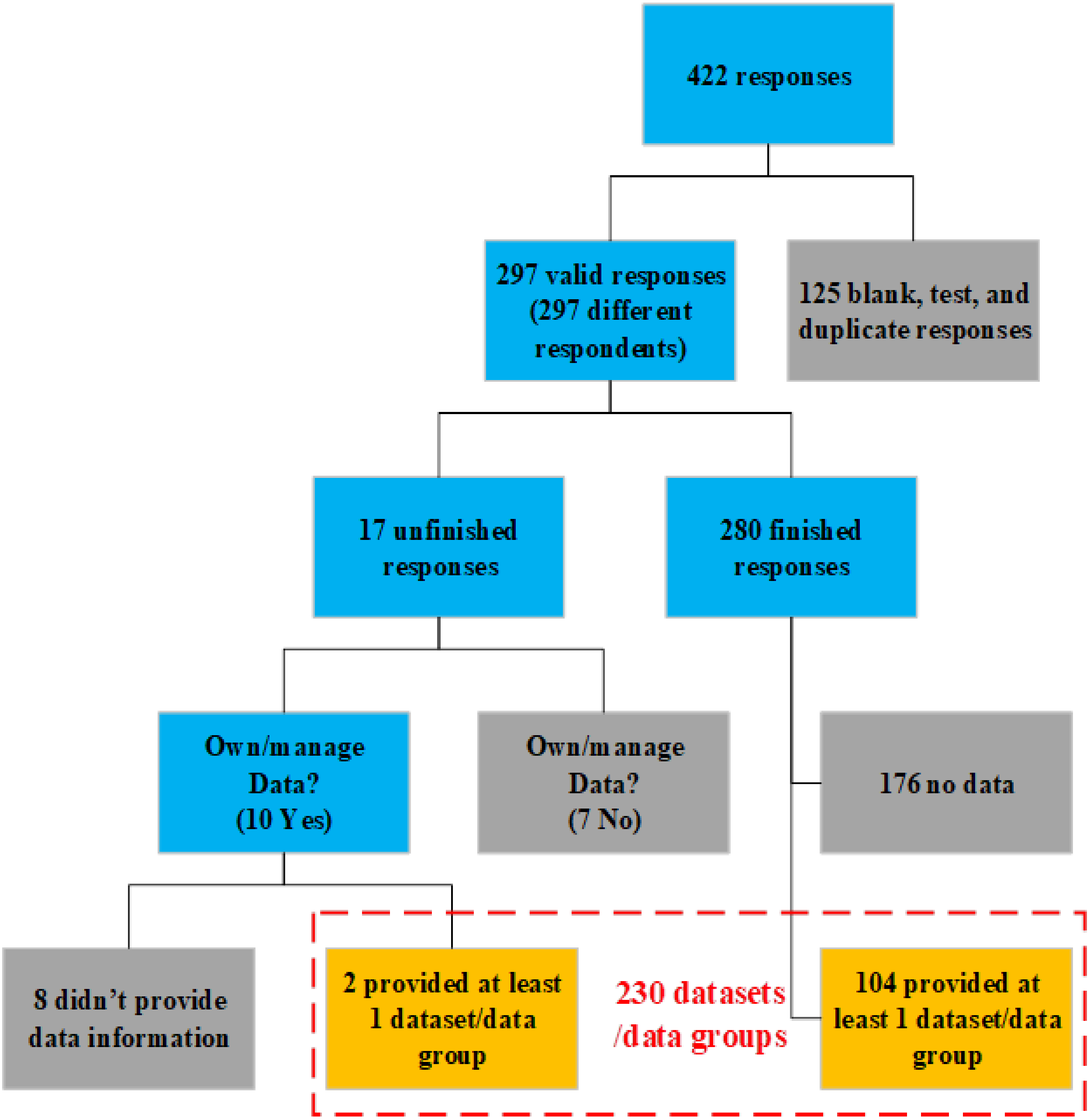

422 responses
297 valid responses (297 different respondents)
125 blank, test, and duplicate responses
17 unfinished responses
280 finished responses
Own/manage Data? (10 Yes)
Own/manage Data? (7 No)
176 no data
8 didn't provide data information
2 provided at least 1 dataset/data group
230 datasets /data groups
104 provided at least 1 dataset/data group

Figure 13. Profiles of the Survey Responses

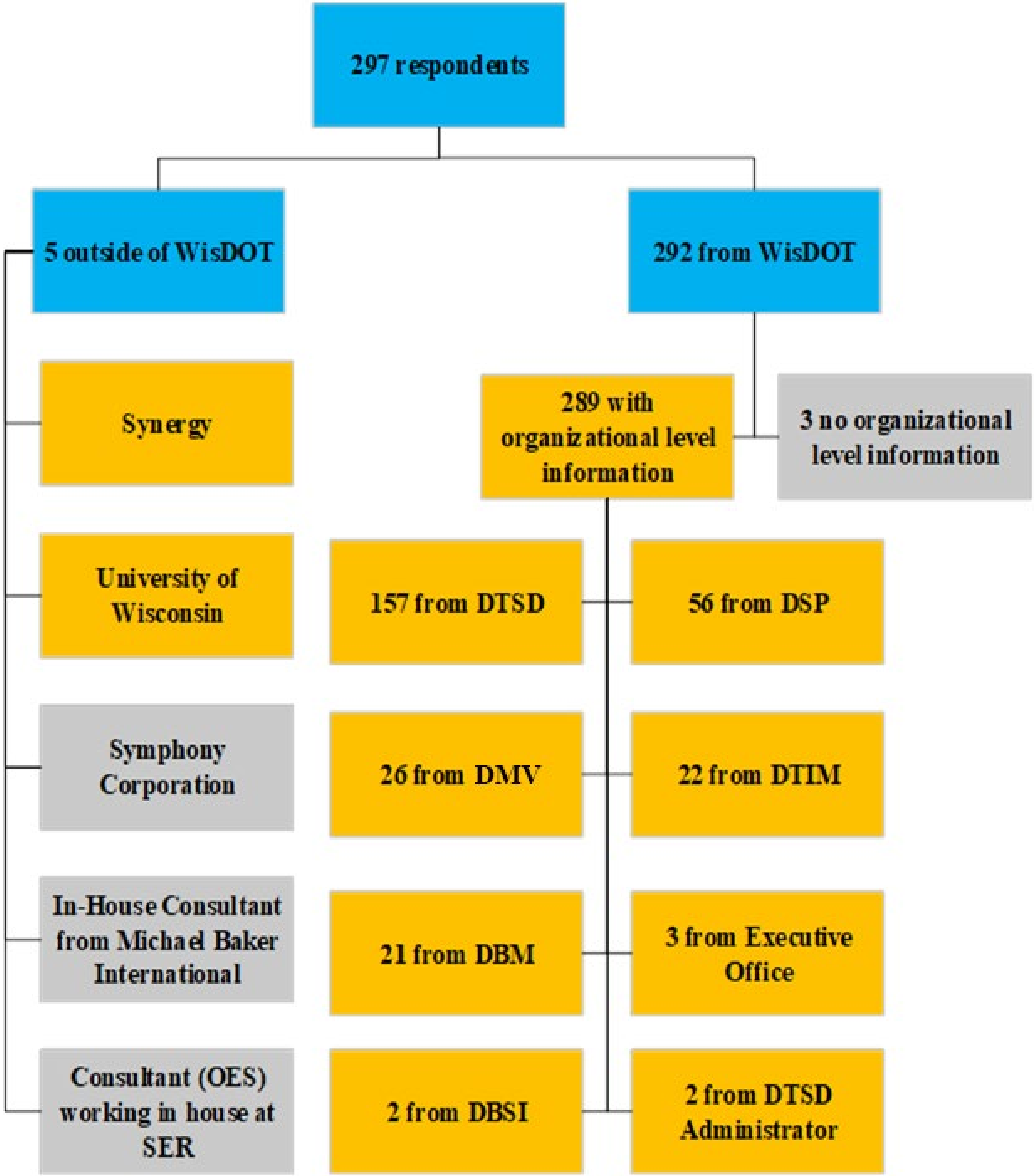

297 respondents
5 outside of WisDOT
292 from WisDOT
Synergy
University of Wisconsin
Symphony Corporation
In-House Consultant from Michael Baker International
Consultant (OES) working in house at SER
289 with organizational level information
3 no organizational level information
157 from DTSD
56 from DSP
26 from DMV
22 from DTIM
21 from DBM
3 from Executive Office
2 from DBSI
2 from DTSD Administrator

Figure 14. Organizational Distributions of Respondents

Furthermore, Figure 15 displays the information regarding datasets/data groups by different organizations. Similarly, most datasets/data groups are owned by DTSD with a total number of 121, and DTIM comes in second place having a total number of 47 datasets/data groups; \Compared to the number of 56 respondents, only 24 datasets/data groups have been provided by respondents from DSP. The remaining 38 datasets/data groups are provided by DBM (18), DMV (15), DBSI (2), University of Wisconsin (2), and Synergy (1).

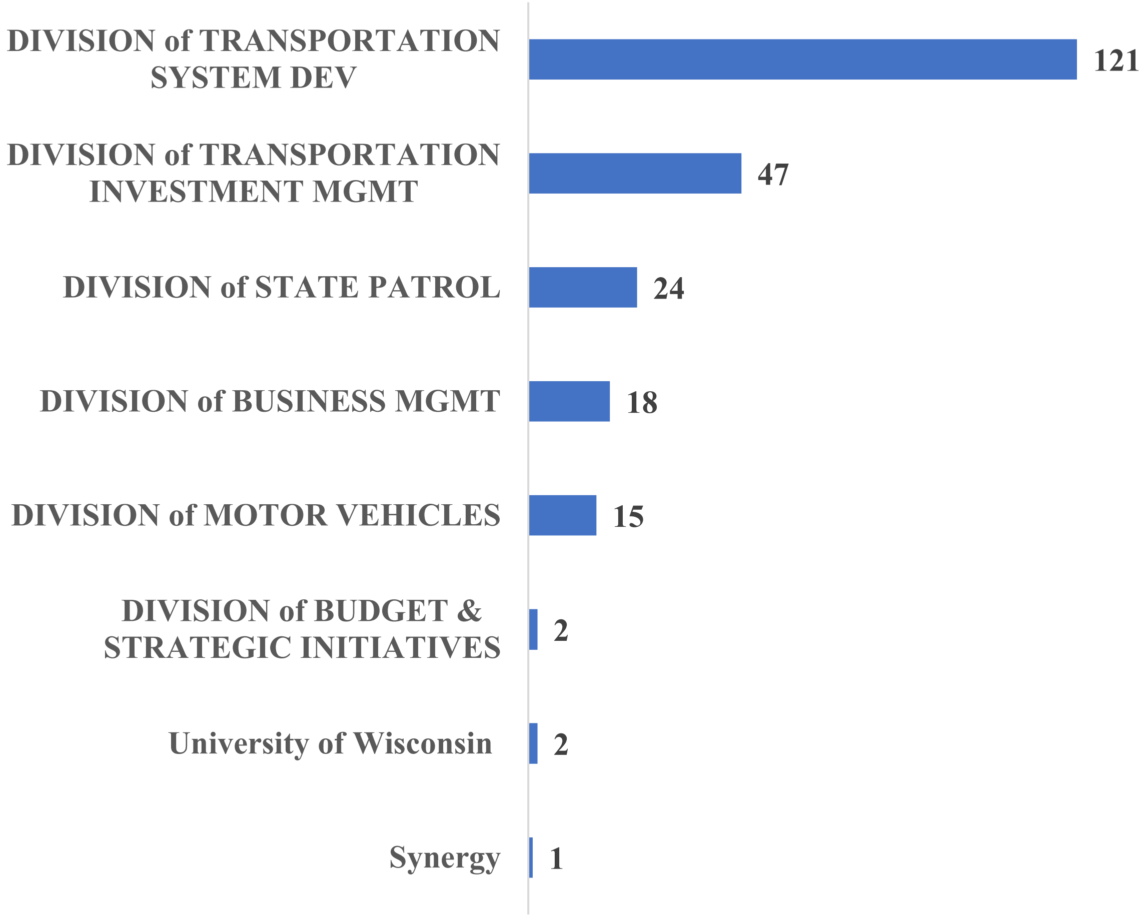


Figure 15. Datasets/Data Group by Organizations

Beyond the overview of the survey results presented, the following sections will provide analysis of the information focused on several areas including:

- Platform/format
- Location/storage
- Business metadata (including completeness of the metadata)
- Classification
- Data subject/domain

4.2 Platform and Locations

The platform defined in the survey is the software or application used for storing the dataset/data group. The pre-defined platforms in the survey include: 1) Oracle, 2) DB2, 3) Excel, 4) Access, 5) SQL Server, and 6) SAS Data Set. A freeform "Other" option was also provided in this survey question.

Figure 16 illustrates the results of platform distributions for surveyed datasets/data groups. Multiple selections were allowed for this question; therefore, the total number of platforms is slightly greater than the total number of the 230 surveyed datasets/data groups. Based on discussions with WisDOT's BITS EDS section, most data should be expected to be stored with either Oracle or DB2. However, it can be seen from Figure 16, though 78 datasets/data groups are stored by using Oracle and 12 datasets/data groups are stored by using DB2, the total number of such datasets/data groups are far less than half of the total 230 datasets/data groups identified in this survey. Despite the provided predefined answers, 57 datasets/data groups are claimed to be stored by using other non-standardized platforms, and within these 57 responses, there are 6 that do not know what platforms they are using. In other words, non-standardized platforms (including Excel, Access, SQL Server, SAS Data Set) are heavily used by WisDOT employees who oversee the data assets. It is worth mentioning that where data is stored has a great impact on data security and the capability and capacity of data linkage and data analysis, which would eventually affect the efficiency and effectiveness of the decision-making process within WisDOT.

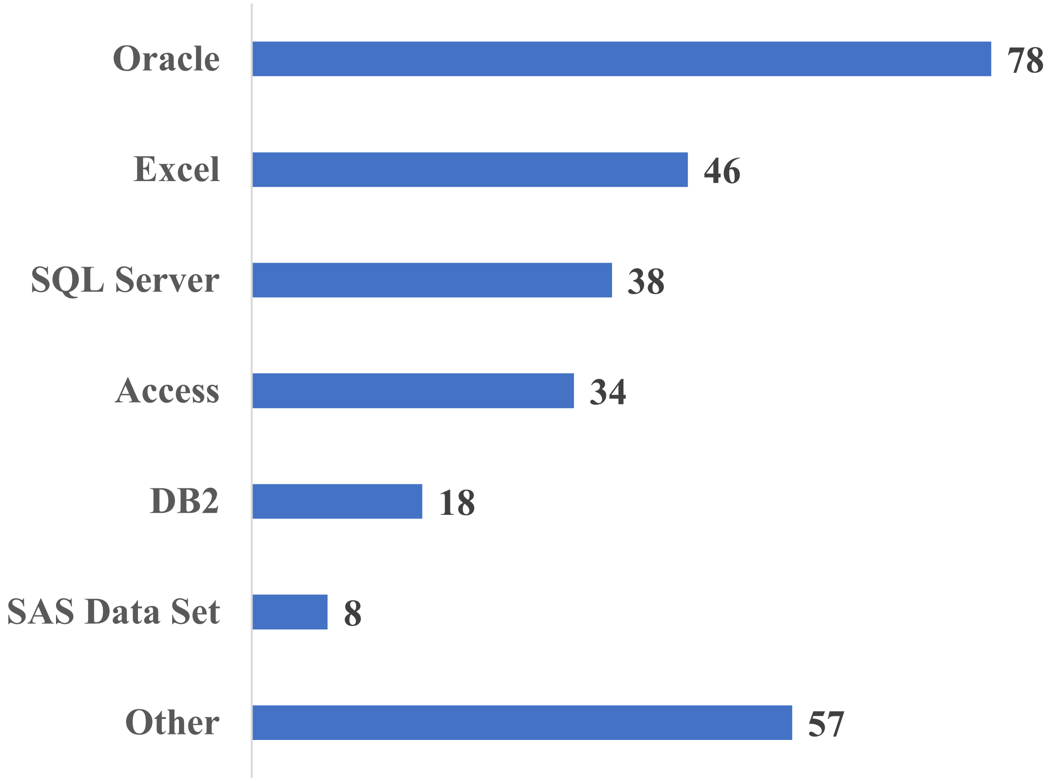


Figure 16. Results of Platform Distributions for Surveyed Datasets/Data Groups

In addition to the platform used for storing the datasets/data groups, the survey also included a specific question to understand the location of the data. The pre-defined answers include: 1) WisDOT server/Division of Enterprise Technology (DET) server, 2) Network drive, 3) Cloud service, 4) Business partner's server, and 5) Desktop/laptop. Similarly, "Other" and "Unsure" options were also provided for selection, and multiple selections were also allowed, which means the total number summed in the table is slightly greater than the 230 surveyed datasets/data groups.

Table 7 shows the summary data storage locations by storage types. It can be seen that the majority of the data have been stored in WisDOT server/Division of Enterprise Technology (DET) server, and though some respondents have selected "Other", they also indicated those datasets/data groups are stored in other private WisDOT servers/applications (i.e., answers of "DSP server", "DSP applications", and "Private Server, backed up by WisDOT"). However, other than the abovementioned storage locations, the number of "Other" selections might still lead to security concerns about data storage. From another perspective, the results might also indicate an issue of data being stored in multiple places throughout the whole organizations. The results from this question also indicate the need for enterprise level Data Governance.

Table 7. Summary of Data Storage Locations by Storage Types

Location of Data	No.
WisDOT server/Division of Enterprise Technology (DET) server	**117**
Network drive	**49**
Cloud service	**48**
Business partner's server	**32**
Desktop/laptop	**14**
Other	**39**
DSP server	13
BOX	10
DSP application	2
External hosted	2
Jackalope hosted by vendor High Desert Traffic	1
Internal Network Storage	1
Private Server, backed up by WisDOT	1
DET mainframe	1
SharePoint	1
with shop tech	1
Box, PeopleSoft, some in FIIPS, some in PMP	1
Local install to laptop	1
field counting devices	1
**Not provide specific location information*	3
Unsure	**7**

4.3 Business Metadata

In contrast to technical metadata, which is data used in the storage and structure of the data in a database or system, the business metadata describes the meaning of data, by defining terms in everyday language without regard to technical implementation. In order to retrieve such information, two specific questions were considered and designed to survey the respondents for understanding if the business metadata has been developed or not, and examining the completeness of the developed business metadata.

Figure 17 displays the number of datasets/data groups with and without business metadata. The results from this question are basically evenly split, which means half of the data might not be easily interpreted if irregular terms or coding schema are applied in the data. At the same time, such lack of metadata might also be a barrier for data linkage/sharing and prevents data analysts/engineers/researchers from querying, retrieving, and analyzing the data.

As for those 119 datasets/data groups which have developed business metadata, the results of the metadata completeness are illustrated in Figure 18. As seen in Figure 18, 36 datasets/data groups are claimed to have a perfectly complete business metadata documented. Total of 55 datasets/data groups claim to have relatively complete business metadata, more than 70% complete (i.e., 25 with more than 90%, 17 with more than 80%, and 13 with more than 70%). Compared to the total number of 230 identified datasets/data groups, such low numbers of metadata completeness point to the need for an enterprise-level business metadata development plan in WisDOT.

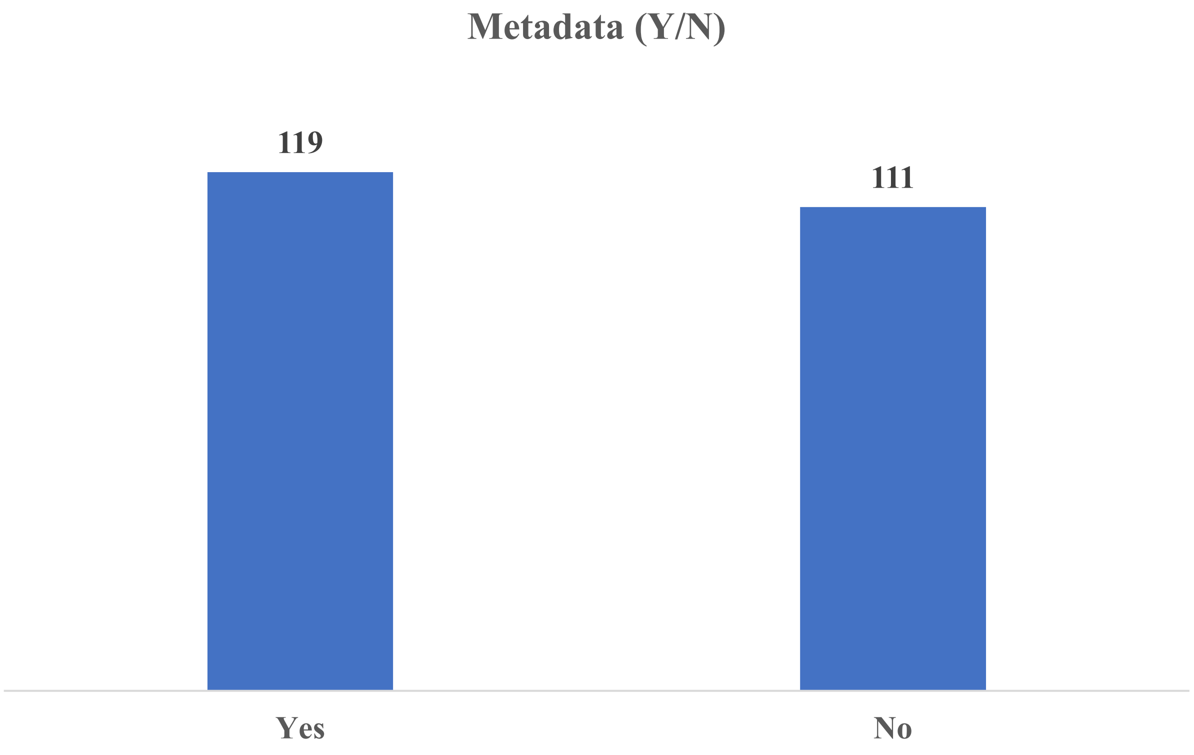


Figure 17. Number of Business Metadata Developed vs. Not

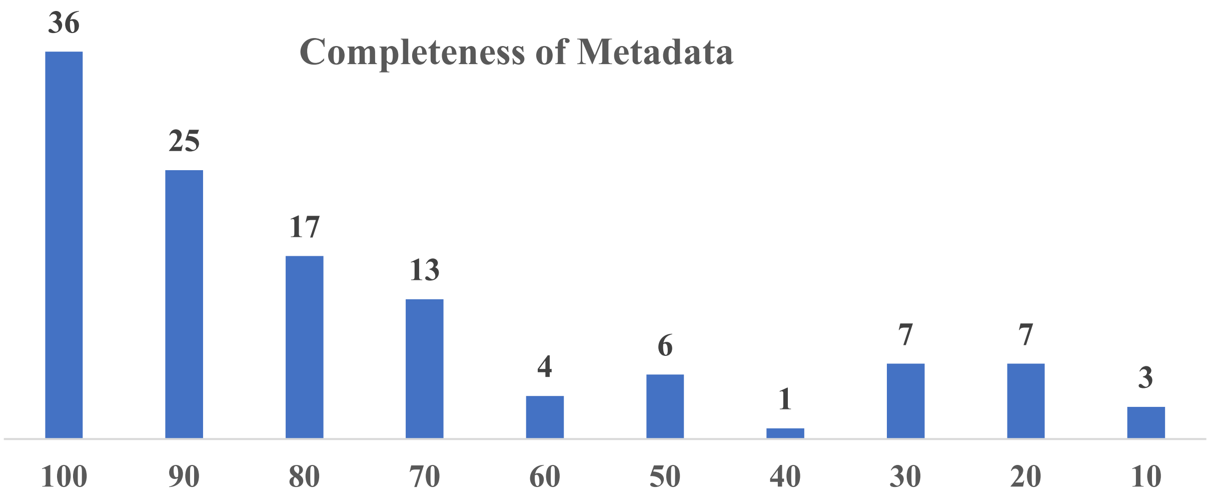


Figure 18. Distributions of Completeness of Metadata (%)

4.4 Data Security/Classifications

One of the most critical aspects of data management is regarding the sensitive information contained in data assets, which is also considered highly related to data security. Data classification supports the use of appropriate security, privacy, and compliance measures to protect: 1) Data confidentiality = prevent adverse impacts due to unauthorized data disclosure, 2) Data integrity = prevent adverse impacts due to unauthorized data modification/destruction, and 3) Data availability = prevent adverse impacts due to disruption of access or use of data.

Based on the level of sensitive information involved, a question related to the data classification provided 4 pre-defined answers, including: 1) Classified information, 2) Restricted information, 3) Sensitive information, and 4) Public information. The detailed descriptions for each option can be found in the Appendix A of the survey. Additionally, the option of "Unsure" was also provided. Table 8 shows the summary of data classifications by organization. It can be seen that 32 datasets/data groups have been identified as "Classified" and 15 datasets/data groups have unsure data classification by the respondents.

Therefore, examinations of answers of both "Classified information" and "Unsure" have been further conducted. According to the results, among 32 classified datasets/data groups, 12 do not have business metadata documents developed. Only 7 out of 32 have a complete business metadata document. For the 15 datasets/data groups identified with unsure status of their data classification, 2 out of the 15 have complete business metadata developed, which means these respondents should clearly know their data classification. However, their "unsure" answer points to the need to elevate data as a valuable asset in the DOT. Such phenomenon exhibits the knowledge gap regarding the basic terminologies of data governance and management. Further explorations should be done to investigate the reasons and potential training program might be provided to increase the awareness of the responsibility of data ownership in terms of data sensitivity.

Table 8. Summary of Data Classifications by Organization

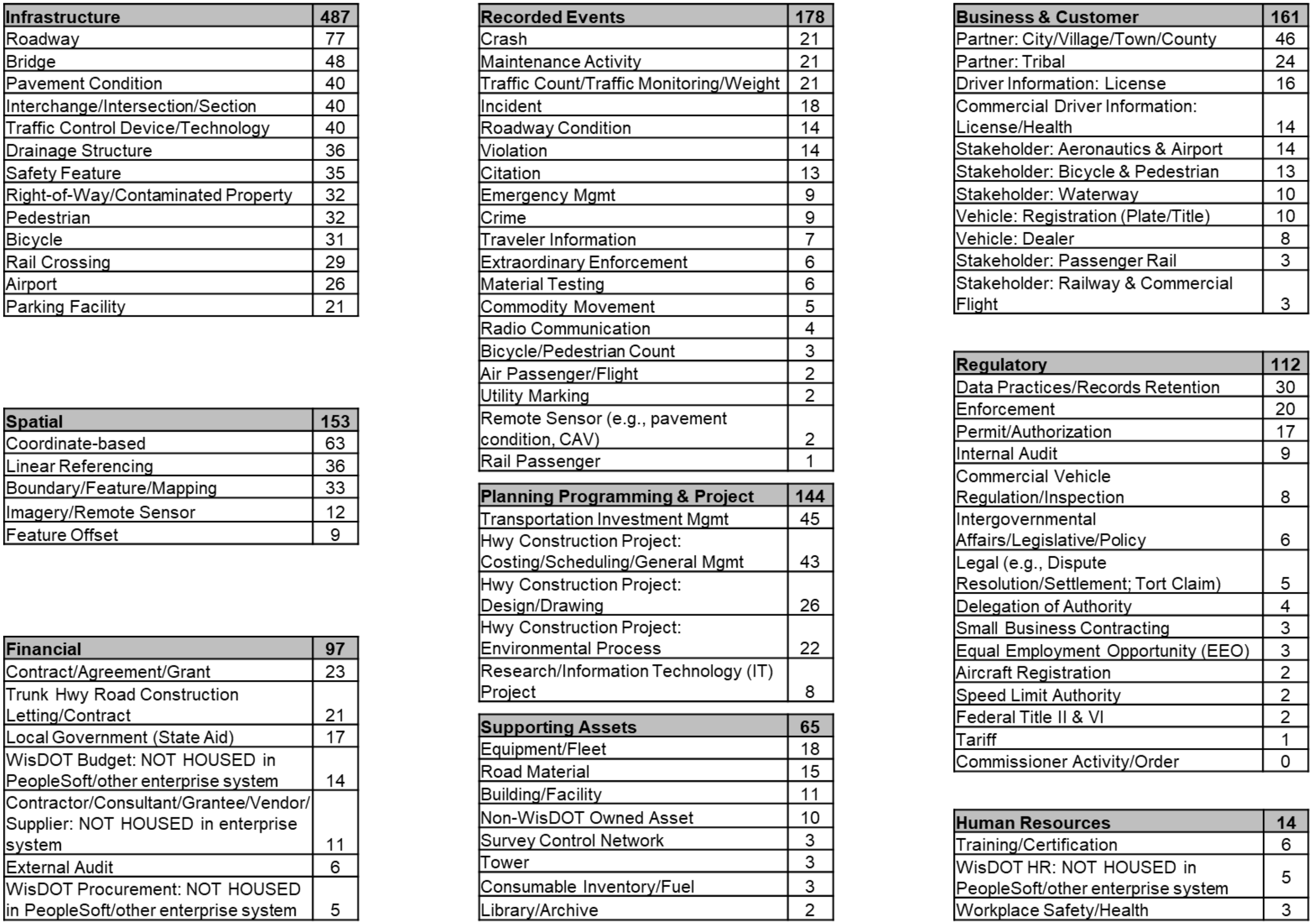
Organization	Public Information	Sensitive Information	Restricted Information	Classified Information	Unsure	Total
DIVISION of BUDGET & STRATEGIC INITIATIVES	2	0	0	0	0	2
DIVISION of BUSINESS MGMT	5	4	2	7	0	18
DIVISION of MOTOR VEHICLES	1	2	2	6	4	15
DIVISION of STATE PATROL	2	8	4	10	0	24
DIVISION of TRANSPORTATION INVESTMENT MGMT	13	22	10	0	2	47
DIVISION of TRANSPORTATION SYSTEM DEV	61	34	8	9	9	121
University of Wisconsin	0	2	0	0	0	2
Consultant	0	0	1	0	0	1
Total	84	72	27	32	15	230

4.5 Data Domains/Subject Areas

The question for surveying the data domains and subject areas is designed to accomplish the purpose for achieving the future data catalog/inventory by WisDOT. The provided data domains and data subject areas in the survey have been adopted from the practice of MnDOT's data governance plan and modified by the research team to better incorporate the data management schema applied by WisDOT, which include 9 domains 85 subject areas (detailed options can be found in Figure 19 and in the Appendix A of the detailed survey). Such information could be beneficial to many aspects of data governance, including but not limited to: 1) data management, 2) data query, and 3) data sharing.

Figure 19 shows the summary of statistical results of data domains and subject areas for the surveyed datasets/data groups. A general observation shows the diversity of the data owned/managed by WisDOT and almost all data subject areas were selected at least once in the survey. This validates the effectiveness of the designed survey, which is able to capture the majority of characteristics of data throughout WisDOT. In addition, there were 59 datasets/data groups that were identified as "Other – Data Domains/Data Subjects" that have not been listed in the survey. Potentially, such inputs from the respondents offer great values for tailoring WisDOT's own data catalog/inventory in the near future.

Once again, the results of this question show a high level of complexity associated with WisDOT's data assets. This should be considered when developing a data catalog/inventory plan and will require effort and resources.



Infrastructure	487
Roadway	77
Bridge	48
Pavement Condition	40
Interchange/Intersection/Section	40
Traffic Control Device/Technology	40
Drainage Structure	36
Safety Feature	35
Right-of-Way/Contaminated Property	32
Pedestrian	32
Bicycle	31
Rail Crossing	29
Airport	26
Parking Facility	21

Spatial	153
Coordinate-based	63
Linear Referencing	36
Boundary/Feature/Mapping	33
Imagery/Remote Sensor	12
Feature Offset	9

Financial	97
Contract/Agreement/Grant	23
Trunk Hwy Road Construction Letting/Contract	21
Local Government (State Aid)	17
WisDOT Budget: NOT HOUSED in PeopleSoft/other enterprise system	14
Contractor/Consultant/Grantee/Vendor/Supplier: NOT HOUSED in enterprise system	11
External Audit	6
WisDOT Procurement: NOT HOUSED in PeopleSoft/other enterprise system	5

Recorded Events	178
Crash	21
Maintenance Activity	21
Traffic Count/Traffic Monitoring/Weight	21
Incident	18
Roadway Condition	14
Violation	14
Citation	13
Emergency Mgmt	9
Crime	9
Traveler Information	7
Extraordinary Enforcement	6
Material Testing	6
Commodity Movement	5
Radio Communication	4
Bicycle/Pedestrian Count	3
Air Passenger/Flight	2
Utility Marking	2
Remote Sensor (e.g., pavement condition, CAV)	2
Rail Passenger	1

Planning Programming & Project	144
Transportation Investment Mgmt	45
Hwy Construction Project: Costing/Scheduling/General Mgmt	43
Hwy Construction Project: Design/Drawing	26
Hwy Construction Project: Environmental Process	22
Research/Information Technology (IT) Project	8

Supporting Assets	65
Equipment/Fleet	18
Road Material	15
Building/Facility	11
Non-WisDOT Owned Asset	10
Survey Control Network	3
Tower	3
Consumable Inventory/Fuel	3
Library/Archive	2

Business & Customer	161
Partner: City/Village/Town/County	46
Partner: Tribal	24
Driver Information: License	16
Commercial Driver Information: License/Health	14
Stakeholder: Aeronautics & Airport	14
Stakeholder: Bicycle & Pedestrian	13
Stakeholder: Waterway	10
Vehicle: Registration (Plate/Title)	10
Vehicle: Dealer	8
Stakeholder: Passenger Rail	3
Stakeholder: Railway & Commercial Flight	3

Regulatory	112
Data Practices/Records Retention	30
Enforcement	20
Permit/Authorization	17
Internal Audit	9
Commercial Vehicle Regulation/Inspection	8
Intergovernmental Affairs/Legislative/Policy	6
Legal (e.g., Dispute Resolution/Settlement; Tort Claim)	5
Delegation of Authority	4
Small Business Contracting	3
Equal Employment Opportunity (EEO)	3
Aircraft Registration	2
Speed Limit Authority	2
Federal Title II & VI	2
Tariff	1
Commissioner Activity/Order	0

Human Resources	14
Training/Certification	6
WisDOT HR: NOT HOUSED in PeopleSoft/other enterprise system	5
Workplace Safety/Health	3

Figure 19. Statistical Results of Data Domains and Subjects for the Surveyed Datasets/Data Groups

5 REVIEW OF DATA CATALOG SOFTWARE VENDORS

This section reviews industry-leading data catalog software/tools and uses published literature to put forward relevant ranking criteria to evaluate data cataloging software/tools including compatibility criteria with the existing software ecosystem at WisDOT.

A data catalog, or more specifically an enterprise data catalog, is an organized inventory of metadata about data assets in an organization. As the amount of data increases in an organization, the need for organized data management increases in terms of: data accessibility, quality, and timeliness of updates. For instance, state DOTs use structured transportation datasets to run their organizations, and transportation data is generally created in a free and unrestricted manor to meet the needs of individual business units. Without a data catalog in place for data management, it is virtually impossible for an organization the size of a state DOT to discover unknown data benefits and connections within their large amount of data. Therefore, this is not a question of if a DOT should have a data catalog, but rather when will they have a data catalog. Moreover, while there is a cost to create and maintain metadata in a data catalog, the benefit of increased exploration, connections, and productivity outweighs the cost.

Data catalog applications are tools for storing and searching metadata. Data catalog tools should be the go-to tool for organizations with complex and distributed data capture, storage, and use requirements, such as a state DOT. The key feature of a data catalog application is to provide metadata context to the user in a way that allows, for example, different departments within WisDOT (both IT and Non-IT) to discover, understand, and even analyze relevant data. Table 9 shows the 30 data catalog tools recommended by DBMS.COM, which are advanced data cataloging software that can solve data profiling, data lineage, and data classification problems, as well as open-source data catalog tools at the same time. According to this table, the important functions of data catalog tools should be evaluated through the following aspects:

1. Automated Cataloging: the ability of cataloging input data automatically
2. Business Glossary: the ability of adapting user-defined business terms/vocabularies
3. Commenting/Community: the availability of education, community resources, and communication channels
4. Commercial/Free: the affordability of the production of the product
5. Data Classification: the capability of classifying data
6. Data Lineage: the process of understanding, recording, and visualizing data as it flows from data sources to consumption (this includes all transformations the data underwent along the way—how the data was transformed, what changed, and why)
7. Data Profiling: the ability of examining, analyzing, and creating useful summaries of data in order to discover data quality issues, risks and overall trends
8. Rating of Assets: the ability of evaluating the data quality.

It should be noted that data catalogs are driven off of metadata; therefore, consistency in metadata increases the power of a data catalog. Manual metadata creation for a data catalog is time consuming and can contain bias; therefore, automated data cataloging for metadata harvest with applying tools such as Artificial Intelligence (AI) algorithms, could be a basic functional requirement to a catalog application. It is relatively easy for a data catalog application to automatically capture technical metadata (schema, tables, columns, type, etc.) and operational metadata (data about: collection, extract transform load (ETL) process, table use and by whom, etc.). However, the incorporation of business metadata, which is knowledge such as descriptions, comments, and annotation is difficult to automatically populate; therefore, a mechanism for subject matter experts to add to the metadata in the catalog should be provided. As discussed in Section 4 with the results of the analysis on the survey responses, the issue of knowledge gaps of data owners/managers in WisDOT regarding the basic terminologies of data governance and management is evident. This data management knowledge gap could be mitigated if the applied data catalog tool has the ability of adapting user-defined business terms/vocabularies (i.e., “Business Glossary”) to common terms that everyone in an organization can understand. By combining such functions further with the capability of data profiling, it might also be helpful to improve the current situation of the poor metadata management in WisDOT. From another long-term perspective, the ability of evaluating data quality (“Rating of Assets”) could aid data maintenance by identifying problematic data and their associated issues.

In order to provide a comprehensive review of data catalog tools, Table 10 displays another recommendation on the top 10 data catalog tools (Harvey, 2021). Unlike many articles that only provide the basic descriptions of each recommended data catalog tool, this article also provides some more analyses regarding both pros and cons of the data catalog tools. Compared to the previous recommendation, four more catalog tools are provided, including: Alex Solutions, Data.world, Hitachi Vantara, and Infogix.

Table 9. 30 Recommended Data Catalog Tools Adopted from (DBMSTOOLS.COM, n.d.)

Data Catalog Tools	Automated Cataloging:	Business Glossary:	Commenting/ Community:	Commercial/ Free:	Data Classification:	Data Lineage:	Data Profiling:	Rating of Assets:
Collibra Catalog	√	√	√	Commercial	√	√	√	√
Dataedo	√	√	√	Commercial	√	×	√	×
Alation Data Catalog	√	√	√	Commercial	√	√	√	-
Informatica Enterprise Data Catalog	√	√	√	Commercial	√	√	√	√
Redgate SQL Data Catalog	√	×	√	Commercial	√	×	√	×
Lumada Data Catalog	√	√	√	Commercial	√	√	√	√
IBM Watson Knowledge Catalog	√	√	√	Commercial	√	√	√	√
Talend Data Catalog	√	√	√	Commercial	√	√	√	√
Ataccama Metadata Management & Data Catalog	√	√	√	Commercial	√	√	√	×
Apache Atlas	×	√	×	Free	√	√	×	×
OvalEdge	√	√	√	Commercial	√	√	√	×
Alteryx Connect	√	√	√	Commercial	√	√	√	√
Truedat	√	√	×	Free	√	√	√	×
Cloudera Data Catalog	√	√	√	Commercial	√	√	√	√
Data3Sixty Govern	√	√	√	Commercial	√	√	√	√
Io-Tahoe	√	√	√	Commercial	√	√	√	√
Dawizz MYDATACATALOGUE	√	×	×	Commercial	×	×	√	×
SAP Data Intelligence	×	×	×	Commercial	√	√	√	×
Alex Solutions	×	√	×	Commercial	×	√	√	×
Oracle Cloud Infrastructure Data Catalog	√	√	√	Commercial	√	√	√	√
Google Cloud Data Catalog	√	√	×	Commercial	√	×	×	×
Octopai	√	√	×	Commercial	×	√	×	×
Azure Data Catalog	√	√	√	Commercial	×	×	√	×
Qlik Data Catalyst	√	√	√	Commercial	√	√	√	√
erwin Data Catalog	√	√	√	Commercial	√	√	√	×
Global IDs	√	√	-	Commercial	√	√	√	×
Tree Schema	√	√	√	Commercial	√	√	√	×
Atlan	√	√	√	Commercial	√	√	√	√
Sidecar	√	√	√	Commercial	√	√	×	√
Stemma	√	√	√	Commercial	×	√	√	√

Table 10. Top 10 Recommended Data Catalog Tools Adopted from (Harvey, 2021)

Data Catalog Tools	Pros	Cons
Alation	* Excellent machine learning (ML) capabilities * Good collaboration features * Pioneering innovation	* High pricing * Buggy releases * Poor data lineage features
Alex Solutions	* Broad capabilities * East to deploy and use * Excellent lineage profiling	* Difficult to integrate with business intelligent BI and data science tools * Poor collaboration capabilities * Needs better training
Collibra	* Excellent data intelligence and graph technology * Good for complex environments * Strong partner ecosystem	* Not user-friendly * Cloud-only * Occasional bad service
Data.world	* Upfront pricing * Easy to use * Public benefit corporation	* Immature product * Limited integrations
Erwin	* Broad data governance capabilities * Good data modelling capabilities * Large ecosystem	* Difficult deployment * High pricing * Not user-friendly
Google Cloud Data Catalog	* Integration with other Google Cloud products * Highly scalable * Affordable	* Doesn't integrate with all data sources * Difficult to estimate price accurately
Hitachi Vantara	* Advance ML and behavioral intelligence * Excellent lineage analysis * User-friendly	* Limited data governance abilities * Limited connectors * Poor collaboration capabilities
Infogix	* Wide range of features * Quantifies data value * Easy to use	* Limited analytics capabilities * Poor handling of large data sets * Inadequate documentation
Informatica	* Integration with other Informatica tools * Metadata intelligence engine * Highly scalable	* Difficult deployment * Limited data governance capabilities * High Total-Cost-of-Ownership (TCO)
IBM	* Integration with other IBM products * Flexible deployment options * Upfront pricing for cloud deployment	* Not user-friendly * Difficult deployment * High pricing

6 CONCLUSIONS AND FUTURE WORK

The WisDOT Data Inventory\Catalog research project completed by the Institute for Physical Infrastructure and Transportation (IPIT) at the University of Wisconsin-Milwaukee (UWM) focused on finding digital datasets and pertinent information about those datasets throughout WisDOT. The IPIT team members met with WisDOT team members on a biweekly basis throughout the duration of the project. The project team reviewed literature related to transportation data management, and more widely, best practices in data management for organizations of comparable size. From this review, data discovery/inventory was identified as the first step in understanding the depth and breadth of data assets within an organization. The main mechanism employed to find datasets at WISDOT during this project was a Qualtrics survey designed by the project team. The survey comprised 12 questions about each dataset and was distributed to 580 people, both within WisDOT and to consultants working for WisDOT. A return rate of 51.20% was achieved. It should be noted that participants were instructed to pass the survey link to appropriate data managers; therefore, the research team was not expecting everyone who received the survey to complete the survey. Through this process the research team identified 230 datasets associated with 106 people. The survey responses came from business units across the department and touched all areas of WisDOT data. The responses from the survey were aggregated and are presented in Section 4. From the analysis of the survey results several conclusions can be drawn.

As WisDOT strives to be a data-driven decision-making organization, centralized and consistent information about datasets throughout the entire enterprise becomes increasingly important. Much of the data at WisDOT have been collected, stored, and analyzed to meet specific goals or requirements. While this approach works, it creates siloed data, individualized schema, ad hoc metadata, and could be a security risk if the data is not stored or used appropriately. Alternatively, the benefits to an organization with robust data governance and data catalog are: multi-dataset analytics, new insights into data, new connections between data, the ability for data discovery, and improved data quality and data security.

Based on this work, the following conclusions are advanced:

- Data ownership is appropriately distributed in business units that need these data, but there is a lack of consistency in the manner in which data is stored, classified, and described. The lack of standardization in metadata is inconsistent with the goal of being a data-driven organization.
- Data owners are very familiar with their data, but are unfamiliar with basic data governance terminology and practices. As would be expected, WisDOT has many domain experts that focus on transportation and business data needed in their area. These employees were hired and trained on specific subject matter in their units. It is evident that individual data owners lack a "big picture" view along with the knowledge, training, bandwidth and incentives needed to properly create and manage metadata that add value and minimize risk to the enterprise.

- Currently there is no centralized data governance structure or data catalog software at WisDOT. This is evident by reviewing the variety of answers related to: data storage and platforms, range of documented metadata and schema, and responses related to data classification and data security. It was also observed that some respondents did not know the primary source of the data they reported, which can lead to inaccurate analyses based on outdated data.

Future work

Future work recommendations are divided into: Immediate action, and Next steps. The overarching goal of these recommendations is to minimize data risk and maximize enterprise data value.

An immediate action is to investigate self-identified classified data from the survey. This data should be investigated for: 1) is the data truly classified, 2) is metadata associated with classified data available and current.

Following the classified data, other anomalies identified during the data inventory analysis should be individually investigated. These include unconventional applications, uses, and storage methods of WisDOT data.

Next step recommendations are centered around data governance and data cataloging. The creation of rules, processes, and accountability around data is the next step to move WisDOT closer to the benefits of robust information management. Utilizing the best practices for data governance from DOTs and other similar sized agencies, a draft data governance framework should be developed. The draft data governance framework should strive to: harmonize data sources, properly control access, document ownership, and have a mechanism to capture both technical and descriptive information about the who, what, where, when, and why of enterprise data. During the drafting of the data governance documentation, existing WisDOT data rules and policies should be reviewed and included along with identifying any data rules other agencies have that would benefit WisDOT.

Enterprise data governance requires champions at the highest levels of an organization along with a data governance leader, committee, documentation, and implementation rules about "What and how should things related to data happen." It is expected that a Data Governance Director would be a full time equivalent. In addition, resources will be needed for a data catalog software package and the initial training and population of the catalog.

As data governance is being developed, a data catalog system should be selected and stood up. A dedicated person whose job is a Data Catalog Administrator, should oversee and facilitate the collection, population, and use of the data catalog information. It is expected that a Data Catalog Administrator would be a full time equivalent. Classified as well as mission critical data should be cataloged first, followed by datasets that already have complete or nearly complete metadata. Finally, standalone and off-site datasets should be cataloged. During this process, data governance rules should be implemented which will move WisDOT to standardize data management across the agency and increase the value of its data.

REFERENCES

Albee, M., Hamilton, I., & Chestnutt, C. (2020). *Data Governance: Ohio's People, Processes, and Technology* (No. FHWA-SA-20-059). United States. Federal Highway Administration. Office of Safety.

Bhat, C. R., Claudel, C., Ruiz-Juri, N., Perrine, K. A., Lei, T., Long, K., & Mohamed, A. (2019). *Weather-Savvy Roads: Sensors and Data for Enhancing Road Weather Management* (No. FHWA/TX-19/5-9053-01-1). University of Texas at Austin. Center for Transportation Research.

Causseaux, J. (2019). Florida DOT Implementation of Data Governance. In *GIS-T Conference, April*.

Christian, C. (2020). *FDOT Data Governance Initiative: Managing Vital Data Assets* (No. FHWA-HIF-20-064). United States. Federal Highway Administration.

DBMSTOOLS.COM. (n.d.). *30 Data catalogs—DBMS Tools*. Retrieved March 4, 2022, from https://dbmstools.com/categories/data-catalogs

Data Coordinator Guidance. (n.d.). Retrieved from https://datasf.org/resources/data-inventory-guidance/

Florida Department of Transportation (n.d.). Transportation Data Portal. Accessed via http://www.fdot.gov/agencyresources/mapsanddata.shtm

Gharaibeh, N., Oti, I., Schrank, D., & Zmud, J. (2017). *Data management and governance practices* (No. Project 20-05, Topic 47-05).

Green, M., & Anthony, L. (2018). *Data governance & data management case studies of select transportation agencies*. https://www.gis.fhwa.dot.gov/reports/GIS_Data_Governance_and_Data_Management_Case_Studies.pdf

Halevy, A., Korn, F., Noy, N. F., Olston, C., Polyzotis, N., Roy, S., & Whang, S. E. (2016). Goods: Organizing google's datasets. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 795-806).

Harvey, C. (2021, January 27). *Top 10 Data Catalog Software Solutions*. Datamation. https://www.datamation.com/big-data/top-10-data-catalog-software-solutions/

Minnesota Department of Transportation. Data Business Plan. Draft. Retrieved from http://www.dot.state.mn.us/tda/Data Business Plan Final Draft.docx

What Is a Data Catalog and Why Do You Need One? (n.d.). Oracle Corporation. Retrieved March 7, 2022, from https://www.oracle.com/big-data/what-is-a-data-catalog/

Ogle, J. H., Sarasua, W., Putman, B., Davis, J., Huyhan, N., & Ziehl, P. (2019). SCDOT Asset Data Collection Assessment Final Report.

Pecheux, K. K., Pecheux, B. B., Ledbetter, G., & Lambert, C. (2020). Guidebook for Managing Data from Emerging Technologies for Transportation. *NCHRP Research Report*, (952).

Spy Pond Partners, LLC (Arlington, Massachusetts), & Atkins North America, Inc.,(Orlando, Florida). (2021). *Guidebook for Data and Information Systems for Transportation Asset Management*. Transportation Research Board.

Stillerman, J., Fredian, T., Greenwald, M., & Manduchi, G. (2016). Data catalog project—A browsable, searchable, metadata system. *Fusion Engineering and Design*, *112*, 995-998.

Vandervalk, A., Almario, R., Pasumarthy, P., & Snyder, D. (2017). *Maryland State Highway Administration Pilot of the Data Business Plan Guide for State and Local Departments of Transportation: Data Business Plan* (No. FHWA-HOP-18-010). United States. Federal Highway Administration.

Vandervalk, A., Snyder, D., Hajek, J. K., & Systematics, C. (2013). *US DOT roadway transportation data business plan (phase 1): data business plan* (No. FHWA-JPO-13-084). United States. Federal Highway Administration. Office of Operations.

Vandervalk, A., Cronin, B., & Thompson, C. (2020). *Advancing practices for data governance, information management, and managing the impact of digitization on DOT workforces* (No. 20–44(11)). NCHRP. https://agencyadmin.transportation.org/wp-content/uploads/sites/12/2020/11/NCHRP_20-4411_FinalReport_20200408.pdf

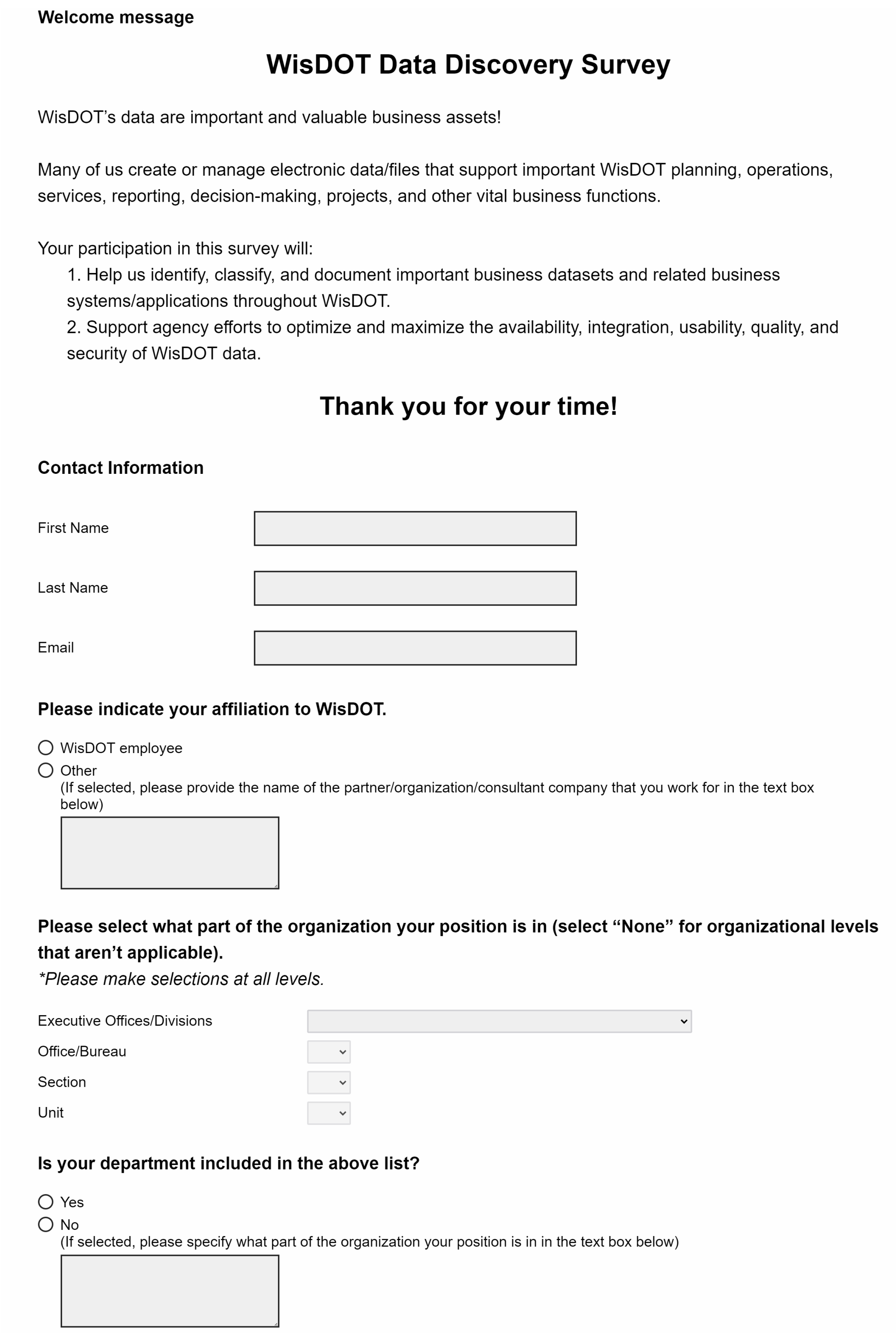
Varshney, S. (n.d.). A Step-by-step Guide to Build A Data Catalog. Retrieved from https://www.datasciencecentral.com/profiles/blogs/a-step-by-step-guide-to-build-a-data-catalog

Wells, D. (2019). Introduction to Data Catalogs.

Wells, D. (2021). A Data Architect's Guide to the Data Catalog. Retrieved from https://www.alation.com/blog/a-data-architects-guide-to-the-data-catalog/

APPENDIX A

Data Discovery Survey



Welcome message

WisDOT Data Discovery Survey

WisDOT's data are important and valuable business assets!

Many of us create or manage electronic data/files that support important WisDOT planning, operations, services, reporting, decision-making, projects, and other vital business functions.

Your participation in this survey will:

1. Help us identify, classify, and document important business datasets and related business systems/applications throughout WisDOT.
2. Support agency efforts to optimize and maximize the availability, integration, usability, quality, and security of WisDOT data.

Thank you for your time!

Contact Information

First Name

Last Name

Email

Please indicate your affiliation to WisDOT.

○ WisDOT employee
○ Other
(If selected, please provide the name of the partner/organization/consultant company that you work for in the text box below)

Please select what part of the organization your position is in (select "None" for organizational levels that aren't applicable).
**Please make selections at all levels.*

Executive Offices/Divisions

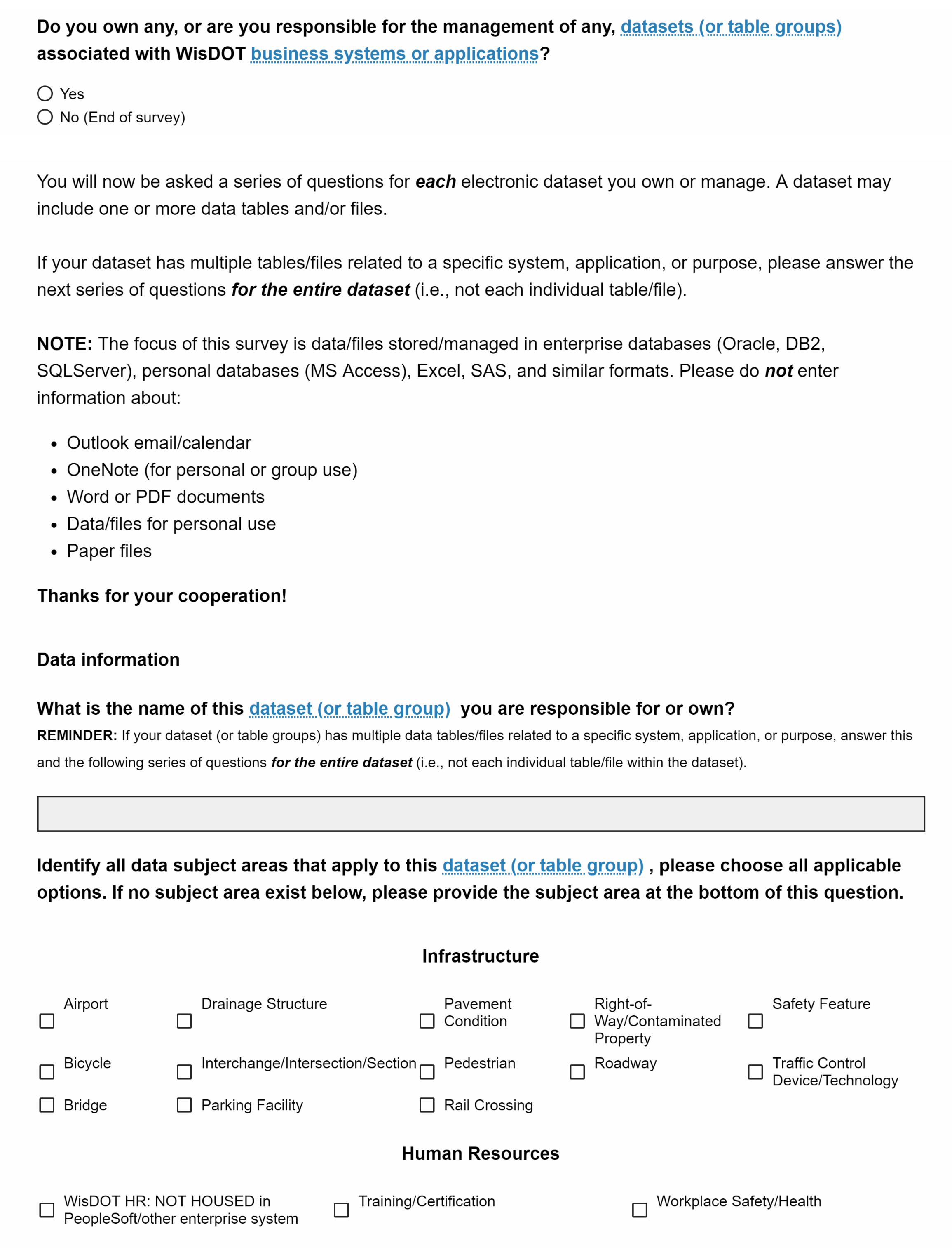
Office/Bureau

Section

Unit

Is your department included in the above list?

○ Yes
○ No
(If selected, please specify what part of the organization your position is in in the text box below)



Do you own any, or are you responsible for the management of any, datasets (or table groups) associated with WisDOT business systems or applications?

○ Yes
○ No (End of survey)

You will now be asked a series of questions for ***each*** electronic dataset you own or manage. A dataset may include one or more data tables and/or files.

If your dataset has multiple tables/files related to a specific system, application, or purpose, please answer the next series of questions ***for the entire dataset*** (i.e., not each individual table/file).

NOTE: The focus of this survey is data/files stored/managed in enterprise databases (Oracle, DB2, SQLServer), personal databases (MS Access), Excel, SAS, and similar formats. Please do ***not*** enter information about:

- Outlook email/calendar
- OneNote (for personal or group use)
- Word or PDF documents
- Data/files for personal use
- Paper files

Thanks for your cooperation!

Data information

What is the name of this dataset (or table group) you are responsible for or own?
REMINDER: If your dataset (or table groups) has multiple data tables/files related to a specific system, application, or purpose, answer this and the following series of questions ***for the entire dataset*** (i.e., not each individual table/file within the dataset).

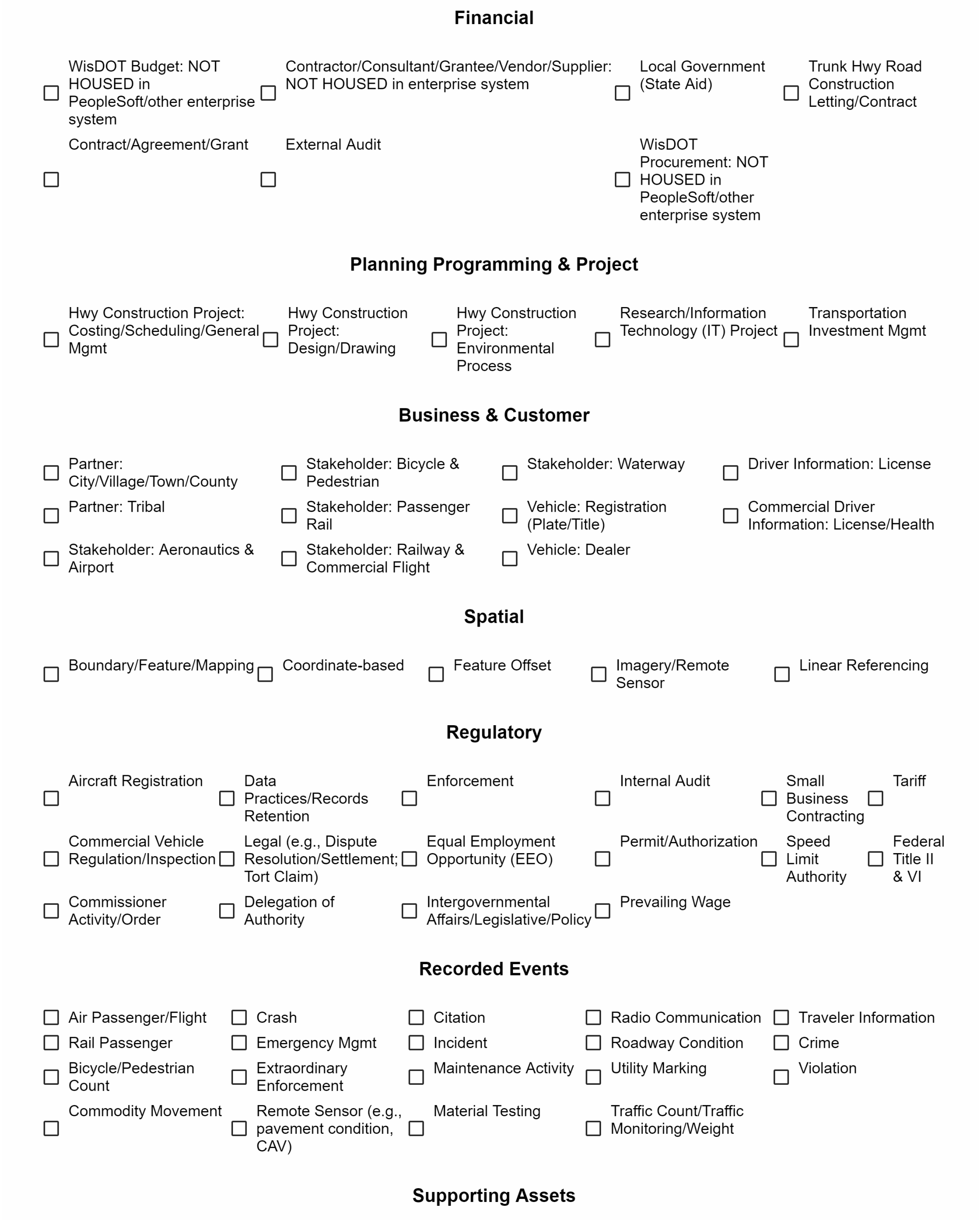
Identify all data subject areas that apply to this dataset (or table group) , please choose all applicable options. If no subject area exist below, please provide the subject area at the bottom of this question.

Infrastructure

□ Airport
□ Bicycle
□ Bridge
□ Drainage Structure
□ Interchange/Intersection/Section
□ Parking Facility
□ Pavement Condition
□ Pedestrian
□ Rail Crossing
□ Right-of-Way/Contaminated Property
□ Roadway
□ Safety Feature
□ Traffic Control Device/Technology

Human Resources

□ WisDOT HR: NOT HOUSED in PeopleSoft/other enterprise system
□ Training/Certification
□ Workplace Safety/Health



Financial

□ WisDOT Budget: NOT HOUSED in PeopleSoft/other enterprise system
□ Contractor/Consultant/Grantee/Vendor/Supplier: NOT HOUSED in enterprise system
□ Local Government (State Aid)
□ Trunk Hwy Road Construction Letting/Contract
□ Contract/Agreement/Grant
□ External Audit
□ WisDOT Procurement: NOT HOUSED in PeopleSoft/other enterprise system

Planning Programming & Project

□ Hwy Construction Project: Costing/Scheduling/General Mgmt
□ Hwy Construction Project: Design/Drawing
□ Hwy Construction Project: Environmental Process
□ Research/Information Technology (IT) Project
□ Transportation Investment Mgmt

Business & Customer

□ Partner: City/Village/Town/County
□ Stakeholder: Bicycle & Pedestrian
□ Stakeholder: Waterway
□ Driver Information: License
□ Partner: Tribal
□ Stakeholder: Passenger Rail
□ Vehicle: Registration (Plate/Title)
□ Commercial Driver Information: License/Health
□ Stakeholder: Aeronautics & Airport
□ Stakeholder: Railway & Commercial Flight
□ Vehicle: Dealer

Spatial

□ Boundary/Feature/Mapping
□ Coordinate-based
□ Feature Offset
□ Imagery/Remote Sensor
□ Linear Referencing

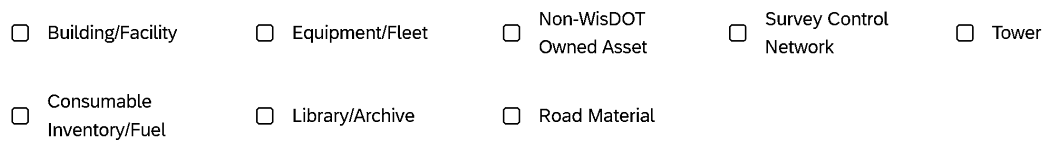
Regulatory

□ Aircraft Registration
□ Data Practices/Records Retention
□ Enforcement
□ Internal Audit
□ Small Business Contracting
□ Tariff
□ Commercial Vehicle Regulation/Inspection
□ Legal (e.g., Dispute Resolution/Settlement; Tort Claim)
□ Equal Employment Opportunity (EEO)
□ Permit/Authorization
□ Speed Limit Authority
□ Federal Title II & VI
□ Commissioner Activity/Order
□ Delegation of Authority
□ Intergovernmental Affairs/Legislative/Policy
□ Prevailing Wage

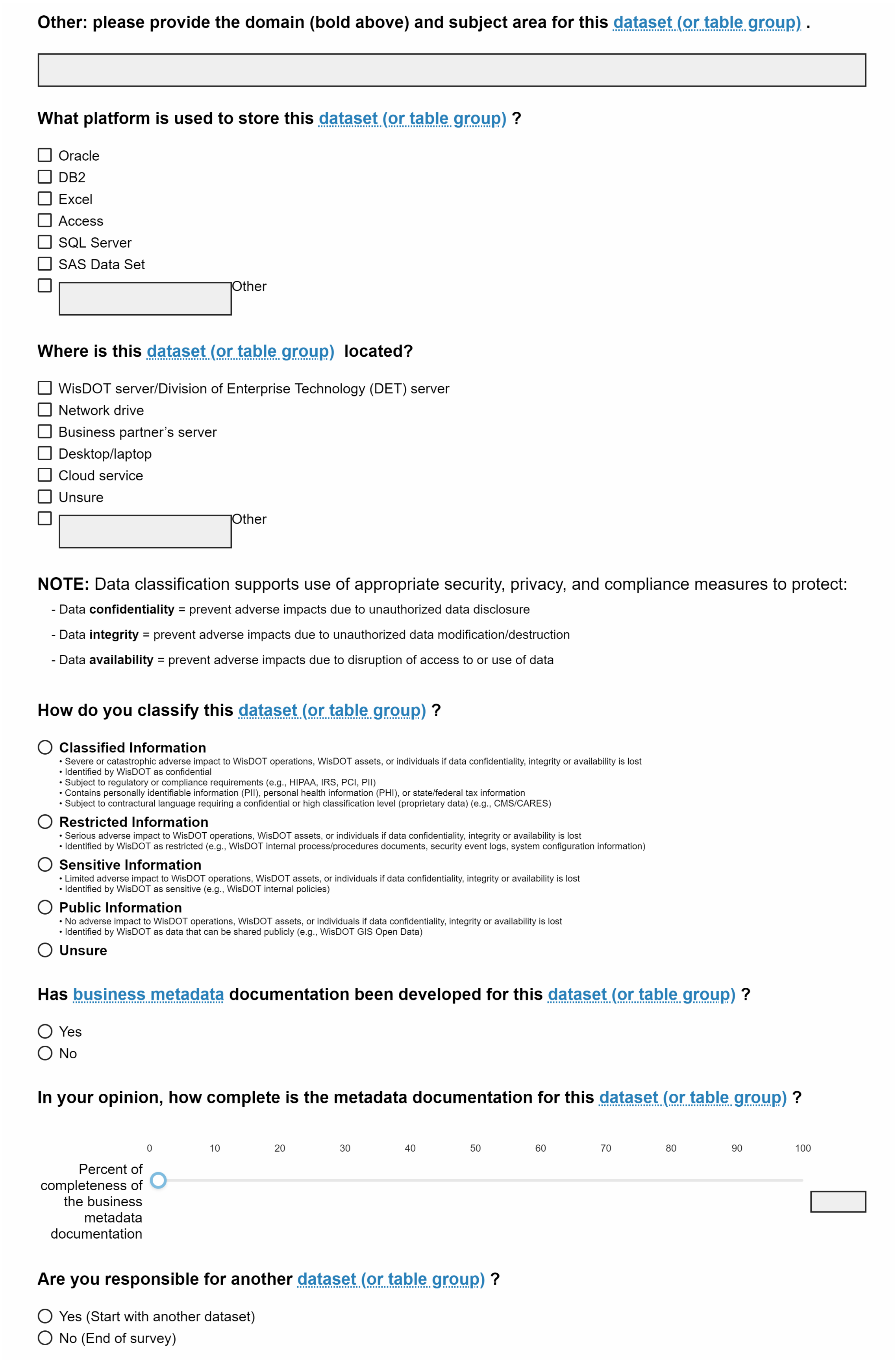
Recorded Events

□ Air Passenger/Flight
□ Crash
□ Citation
□ Radio Communication
□ Traveler Information
□ Rail Passenger
□ Emergency Mgmt
□ Incident
□ Roadway Condition
□ Crime
□ Bicycle/Pedestrian Count
□ Extraordinary Enforcement
□ Maintenance Activity
□ Utility Marking
□ Violation
□ Commodity Movement
□ Remote Sensor (e.g., pavement condition, CAV)
□ Material Testing
□ Traffic Count/Traffic Monitoring/Weight

Supporting Assets



□ Building/Facility
□ Equipment/Fleet
□ Non-WisDOT Owned Asset
□ Survey Control Network
□ Tower
□ Consumable Inventory/Fuel
□ Library/Archive
□ Road Material



Other: please provide the domain (bold above) and subject area for this dataset (or table group) .

What platform is used to store this dataset (or table group) ?

- □ Oracle
- □ DB2
- □ Excel
- □ Access
- □ SQL Server
- □ SAS Data Set
- □ Other

Where is this dataset (or table group) located?

- □ WisDOT server/Division of Enterprise Technology (DET) server
- □ Network drive
- □ Business partner's server
- □ Desktop/laptop
- □ Cloud service
- □ Unsure
- □ Other

NOTE: Data classification supports use of appropriate security, privacy, and compliance measures to protect:

- Data **confidentiality** = prevent adverse impacts due to unauthorized data disclosure
- Data **integrity** = prevent adverse impacts due to unauthorized data modification/destruction
- Data **availability** = prevent adverse impacts due to disruption of access to or use of data

How do you classify this dataset (or table group) ?

○ **Classified Information**
- Severe or catastrophic adverse impact to WisDOT operations, WisDOT assets, or individuals if data confidentiality, integrity or availability is lost
- Identified by WisDOT as confidential
- Subject to regulatory or compliance requirements (e.g., HIPAA, IRS, PCI, PII)
- Contains personally identifiable information (PII), personal health information (PHI), or state/federal tax information
- Subject to contractual language requiring a confidential or high classification level (proprietary data) (e.g., CMS/CARES)

○ **Restricted Information**
- Serious adverse impact to WisDOT operations, WisDOT assets, or individuals if data confidentiality, integrity or availability is lost
- Identified by WisDOT as restricted (e.g., WisDOT internal process/procedures documents, security event logs, system configuration information)

○ **Sensitive Information**
- Limited adverse impact to WisDOT operations, WisDOT assets, or individuals if data confidentiality, integrity or availability is lost
- Identified by WisDOT as sensitive (e.g., WisDOT internal policies)

○ **Public Information**
- No adverse impact to WisDOT operations, WisDOT assets, or individuals if data confidentiality, integrity or availability is lost
- Identified by WisDOT as data that can be shared publicly (e.g., WisDOT GIS Open Data)

○ **Unsure**

Has business metadata documentation been developed for this dataset (or table group) ?

- ○ Yes
- ○ No

In your opinion, how complete is the metadata documentation for this dataset (or table group) ?

Percent of completeness of the business metadata documentation

Are you responsible for another dataset (or table group) ?

- ○ Yes (Start with another dataset)
- ○ No (End of survey)