STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT DOCUMENTATION PAGE

TR0003 (REV 10/98)

ADA Notice

For individuals with sensory disabilities, this document is available in alternate formats. For information call (916) 654-6410 or TDD (916) 654-3880 or write Records and Forms Management, 1120 N Street, MS-89, Sacramento, CA 95814.

1. REPORT NUMBER	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER
CA22-3834		

4. TITLE AND SUBTITLE	5. REPORT DATE
Statewide Data Collection Strategic Plan	June 2022
	6. PERFORMING ORGANIZATION CODE

7. AUTHOR	8. PERFORMING ORGANIZATION REPORT NO.
Jane F Macfarlane, Anthony Patire, Rick Wagner, Viswanath Nandigam	

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. WORK UNIT NUMBER
UC Berkeley 109 McLaughlin Hall Berkeley, CA 94720	11. CONTRACT OR GRANT NUMBER 65A0810

12. SPONSORING AGENCY AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
California Department of Transportation Division of Research, Innovation and System Information, MS-83 1727 30th Street, 3rd Floor Sacramento CA94273-0001	Final Report, Feb 2021 - June 2022
	14. SPONSORING AGENCY CODE Caltrans DRISI

15. SUPPLEMENTARY NOTES

16. ABSTRACT

With the current digital transformation that is occurring across all Department of Transportation (DOT) agencies, data collection activities will play a crucial role in facilitating the organizational processes that support Caltrans' goals. Transportation data needs to be widely understood, organized, communicated, and shared to make possible truly integrated planning, asset management, and operational strategies. And, as importantly, the data assets of Caltrans should be resourced and managed at the same priority as its physical assets. This plan focuses on Cultivating Excellence in the area of data collection in support of these organizational processes. The focus is to define a strategic plan for developing the foundational processes for creating a data-informed culture within the organization and leveraging existing technologies for implementing an enterprise-wide approach to data collection and procurement. The proposed approach for transitioning to a data-informed culture with robust processes for data collection is composed of three core processes: Document domain specific business processes and the associated data use; Build domain ontologies and data dictionaries for the datasets used in the business processes; Establish a Caltrans data registry with links to enterprise data catalogs. The plan defines an organizational process as a feedback loop, recognizing that as business processes change and new data sources emerge, the foundational knowledge base that captures Caltrans' core organizational processes and supporting data resources must also change. The vision is that "Data is collected once and is usable and findable by all" and the mission is to "Enable Caltrans to maximize the value of collected and procured data through proactive stakeholder outreach and feedback."

17. KEY WORDS	18. DISTRIBUTION STATEMENT
transportation data collection, data management, data registry, transportation ontologies	The readers can freely refer to and distribute this report. If there is any questions, please contact one of the authors.

19. SECURITY CLASSIFICATION *(of this report)*	20. NUMBER OF PAGES	21. COST OF REPORT CHARGED
No security issues	122	Free for E-copy

Disclaimer Statement

This document is disseminated in the interest of information exchange. 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 or policies of the State of California or the Federal Highway Administration. This publication does not constitute a standard, specification or regulation. This report does not constitute an endorsement by the Department of any product described herein.

For individuals with sensory disabilities, this document is available in Braille, large print, audiocassette, or compact disk. To obtain a copy of this document in one of these alternate formats, please contact: the Division of Research, Innovation, and System Information, MS-83, California Department of Transportation, Division of Research, Innovation, and System Information, P.O. Box 942873, Sacramento, CA 94273-0001.

Caltrans Strategic Plan for Data Collection

Jane Macfarlane and Anthony Patire
University of California, Berkeley

Rick Wagner and Viswanath Nandigam
University of California San Diego

June 16, 2022

Executive Summary

This report documents a one-year effort to develop a Strategic Plan for Data Collection for Caltrans. The objective of this effort was to create a strategic plan that leverages existing technologies for implementing an enterprise-wide approach to field asset and mobility data collection and procurement that minimizes duplication and maximizes data value and operational efficiency through a collect once, use many times approach.

Caltrans Statewide Data Collection Strategic Plan

Caltrans' overall strategic plan has a mission, a vision, and a set of goals:

- Safety First,
- Cultivate Excellence,
- Enhance and Connect the Multi-modal Transportation Network,
- Lead Climate Action, and
- Advance Equity and Livability

The aim of this effort is to create a strategic plan that aligns with the above goals. With the current digital transformation that is occurring across all Department of Transportation (DOT) agencies, data collection activities will play a crucial role in facilitating the organizational processes that support Caltrans' goals. Transportation data needs to be widely understood, organized, communicated, and shared to make possible truly integrated planning, asset management, and operational strategies. And, as importantly, the data assets of Caltrans should be resourced and managed at the same priority as its physical assets. This plan focuses on **Cultivating Excellence** in the area of data collection in support of these organizational processes. The intent is to define a strategic plan for developing the foundational processes for creating a data-informed culture within the organization and leveraging existing technologies for implementing an enterprise-wide approach to data collection and procurement.

What Is The Need?

Caltrans lags behind other State DOTs by not implementing a unified and coordinated statewide approach to data collection and management. This plan will help to minimize duplication and maximize data value and operational efficiency through a collect once, use many times approach.

Stakeholders

The leaders of the Enterprise Data Steward Committee are:

- Chad Baker,
- Walter Yu, and
- April Nitsos.

The stakeholders for this effort comprise the participants listed in the following table:

Name	District/Division: Program	Classification
Samer Batarseh	HQ: Transportation Planning, Planning and Modal	TE Range C
Joshua Davis	District 4: Project Delivery East	Landscape Assoc, Range D
Harold Feinberg	HQ: DRISI Planning and Modal	Research Data Supervisor II
Dario Moreno	Dist 5: Planning and Modal	Research Data Specialist II
Kitae Nam	HQ: Local Assistance Planning and Modal	Senior TE
Kevin Riley	HQ: Traffic Operations	Sup TE

Objectives

There were four specific objectives for the proposed work effort:

- Assess the current state of Caltrans' data collection so that a model of current activities associated with data collection can be described.
- Develop a future model, in coordination with Caltrans stakeholders, that describes the priorities for future data collection and the expected policies and cultural guidelines for data collection.
- Provide the core content for the Strategic Plan which is the mechanism for transitioning from the current state to a future state in which data collection is leveraged for maximum use, reuse, and collaboration among stakeholders in the organization.
- Aid in the development of communications focused on the cultural guidelines and expected outcomes of data collection activities across organizational boundaries.

The approach taken consisted of:

- conducting interviews with key personnel to understand multiple purposes of data collection activities,
- identifying and documenting existing capabilities, and
- identifying deficiencies in current data collection activities, e.g. freshness of data, quality of data.

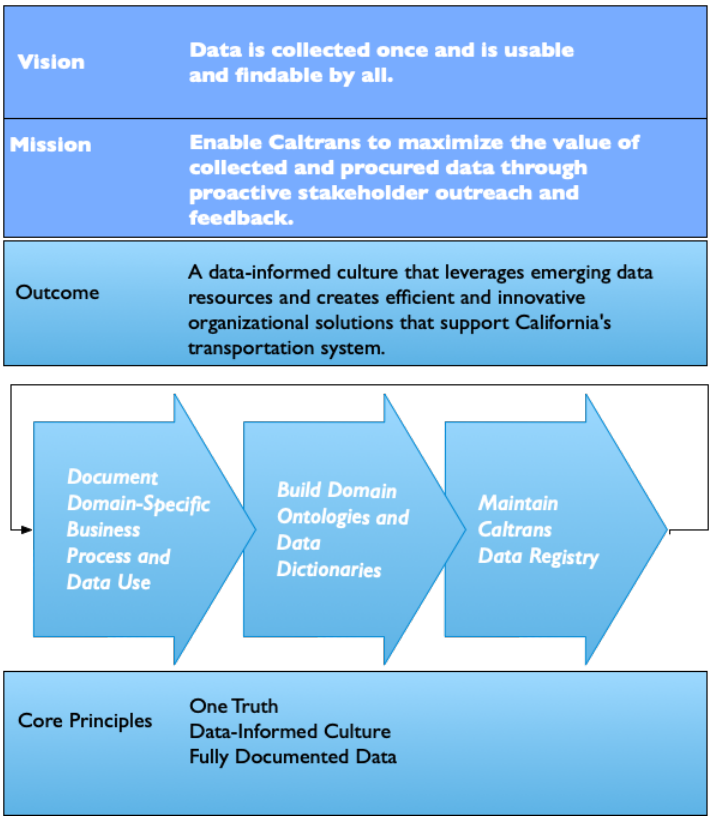
CORE ISSUES AND PROPOSED ACTIONS

Issue: Siloed data procurements.	**Action:** Increase awareness of organizational data resources.
Issue: Lack of ontologies for describing the core data needs that support business processes.	**Action:** Establish a cross-organizational governance process for establishing data resource ontologies and associated metadata mappings.
Issue: Siloed analytics.	**Action:** Establish a cross-organizational process for identifying data integration opportunities and analytics reuse.

The proposed approach for transitioning to a data-informed culture with robust processes for data collection is composed of three core processes:

- Document domain specific business processes and the associated data use.
- Build domain ontologies and data dictionaries for the datasets used in the business processes.
- Establish a Caltrans data registry with links to enterprise data catalogs.

The figure below shows the full process as a feedback loop, recognizing that as business processes change and new data sources emerge, the foundational knowledge base that captures Caltrans' core organizational processes and supporting data resources must also change. The vision is that "Data is collected once and is usable and findable by all" and the mission is to "Enable Caltrans to maximize the value of collected and procured data through proactive stakeholder outreach and feedback." Core principles that should drive the data collection processes are: build an infrastructure that encourages a "One Truth" approach that discourages siloed data procurements and siloed analytics that create multiple representations of the same datasets; promote a data-informed culture that makes data-driven decisions using easily accessible data; and enable full provenance management resulting in fully documented data. The end goal is to build a data-informed culture that leverages emerging data resources and creates efficient and innovative organizational solutions that support California's transportation system.



Key Recommendations

The final section of the plan provides a list of recommendations for moving forward. Three key recommendations regarding important fundamental infrastructure requirements and an associated reporting/process change for data procurement that is necessary for a robust data collection strategy are:

RECOMMENDATION: Caltrans should establish a centralized or federated registry listing key data repositories, data catalogs, and other data resources with reuse potential outside of their originating business unit.

RECOMMENDATION: Caltrans should establish central or federated data repositories tied to either business areas (i.e., physical structure management, transportation system management) and/or core business units (i.e., Districts). Data within these repositories should be discoverable through data catalogs using standards-based metadata. New repository implementations should plan to handle a mix of structured and unstructured data to accommodate future data needs.

RECOMMENDATION: Caltrans should require that all corporate data be curated in a data repository within a reasonable amount of time. In addition, the metadata for existing data used in business processes should be documented and integrated into topical ontologies and added to the data catalogs. The process for beginning this work is described in this report, which is composed of two initiatives: the first is to build teams that will be responsible for documenting the business processes and defining the ontologies, and second is to initiate a workforce development effort that provides educational courses focused on data management. A Caltrans-wide communications effort to raise awareness of the importance of building excellence in data management should be a key component of the workforce development activities.

Contents

List of Figures

List of Tables

1 Background

1.1 Introduction

The rapid pace of digitization across all modes of transportation is creating a massive amount of data. This is creating an opportunity for Departments Of Transportation (DOT) agencies to significantly improve productivity and effectiveness of organizational processes. Understanding how this emerging data can best inform the business processes of the organization is key to leveraging the wealth of data that is being made available.

What is the need? Caltrans lags behind other state DOTs by not implementing a unified and coordinated statewide approach to data collection and management. This plan will help to minimize duplication and maximize data value and operational efficiency through a collect once, use many times approach.

In order to optimize the value of data within Caltrans, it must develop a coordinated and repeatable process to evaluate and prioritize what, where, when, why and how data should be acquired; establish data storage, discovery and access solutions; and define the roles and responsibilities of stakeholders that will provide the core base of expertise for establishing the value proposition of the data in context of real-world business needs.

1.2 Stakeholders

The leaders of the Enterprise Data Steward Committee are:

- Chad Baker,
- Walter Yu, and
- April Nitsos

The stakeholders for this effort comprise the participants listed in Table 1.

Table 1: Project Panel Members

Name	District/Division:Program	Classification
Samer Batarseh	HQ: Transportation Planning, Planning and Modal	TE Range C
Joshua Davis	District 4: Project Delivery East	Landscape Assoc, Range D
Harold Feinberg	HQ: DRISI Planning and Modal	Research Data Supervisor II
Dario Moreno	Dist 5: Planning and Modal	Research Data Specialist II
Kitae Nam	HQ: Local Assistance Planning and Modal	Senior TE
Kevin Riley	HQ: Traffic Operations	Sup TE

1.3 Scope

This specific effort was focused on the creation of a strategic plan for facilitating data collection and data use within the agency. Data collection is an initial step in a comprehensive data management strategy that facilitates the digital transformation of the organization. This first phase effort is to determine the current state of data collection and define a process for moving to a cohesive, coordinated approach to data collection.

Impact of the COVID-19 Pandemic. The COVID-19 pandemic has significantly reduced the desirability to conduct in person workshops. To adjust for this impact, a Qualtrics survey was created to replace the face-to-face workshops. While a different approach, the overall outcome was likely improved by achieving a greater reach across the organization and obtaining more detailed information than would have likely been gathered in a workshop environment.

1.4 Definitions

Dataset. Generally speaking, a dataset is a logical collection of one or more related files. For example, a project can have accounting spreadsheets, engineering drawings, field surveys, and contracts. These can be assembled at the end of the project into a dataset for archival. Datasets can also be stored as tables within a database or only contain a single file, like a video.

Metadata. Metadata is the set of attributes used to describe a resource, such as a file or dataset. Common elements include the resource creator, the date of creation, and a title. Several standards exist for defining metadata attributes, such as Schema.org, ISO 19115 for Geographic Information, and the Federal Geographic Data Committee (FGDC) metadata standard. The use of metadata standards enables rapid discovery of data through common search terms, enables interoperability, and promotes data reuse.

The choice of metadata defines the types of data discovery or searches that can be done. Organizations frequently combine standards and add additional terms. For example, attributes like District, Metropolitan Planning Organization (MPO), Regional Transportation Planning Agency (RTPA), and county could be useful terms to search for in project datasets.

Data Catalog. A data catalog provides a searchable list of data assets. It uses metadata to enable an organization's users to quickly search and identify data of interest. The data can be colocated with the catalog or within a distributed data catalog.

Data catalogs are implemented using metadata as described above. Organizations can have several data catalogs, each describing specific classes of data. In this case, common metadata terms are critical for interoperability.

A data catalog may or may not provide access to the data it lists, although for each dataset it should provide information on how to access the data. For example, a data catalog could provide a listing of financial datasets that require approval for access. In this case, the metadata describing the dataset should include a reference on how to request access.

Data catalogs must be dynamic and be able to adapt and grow with an organization. It is important to have a data strategy. Without a strategy, as data becomes more available and increasingly complex, the result is a fragmented collection of data sources that are siloed, not well documented, and are difficult to find. This can lead to duplicated efforts to collect data, distrust in data that has been collected, and misuse of the data.

Data Repository. A data repository is a curated and managed system for storing organizational data. This distinguishes it from data storage used for active or working data, like a departmental file server. Data repositories have a defined level of availability, along with policies for data access and curation. The data stored in the repository should be listed in one or more data catalogs, and the repository itself should be listed in a registry.

Data repositories can be tightly coupled to a particular data catalog, enabling users to both discover and access data seamlessly. Similarly, data repositories can provide integrated analysis or visualization capabilities, potentially supporting decision-making.

Registry. A registry is a comprehensive inventory of high-level data catalogs, data repositories, databases, and other resources within, or accessible by, the organization. Registries can provide a coarse-grained view of the data within a given resource or actually query the resource to look for individual datasets. Registries can provide a single point for data discovery, without requiring a monolithic data catalog describing granular resources in detail.

Curation. Curation of data is key to the process of discovery. The Inter-university Consortium for Political and Social Research (ICPSR), an international leader in data stewardship, describes data curation as: "The process of 'caring' for data, including organizing, describing, cleaning, enhancing, and preserving data for public use." In the context of this report, we are considering curation to be

the process of ensuring that the data collected is quality controlled, validated, and approved for use by others.

Ontology. An ontology describes the core information concepts that drive the business processes and the relationships among the concepts. An ontology is used to ensure consistent meaning and naming across disparate disciplines, organizations, and IT systems. It can include several taxonomies.

Taxonomy. A taxonomy is a knowledge classification system that organizes the concepts into hierarchical categories. A taxonomy is a description of how the concepts are related.

1.5 Caltrans-Specific Corporate Definitions

Corporate data must follow enterprise data governance practices. Corporate data has four types.

Master Data Data that identifies and describes the people, places, and things that are fundamental to agency activities and that appear across multiple agency information systems.

Reference Data Code tables with permissible values for other data fields (e.g., county codes and names).

Control Agency Required Data Data required by control agencies that impact agency funding or resource allocation decisions (e.g., SB1 reports or Federal Highway Performance Monitoring System Data).

Shared Data Caltrans-sourced data used across multiple Caltrans business units or data of value to agency partners and the traveling public (e.g., traffic counts).

Purchased/Licensed Data Datasets that were produced as a result of state grants, contracts, and cooperative agreements (e.g., Streetlight travel data)

1.6 Organization of the Report

The remainder of the report is organized as follows: Section 2 provides a discussion of the Caltrans data collection mission, vision and goals. In addition, it outlines target characteristics for an organization with mature data collection and management processes. Section 3 provides a description of the work effort associated with defining the current state, including a description of the survey implemented to gather information about current data, for example, its findability and usability. Also included is an initial assessment of data health for existing data assets. Section 4 provides a prioritized plan for moving from the current state towards the desired future state. Finally, Section 5 describes a range of technologies and tools that could be helpful to Caltrans at its current stage of data maturation. The Appendices A-C include currently published guidance documents for creating metadata and data dictionaries and current roles and responsibilities for the Enterprise Data and Geospatial Governance. Appendix D and E details the questions that were used for an organizational survey focused on evaluating the current state of the data collection activities and the results of the initial survey.

2 Caltrans Data Collection Mission, Vision and Goals

2.1 Mission

Enable Caltrans to maximize the value of collected and procured data through proactive stakeholder outreach and feedback.

2.2 Vision

Data is collected once and is usable and findable by all.

2.3 Goals

It is not possible to overstate the importance of cultivating a data-informed culture that proactively maintains data infrastructure such that current, reliable, and actionable information is findable, accessible, usable and reusable across the organization, and shareable with partners and stakeholders. Success in building a data-informed culture requires a foundational data collection strategy that introduces the required discipline to manage and democratize the data.

Goals for a data collection strategy should:

1. Promote multiple stakeholder planning, collection, and use of data to ensure that the collection provides maximum value to Caltrans
2. Encourage collaboration
3. Improve data governance compliance
4. Provide tools that reduce the burden of curation
5. Facilitate the ability to quickly search the data catalog and discover existing data sources
6. Share analytical methods and tools for extracting value from data sources

2.4 Description of Future State

This section describes characteristics of a mature organization with respect to data collection and management. Best practices are summarized according to the categories defined in *NCHRP Report 814: Data to Support Transportation Agency Business Needs: A Self-Assessment Guide.* [BNAoSM15]

Although this discussion is presented within a larger context, the scope is firmly focused on those aspects of data collection and reuse that are of greatest relevance to Caltrans given its current level of data maturity. Emphasis and detail are only provided for the areas that most directly support the vision, such as creating a collaborative culture, creating ontologies, establishing and maintaining mechanisms to track data provenance as the data is used and reused in the organization, and cultivating the ability to manage a progressively greater variety of data.

2.4.1 Data Strategy and Governance

Data governance is created by defining activities that implement and enforce control over the data management process and data assets. With an organization as large as Caltrans, transparency and accountability associated with data procurement and data generation becomes very important. Organizational roles and responsibilities, namely transparency in data ownership and encouragement of data reuse across organizational lines, will help build the organizational culture that supports the vision to procure once and use many times. Additionally, identifying the data dimensions are important in order to ensure that the process for data management and technology selection is well advised.

An organization with a well-defined and mature level of data strategy and governance should address the following considerations:

Strategy and Direction. A strategy for data management exists, collaboration is facilitated across the agency, and a process exists to make continual improvements.

Roles and Accountability. Data stewards who have the necessary skills and authority to carry forward the data management strategy are designated. Clear roles are defined for data stewards and for those accountable for data quality, accessibility, and usability. Data stewards are empowered to collaborate both internally and externally from the agency.

Policies and Procedures. Data collection policies exist and processes are in place for review, and a procedure for ingestion into a data catalog is followed.

Data Asset Inventory and Value. The data catalog is continually updated. Performance metrics are generated about the data to ascertain their primary users, how often it is being used, and for what purpose.

The challenge is to develop a process that is not onerous for the data contributors. An overly controlled process will inhibit acceptance of the policies. Tools that promote policy compliance at the point of creation, such as metadata wizards, can reduce the friction and develop good data management habits.

For the purpose of this strategic plan, emphasis is placed on the need to establish a data catalog to track data asset inventory. The data catalog should provide performance measures and a means to answer some basic questions, such as:

- When was the data captured?
- When was the data updated?
- What was the source of the data?
- Did other processes generate the data?
- What transformations have been applied to the data?
- How is the data validated?
- Is the data fit for the purpose it has been purchased, collected or created for?

The process for data management and the technology used for supporting and enforcing data governance will help reduce the organizational load of management.

Relationships with Data Customers. A feedback loop exists between the data producers and consumers to identify and prioritize possible improvements.

Data Management Sustainability. Institutional knowledge about data management processes are standardized and passed on as staffing evolves. Specialized knowledge about critical datasets is well documented. To pass on data knowledge, tacit knowledge of organizational groups must be turned into explicit knowledge that crosses organizational boundaries. This will drive usage, build trust across groups, and encourage expanded views of how to use the data source. An important activity for codifying tacit knowledge is to encode explicit knowledge in metadata.

2.4.2 Data Life Cycle Management

A key element to consider during the life cycle of data is to strive always to keep 'one truth'. Often in large organizations, data is transferred from person to person via email or common servers and downloaded to individual machines. Without metadata attached to this data that tracks versioning, over time there can be many copies of the same data. Without metadata tracking that updates the data provenance, there is no mechanism to determine if this proliferation of the data has resulted in different datasets or many copies of the same dataset.

An organization with a well-defined and mature level of data life cycle management should address the following considerations:

Data Updating. Rules are defined to coordinate data additions, updates, and deletions.

Data Access Control. Data protection, security, and access control are built into the systems that store and share the data.

Data Findability and Documentation. Data catalogs are searchable, and processes are in place to keep the catalogs current. Data catalogs are accessible and well-known throughout the agency. Data beyond workgroups is centralized. Useful tools include data catalogs, SharePoint, OneDrive, and GitHub, to name a few.

Data Backups and Archiving. Although out of scope for this strategic plan, a mature organization will have procedures for backups and archiving as part of a business recovery plan.

Data Change Management. Processes are in place to anticipate data changes. Self-consistency is maintained when changes are propagated to downstream users.

Data Delivery. Data is available in formats that facilitate its efficient usage without the need for one-off, data-wrangling fire-drills.

The tracking of data provenance is a crucial part of the data life cycle. Data is necessarily processed, cleaned, aggregated and used for many different purposes. Downstream users need to know whether data is appropriate for their intended application - data is fit for purpose. Having access to data provenance information enables the user to make this determination with confidence, and encourages data reuse.

2.4.3 Data Architecture and Integration

Standards, processes, tools, and mechanisms are necessary to facilitate data inter operability and integration within the organization and with external partners.

Location Referencing. Location referencing methods are standardized. Automated tools exist to conflate data between different network definitions and location specification standards. Processes are in place to find and fix errors in network and location mapping.

Geospatial Data Management. Standardized methods and tools are available to integrate GIS data with other types of data.

Data Consistency and Integration. Standard operational procedures exist for consistent coding of common link fields across different datasets to enable cross-referencing. Externally procured datasets can be linked to internal datasets.

Temporal Data Management. Standard operational procedures exist for consistent coding of date and time data elements to facilitate analysis of "snapshots" and trends.

2.4.4 Data Collaboration and Sharing

A collaborative data sharing environment that supports cross-organizational business process goals is crucial. An effective data cataloging environment can provide insights into who uses collected data and for what purpose.

Internal Agency Collaboration. Data partnerships are encouraged within the agency to improve sharing, efficiencies, and best practices. In addition, performance metrics are gathered to understand how data flows within the organization. Opportunities are found to improve the value generated by the data.

External Agency Collaboration. Data partnerships are encouraged with external partners to improve sharing, efficiencies, and best practices.

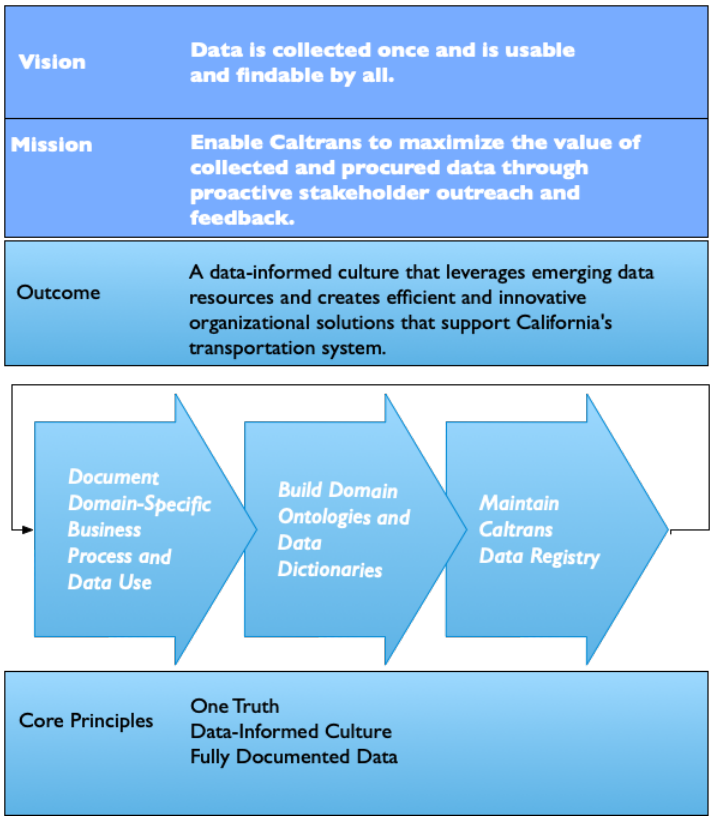


Figure 1: Strategy for Data Collection

2.4.5 Data Quality

It is of crucial importance to guarantee that data is usable for a purpose. This is not a serial process, but rather a practice that needs to permeate the entire organization. Ultimately, a cultural change that encourages a focus on data quality is required for digital transformation.

Data Quality Measurement and Reporting. Automated processes are in place to track performance measures about data, including its timeliness, completeness, validity, and accuracy.

Data Quality Assurance and Improvement. Automated processes are in place to find and flag potential errors, mis-configurations, and inconsistencies. Data stewards are held responsible to maintain data quality standards.

2.4.6 Future State Overview

The following sections of the report will (1) address our understanding of the current state that was gained via a survey of key data users in Caltrans, and (2) describe a recommended path to building an infrastructure to support the mission and vision of this plan.

Of the elements of a full Data Management strategy described previously, Data Collection touches on data strategy and governance, data life cycle management, data access control, data updating, data change management, and data architecture and integration. Establishing a robust system for Data Collection will provide the core foundation for future efforts on leveraging the data, using more sophisticated analytics, and increasing collaboration..

Figure 1 shows the full process as a feedback loop, recognizing that as business processes change and new data sources emerge, the foundational knowledge base that captures Caltrans' core organizational processes and supporting data resources must also change. The vision is that "Data is collected once and is usable and findable by all", and the mission is to "Enable Caltrans to maximize the value of

collected and procured data through proactive stakeholder outreach and feedback." Core principles that should drive the data collection processes are: build an infrastructure that encourages a "One Truth" approach that discourages siloed data procurements and siloed analytics that create multiple representations of the same datasets; promote a data-informed culture that makes data-driven decisions using easily accessible data; and enable full provenance management resulting in fully documented data. The end goal is to build a data-informed culture that leverages emerging data resources and creates efficient and innovative organizational solutions that support California's transportation system.

Roles and responsibilities for those engaged in these processes are described further, and actionable recommendations are provided.

3 Current State Assessment

This section describes the current state of data usage and management within Caltrans and the challenges that may impede the implementation of a "collect once, use many times" approach to data collection at a state-wide scale.

A number of interviews were held with key stakeholders to establish an overall picture of the data challenges and current initiatives underway. In addition, a selection of Caltrans employees were asked to participate in a survey intended to:

- prioritize the highest value data; and,
- identify barriers that hinder the findability, accessibility, and usability of existing datasets

The survey revealed key gaps in Caltrans' processes and organization to successfully implement an efficient and sustainable practice of data curation and deliver the promise of "collect once, use many times."

3.1 Stakeholder Interviews

Multiple interviews were conducted to understand issues and challenges related to data collection and sharing by internal stakeholders of Caltrans. Insights were gathered from these interviews, which were valuable in understanding some of the challenges related to data collection and sharing.

Data sharing within Caltrans is currently not a high priority. Multiple departments within Caltrans collect data for their internal needs but have a difficult time sharing the data with other internal departments. Requests for data is not a priority and can take several days or weeks to fulfill. There is a lack of common data collection standards and semantics (e.g. there are no file naming conventions across departments within Caltrans), which makes interpreting the data once shared even more challenging and can lead to incorrect hypotheses. This lack of standards is of particular concern when data, generated by two different groups, is found to be inconsistent (e.g., weight units).

There is a critical need for ontology development and consequent use of detailed metadata and data catalogs for efficient discovery, access and reuse of data.

Spatial data has its own set of challenges when it comes to collection, quality assurance and sharing. Spatial formats, like shapefiles, are not as efficient to share, such as via Restful APIs.

Data quality is also an issue. When spatial data is collected from multiple sources including spreadsheets and databases, where manual entry is involved, data conflation is difficult.

Data literacy and awareness are lacking, causing a duplication of effort. There is consensus on the importance of LIDAR data and its applications across multiple use cases within Caltrans. However, after data is collected, coordination and dissemination are lacking, resulting in duplication of effort. Headquarters has a statewide contract where sign information is being extracted from LIDAR data. However, some districts are collecting information manually using mosaic imagery or contracting with traffic police.

Data and technology literacy and awareness can help cut down duplication of efforts and costs.

3.2 Survey Results

3.2.1 Survey Framework

The framework of the survey focused on the current state of data use and discovery in Caltrans. In addition, it aimed to find indicators of what kinds of metadata should be considered as a basis for the future state of data use and discovery. Figure 3.2.1 shows the structure of the survey. The core questions focused on:

1. respondents role in the organization,
2. the kind of data that the respondent used regularly in their work,
3. the data exchanges that the respondent managed, and

4. licensing and access right for the data.

Detailed questions focused on four categories of metadata types: structural that describes where the data is and how it is accessed, administrative/technical information related to data formats, preservation of data that provides quality information and usage information, and descriptive information related to information content of data. Example questions related to these metadata categories are provided below.

Structural. What are the highest value datasets that your organization generates? How do you find information that you need? E.g., ask a Data Steward, consult a data catalog, google external providers, other? Do you have a mechanism to relate multiple datasets across stakeholders? Whom do you obtain content from? (Internal stakeholders and external providers) Whom do you provide data to? (Organizations that you interact with) Is it easily findable, current and usable?

Administrative/Technical. In what format do you obtain data from others in the organization (spreadsheet/csv, document, sql query from database, shapefiles, other)? In what format do you provide data to others in the organization? Do you document the provenance of the data? If so, what tools do you use to communicate the provenance?

Preservation. Do you have a mechanism to confirm that you have the correct version of the data? Is there a mechanism to know that there is only one authoritative copy of the data? Is there a mechanism to know the currency of the data? Rights: Do you have to manage licensing and access rights to get the data you need? What kind of data requires this? How do you manage access to the data (security)?

Descriptive. What content do you need now? What content do you think you will need in the future?

3.2.2 Survey Participants

Survey participants were selected by leaders of the Caltrans Enterprise Statewide Field Data Collection Committee. The selected individuals were asked to complete the survey online. The users were chosen based on their current use of data in the business processes that they manage. Because they were hand-selected by Caltrans management familiar with their roles, the team felt that this group represented a good respondent base for this initial assessment. A total of 49 responses were collected.

Many of the users in the respondent group use a large number of datasets associated with the topical areas in the survey. As such, we are considering many of them *power users: people that use this data daily in their work processes.* In later sections of this report, we provide recommendations for building out the metadata foundation for existing datasets and suggest recruiting these respondents for this purpose.

The participants included 31 respondents from Headquarters and 18 from the Districts. There was at least one representative from districts 1, 2, 3, 4, 6, 8, 9, 10, and 11. There were no respondents from districts 5, 7 and 12. In addition, the survey included:

- 17 individuals from HQ Project Delivery
- 5 individuals from HQ Planning and Modal
- 4 individuals from HQ Maintenance and Operations
- 4 individuals from Finance
- 1 individual from Administration

Given the diversity and size of Caltrans, this is a small sample set and, as such, creates a definite bias inherent in the results.

RECOMMENDATION: Assessment of the survey respondents should be done to determine if an expanded survey group would provide improved insights into the current state. If not, launch an expanded survey in order to extend reach into the organization. The survey is readily accessible for a continued expansion and outreach within Caltrans.

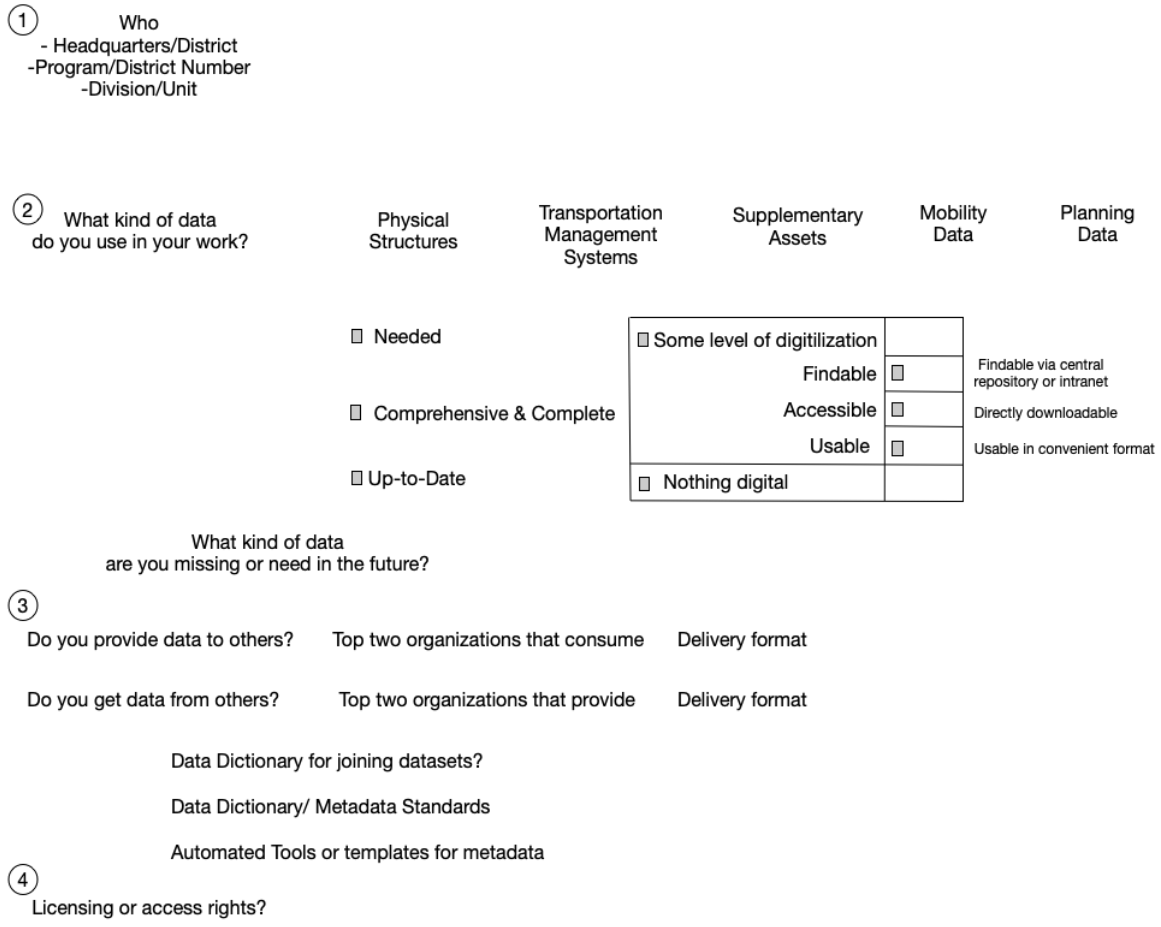


Figure 2: Survey Structure (1) Home Organization (2) Data Types Used in Respondents Work Processes (3) Organizational Exchange of Data (4) Access Limitations

3.2.3 Survey Assessment

Data health is the condition of the data and its ability to support effective, timely decisions and business objectives. Does the organization have the ability to prove that the data is fit for purpose: has it been validated, is it complete, and of sufficient quality to be applied in the business process. The respondents assessed that overall organizational data health is low.

Physical Structures. Overall state of data health is low. Bridges and culvert information appeared to be the best managed datasets. A total of 23 respondents indicated that they are users of asset data associated with physical highway structures. Physical structures include pavement, pavement markings, barriers, sound walls, bridges, culverts, guardrails, crash cushions, static signs, changeable message signs and highway lighting. Within this group, overall awareness of the datasets appeared low.

- **Currency:** Many respondents did not know if the data was up to date. Static signs were most often described as out of date. Bridge and culvert data was described by several as current.
- **Necessity:** Most agreed that physical structure asset data is needed.
- **Completeness:** Pavement, bridge, and culvert data were considered the most complete and comprehensive. Static signs were considered the least complete.
- **Discovery, Access, and Use:** Respondents had very different opinions regarding data discovery, access, and use. Bridge data was commonly regarded as findable and accessible by download, which might indicate that its data health is high, yet Section 4.1.4 identifies that the data is clearly lacking well-defined metadata. Existence of asset data about guardrails, pavement markings, barriers, and crash cushions was commonly unknown or in dispute. Datasets on pavement, bridges, culverts (and perhaps changeable message signs) appeared to be the best managed.

Transportation Management Systems (TMS). Overall state of data heath is low. A total of 16 respondents indicated that they are users of asset data associated with transportation management systems. Transportation management systems include closed circuit cameras, traffic census stations, roadway weather information systems, traffic monitoring detection systems, highway advisory radios, freeway ramp meters, traffic signals, and control cabinets. Within this group, overall awareness of the datasets appeared low.

- **Currency:** Many respondents did not know if the data was up to date.
- **Necessity:** Most agreed that TMS asset data is needed.
- **Completeness:** Many were comfortable with the completeness and comprehensiveness of the data.
- **Discovery, Access, and Use:** Very few respondents reported that the data was accessible and usable by themselves. Respondents rely on their data stewards to help them find the data that they need.

Supplementary Assets. Overall state of data health is low. A total of 19 respondents indicated that they are users of data about supplementary assets. Supplementary assets include vegetation, irrigation and drainage, ADA infrastructure, roadside rest facilities, park and ride infrastructure, truck weigh stations, sidewalks and bicycle and pedestrian facilities.

- **Currency:** Many respondents did not know if the data was up to date.
- **Necessity:** Most agreed that supplementary asset data is needed.
- **Completeness:** Roadside rest facility data is considered the most complete in this category. Data on sidewalks, bicycle and pedestrian facilities is considered the most incomplete.
- **Discovery, Access, and Use:** Respondents had very different opinions about the data. Some found using the data easy while others rely on data stewards and co-workers. Existence and accessibility of vegetation and ADA infrastructure information was unclear.

Mobility Data. Overall state of data health is low. A total of 12 respondents indicated that they are users of mobility data. Mobility data include flow and volume data, speed data, vehicle hours travelled (VHT) and vehicle miles travelled (VMT), trip and mode choice, origin-destination (OD), intersection turning counts, GPS traces, and freight and goods movement.

- **Currency:** Very few respondents thought that mobility data was up to date.
- **Necessity:** Most agreed that mobility data is needed. GPS traces were ranked the lowest.
- **Completeness:** Flow, speed, VHT and VMT data were considered the most complete in this category. Few considered freight, OD and GPS traces to be complete
- **Discovery, Access, and Use:** Freight data was considered the most findable, accessible and usable, with flow, volume, VHT and VMT data not quite there but mostly digitally accessible. Respondents had very different opinions about the other data types or did not know. Awareness of the data is low.

Planning Data. Overall state of data health is low. A total of 14 respondents indicated that they are users of planning data. Planning data includes census tracts, land use and parcels, safety data and injuries, transit data and, other GIS data.

- **Currency:** Census tracts, land use and parcel data was considered the most up to date.
- **Necessity:** Most agreed that planning data is needed.
- **Completeness:** Census tracts, land use and parcel data was largely considered to be complete. Respondents had split opinions about the completeness of transit data.
- **Discovery, Access, and Use:** Census tracts, land use and parcel data was considered the most findable and accessible. Respondents had very different opinions about transit data.

Additional Data Needed. Respondents were asked what additional or improved data they need for managing their business processes. Responses largely fell into two main categories:

- Administrative data for improved management of organization, contracts, and vendors; and,
- Data about the physical world, including emerging IoT sensing capabilities.

Administrative data for management of the organization included the following categories:

- Data for tracking of employees, training, performance, pay, and benefits
- Project management data, including post mile locations, status, cost, and milestones in an accurate and consistent format
- Construction contract information, including contractors, bid items, payments, specifications, change orders, wages and labor compliance
- Cost of equipment, and change orders.
- Financial data, including that from CTIPS, PRSM, and AMS
- Archival records

Physical world data fell into roughly four broad categories. Some of the following will require new data management mechanisms that can handle geospatial and temporal data.

- Geospatial mapping
  - Utility data, above and below ground
  - Land survey data
  - Aerial imagery and LIDAR
  - Standard roadway asset data: road classification, characteristics, maintenance history
  - Infrastructure status monitoring
  - Elevation
  - Curb ramps
  - Monument mapping
  - Current census

- Environment
  - Species occurrence
  - Habitat assessment
  - Wildlife movement related to bridges and culverts
  - Jurisdictional waters
  - Climate change and sea level rise
  - Stormwater treatment infrastructure
  - Water and air quality monitoring
  - Trash collection, sediment, drainage
- TMS elements
  - Extinguishable message signs
  - Changeable message signs
  - Fiber optic cable
  - V-2-1 stations
  - Communications and data hubs
  - Weigh-in-motion stations
- Mobility
  - Better and more accurate freight data
  - More accurate trip, VMT, and vehicle occupancy data
  - Border wait times
  - Travel times for different modes, corridors
  - Boarding and alighting times for transit
  - Fleet mix

Data Generation. Respondents were asked to list the most valuable datasets generated by their business units. Again, responses largely fell into two main categories:

- Administrative data for management of the organization
  - Schedules
  - Financials
  - Reporting
  - Construction and contractors
- Physical world data
  - Civil3D topographic drawings and terrains
  - TMS related data
  - Bike and pedestrian counts
  - ADL burial locations
  - Paleontology sensitivity maps
  - Hazardous waste site maps
  - Contaminated property acquisitions
  - Water quality and hazardous waste
  - Stormwater quality
  - Project emissions and noise levels

Licensing and Access Management. Licensing and access management issues can be difficult to manage and result in having data fall into silos in the organization. Within Caltrans, only a few of the respondents used licensed data or data that required specialized access processes.

Data-Exchange Formats. Respondents reported that spreadsheets and textual documents are the most common data formats used for data exchange.

Table 2: Frequency of Methods Used to Assess Data Fitness

Method	Correct	Authoritative	Current	Reliable
Assume	4	4	3	2
Cross-Check Data	10	0	4	8
Ask Data Steward	2	2	2	1
Ask Source	5	7	3	4
Electronic Signature	3	1	8	3
Knowledge of Process	0	3	2	0

Metadata. The survey revealed a lack of metadata for relating datasets across stakeholders, programs, or divisions. This could be related to a lack of awareness of the data governance efforts that are underway in the organization. As such, an outreach program has been included in the recommendations. In addition, there is a lack of standards and automated tools or templates for generating metadata. When asked why, the most common responses were as follows:

- Data dictionary not applicable (4 responses)
- Have internal systems and do not share (3 responses)
- Data dictionary development is in progress, or future work (3 responses)
- Data dictionary not established (3 responses)

Provenance. Many respondents felt that they are able to obtain correct data and that their data is current. Respondents were split on whether they are able to obtain authoritative and reliable data.

Fitness for Purpose. Respondents reported using many different methods to assess data fitness for purpose, such as inferring or assuming data was clean, cross-checking with other data, asking the data steward, asking the source, having an electronic signature, and having in-depth knowledge of the process used to created the data. Table 2 displays the frequency that a method was reported as being used to check a characteristic of the data.

Overall, respondents appeared to have no mechanism to determine if any data conditioning or analytics had been applied to the data prior to their use. To accommodate having little information about the lineage of the data, many respondents perform a lot of cross-checking to make sure that the data is reasonable. This is a clear opportunity to improve the productivity and effectiveness of data use by providing data lineage through data catalogs and relieving the user of this work effort.

Future Data Needs. Finally, respondents were asked to report what data would be needed to support future needs. Some responses echoed previous answers about data for managing existing business needs. Additional data types were identified as needed in the future:

- Litter abatement
- Bicycle, pedestrian and micromobility information
- More: GIS, WIM, PeMS
- UAV
- Imagery
- Charging stations
- Point cloud
- CAV infrastructure
- Noise monitoring
- More monitoring: water, air quality, pollution
- Road characteristics
- Right of way
- CADD
- Disadvantaged Businesses, Disabled Veteran Businesses
- Derived data: Reduction in GHG, VMT, etc.

3.3 Key Areas for Improvement

Data Asset Inventory. The survey revealed a wide range of beliefs regarding the existence of data across a variety of asset types. There is also a wide variety of opinions regarding ease of digital access. These findings are particularly notable, because the survey participants primarily consisted of individuals identified as "power users" within the organization. It is unlikely that the diversity of beliefs would be reduced by expanding the survey to a larger group of participants. Based on evidence from the interviews and the survey, one key area for improvement is the data asset inventory. It will be crucial to have an actively maintained and current data catalog. Without such a catalog, it is not possible to coordinate effective data sharing and reuse. A data catalog references each of the highest value datasets and is necessary to:

- Make the data findable
- Evaluate policies and processes around the data
- Generate performance metrics about the data, usage, and users
- Manage the life cycle of the data

Data Findability and Delivery. The survey focused on evaluating data findability and usability. As noted above, respondents voiced a wide range of beliefs regarding the existence of data across a variety of asset types. This fact reveals a critical need for the establishment of detailed metadata and data catalogs for efficient discovery, access, and use of data. Data sharing within Caltrans is not a high priority across divisions or districts. Multiple departments collect data for their internal needs but without broader consideration of potential uses and the types of delivery formats that would best serve these uses. A lack of common data collection standards and semantics creates additional barriers to the delivery of data in an immediately usable format.

Location Referencing. With regards to location referencing, Caltrans uses a well-known linear reference system (LRS) that is employed for the Performance Measurement System (PeMS). As noted in other studies [KFM+20], limitations in the expressivity and usage of the PeMS metadata is such that the true physical geolocation (latitude and longitude) of a PeMS sensor is not the same as its reported latitude and longitude provided in the PeMS metadata. As a result of the limitations and inconsistent usage patterns in the metadata, it is:

- Difficult to check PeMS data for consistency and data quality,
- Difficult to use the PeMS data for modeling in integrated corridor management (ICM) applications and,
- Difficult to use the PeMS data when incorporating third-party data.

Data Quality. The results of the survey indicated a concern about data quality. Respondents were unclear about the quality of data associated with physical structures. Similar responses were received regarding other data resources.

4 Roadmap to the Future State

The process to move forward with a Data Collection Strategy in an organization the size of Caltrans requires a digital approach to the development of long-term, robust data management efforts. As the digital transformation of the organization continues, using automated tools for collecting information will provide a foundation for continued assessments of how well the organization is accommodating new approaches to managing data and its impact on the effectiveness of the business processes. The survey conducted for this initial effort provides a first step to both evaluating current practices but also in elevating the importance of this foundational effort. Continued outreach across organizational boundaries to assess the current state and communicate the vision of this plan will be imperative to success.

RECOMMENDATION: Develop the survey further and expand the engagement of the organization. Define a cadence for the survey for continued evaluation of the organization.

RECOMMENDATION: Develop a Communications Plan for the strategy going forward, including a summary of work-to-date that has already been accomplished by the current Enterprise Data and Geospatial Governance activities.

4.1 Prioritize Ontology Development for Business Value

The long-term goal is to develop a comprehensive ontology that will provide the foundation for interoperability across the organization. An ontology describes the core information concepts that drive the business processes and the relationships among the concepts. An ontology is used to ensure consistent meaning and naming across disparate disciplines, organizations, and IT systems. It can include several taxonomies. A taxonomy is a knowledge classification system that organizes the concepts into hierarchical categories. A taxonomy is a description of how the concepts are related.

Building ontologies is not a one-time, serial process. Ontologies are dynamic and they have to be managed as the needs of the organization changes. The goal is to develop a holistic approach to developing the ontologies by constantly considering the information flows among the Caltrans data owners and their business process needs.

4.1.1 Ontology Development

Ontologies can be generated from a business process perspective and also from a data exchange and management perspective. A combination of both approaches generates the most complete view and drives the most value for the organization. From a business process perspective, the goal will be to discover business process information needs that are currently lacking and determine if the information can be derived from existing data resources or whether a data procurement would be justified. This process requires a detailed understanding of the organizations' resources and the information gaps that could potentially be filled by procuring the additional data.

Because Caltrans has an existing data management activity underway, the data exchange and management perspective is a good starting point to develop a data collection strategy. The exercise that must be undertaken is to assemble business process owners for a specific topic of interest, define the organizing principles around the topic, and the context for each process owner. This effort should strive to document workflows for each process owner to tie the principles directly to current processes. These principles will guide the development of the knowledge base, metadata and data catalogs. Furthermore, because the process context is included in the discussion, the domain experts should be able to generate not only a cohesive ontology but also discover data-sharing opportunities, common data analytics that can benefit multiple process owners, and discuss the data from a fit-for-purpose perspective in order to understand how best to collect or procure data for all process owners.

To take an even broader view on data and to ensure that it is fit for purpose, the following questions must be addressed:

- Fit for purpose from a quality standpoint. Is the data statistically sound for the intended use? What is the provenance of the data? Have the salient characteristics of the data been removed by previous processing?
- Fit for purpose from a timing perspective. Is the data current enough for the intended use?
- Fit for purpose for the use case. Is the source of the data understood and relevant to the use case? Are the known biases in the data relevant to the use case?

In addition to evaluating the data resources and current data needs, the business process owners must consider future use. While predicting future needs is difficult, projecting use cases can provide some insights not only into additional data requirements but also potential data integration opportunities that can greatly enhance the data use across the organization.

A challenge for this stakeholder group will be to level the information appropriately. Starting with an existing dataset will help ground the group but might lead to overly detailed discussions. If possible, the discussion should be facilitated by an outside third party to support the leveling. The goal is to avoid overly general terms and zero in on concepts that can translate into reasonable data requirements and not lead to expensive data collection requirements that do not bring business value for the organization.

RECOMMENDATION: Establish domain-centric ontology teams or data resource teams to develop data resource-focused ontologies. The responsibility of each team is to manage the ontology for the specific resource type.

RECOMMENDATION: Assign a data steward for each ontology team. Specific responsibilities of the data steward are to manage the ontology over time; arrange meetings of the ontology team at a frequency that is most relevant for the data resource; facilitate the mapping of the onotology to metadata items that can be represented in a data catalog; ensure the freshness of the representation by extending the ontology and associated metadata as the needs of the organization change; attend cross-organization, ontology integration meetings.

RECOMMENDATION: Contract a third-party facilitator to attend the initial ontology meetings and provide leveling expertise so that the discussions generate concise and comprehensive ontologies. The facilitator should review changes to the resource ontologies over time with the data steward; attend cross-organizational meetings that focus on ontology integration.

RECOMMENDATION: Evaluate a data-cataloging tool to manage the metadata and organizational access.

RECOMMENDATION: Begin with one domain-specific topic area related directly to Caltrans' strategic goals and focus energy on building the ontology and metadata for that topic.

4.1.2 Metadata Mapping

Because of the proliferation of data, as well as the diversity of data types (structured, unstructured, semi-structures, etc.), commercial tools have been developed to build data lakes. These tools provided a mechanism to store and search large amounts of diverse data. As data lakes grow they become difficult to search and often house large amounts of stale data that is difficult to reuse because an organization's ontologies and taxonomies and their integration points change over time. The result is high storage costs for data with unknown value. To avoid this result, a concerted effort should be devoted to metadata management, including developing a data culture within the organization to tend the metadata as a part of the business process so that the metadata and its associated data resource stay current with the needs of the organization.

The onotologies represent the concepts and their relationships for the data resource. Metadata is data that describes a data resource and supports findability and usability of the data. Consistent, structured metadata provides the mechanism for navigating the data resources. There are three types of metadata: structural, administrative, and descriptive [NIS]. Structural metadata describes how the data is organized and is useful for describing relationships between data resources. Administrative

metadata describes the sourcing of the data asset, e.g., the file type, the collection information of when and where the data was sourced. It also includes the licensing, access and protection information. The National Information Standards Organization (NISO) further categorizes administrative metadata into technical, preservation, and rights. Technical administrative data describes the format and how to use the data, preservation metadata describes the temporal management of the data, and rights metadata describes the access rights. Descriptive metadata describes the content. Figure 3 shows types of information that can be used to describe data resources.

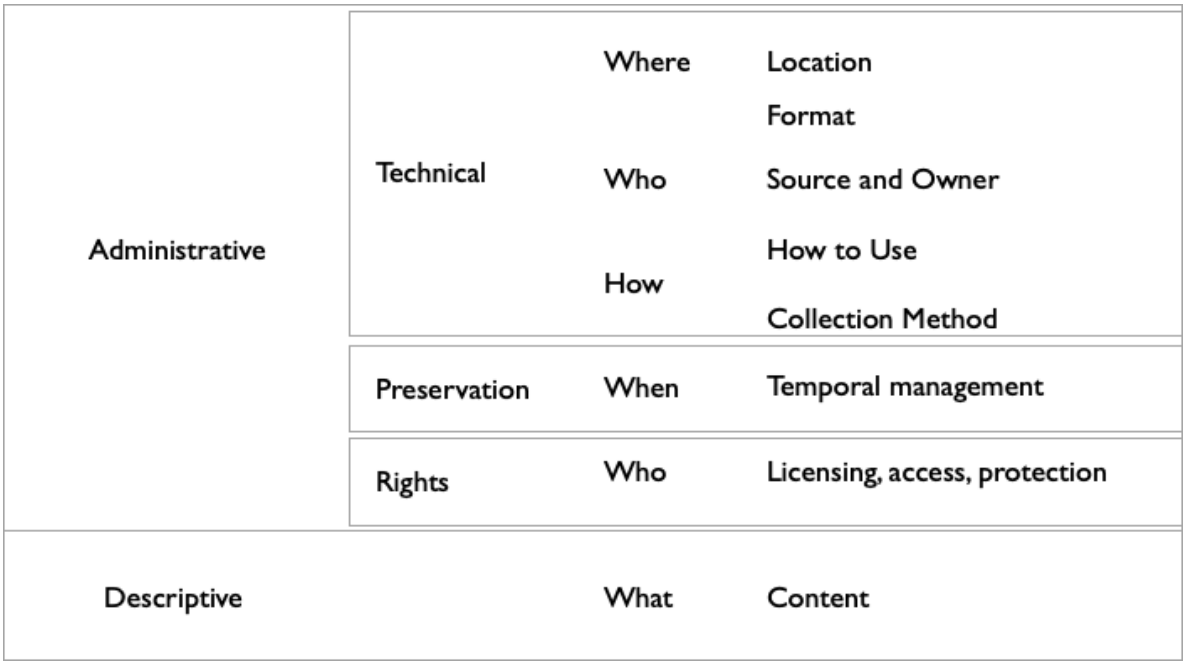


Figure 3: Metadata Structure: Administrative and Descriptive Categories

4.1.3 Data Resource Team Development

Using the existing survey, stakeholder teams can be assembled to build topical ontologies. An initial approach is to assemble all respondents that stated that a data resource is required for their business process. As an example, a stakeholder team whose charge would be to examine Bridge Data collection is suggested in Table 3. It would align well with the Caltrans **"Safety First"** goal.

Fifteen respondents are identified from our initial survey. If the team is considered too large, these stakeholder teams can be culled by picking only one person from each organization and that member takes on the responsibility of communicating within their own organization.

The accessibility of the data to this particular stakeholder group varied, and the understanding of the currency and completeness of the data also varied, as described in Table 4. This is likely a result of how the data is used in each person's process. Users make use of different data elements, or they require data at different temporal frequencies. By addressing user context and documenting the process, the data of interest can examined for fit for purpose for all stakeholders. Only then can data collection requirements be determined. A data steward should then be assigned to the data resource team. An initial choice could be based on how many members in the data resource team belong to a specific organization. A suggested data steward for the Bridge ontology development is Clint Peace, District 11.

An alternate approach would be to focus effort on FHWA requirements. The upcoming MIRE 2026 requirements align well with the Caltrans **"Safety First"** goal. The Fundamental Data Elements (FDE) are well defined and require focused data collection and associated data extraction activities. With the types of source data — likely aerial imagery, LIDAR, or street view data — the raw data

Table 3: Initial Bridge Ontology Generation Team

Name	Organization
Clint Peace	Senior TE, District 11, Project Management
Christopher Dennis	Sr. Engineering Geologist, Headquarters, Project Delivery, Environmental Analysis
Jesus Mora	Acting PM for VDC, Headquarters, Project Delivery, Design
Stefan Sutton	Senior Environmental Planner, Headquarters, Project Delivery, Environmental Analysis
Subu Nujella	Sr. Transportation Engineer, Headquarters, Project Delivery, Design
Dick Fahey	District 4, Transportation Planning
Oscar Aguilar	Project Engineer, District, 11, Project Management
Jimmy Walth	Senior Environmental Planner, Headquarters, Project Delivery, Environmental Analysis
Joseph Watkins	Staff Services Manager I, Headquarters, Project Delivery, Design
Daisy Laurino	TE, Headquarters, Project Delivery, Environmental Analysis
Laura Rose	Associate Transportation Planner, District, 2, Transportation Planning
Jon Bevan	Maintenance Manager II , District, 10, Maintenance
Luz Quinnell	Headquarters, Project Delivery, Environmental Analysis
Patrick Lo	SB 1 Office Chief, Headquarters, Finance, SB1 Program
Kyle Singh	GIS Coordinator, District, 6, Environmental Analysis
Lynn O'Connor	Office Chief, Systems Planning and Goods Movement, District, 10, Transportation Planning
Deniz Ozakcay	GIS Specialist, District, 11, Transportation Planning
Loren Turner	Office Chief, HQ Asset Management, Headquarters, Finance, Transportation Asset Management
Mihai Giurgiulescu	GIS Analyst, District, 9, Project Management

and derived data are good candidates for "collect once and use many times." This, however, is heavily dependent on the state of the new LRS and how well the data has been captured, extracted and conflated into an existing database. This focus would elevate the important role of Caltran's LRS, the data extraction efforts, and integration capabilities.

4.1.4 Existing Metadata and Data Dictionary Guidance

Appendix A describes the current guidance for creating metadata. The guidance maps directly to the metadata structure in Figure 3. Appendix B describes the current guidance for creating a data dictionary.

Data associated with State Bridges was downloaded from the California Open Data Portal, https://data.ca.gov/dataset/state-highway-bridges, along with the associated data descriptions. As can be seen in Table 5, the data dictionary contains the column headers of the downloaded CSV file. While the dataset is rich with information, the column labels do not provide meaningful descriptors. As such, new or unfamiliar users are left to infer the meaning of the data or must find the data owner to confirm their inferences. The state of this partial description implies a lack of a comprehensive data dictionary or a poor linkage to an existing data dictionary. Either case is unfortunate for the individual downloading the data. Improving the data dictionary that is associated with the bridges dataset should be the first step addressed by the data resource team associated with physical structures.

Table 4: Current Evaluation of Bridge Data in Context of the Need

Name	**Level of Digital Accessibility**	**Up to Date**	**Complete**
Clint Peace	Digitally findable, accessible, and usable without reformatting or data wrangling	Don't Know	Yes
Christopher Dennis	Don't Know	Don't Know	Don't Know
Jesus Mora	Digitally findable, accessible, and usable without data wrangling	Yes	Yes
Stefan Sutton	Digitally findable without consulting Data Steward	Don't Know	Yes
Subu Nujella	Digitally findable and accessible by direct download	Don't Know	Yes
Dick Fahey	Digitally findable and accessible by direct download	Don't Know	Yes
Oscar Aguilar	Digitally findable without consulting Data Steward	No	No
Jimmy Walth	Findable via co-worker	No	Yes
Joseph Watkins	Digitally findable and accessible by direct download	Yes	Yes
Daisy Laurino	Digitally findable without consulting Data Steward	Don't Know	No
Laura Rose	Digitally findable and accessible by direct download	Don't Know	Yes
Jon Bevan	Digitally findable and accessible by direct download	Yes	Yes
Luz Quinnell	Digitally findable and accessible by direct download	Yes	Yes
Patrick Lo	Findable via co-worker	Don't Know	–
Kyle Singh	–	Don't Know	–
Lynn O'Connor	Digitally findable and accessible by direct download	Yes	Yes
Deniz Ozakcay	Digitally findable and accessible by direct download	Yes	No
Loren Turner	Digitally findable and accessible by direct download	Yes	Yes
Mihai Giurgiulescu	Digitally findable and accessible by contacting Data Steward	No	Yes

Table 5: Bridge Information Data Dictionary

Name	Type	Inferred Definition?
DIST	Number	no
CO	Text	no
BRIDGE	Text	no
BRIDGE_X	Number	no
BRIDGE_Y	Number	no
LAT	Number	yes - latitude of bridge center?
LON	Number	yes - longitude of bridge center?
NAME	Text	yes - name of bridge e.g Bay Bridge
LOC	Text	no
YRBLT	Number	yes - year bridge was built
HST	Number	no
FAC	Text	no
APWID	Number	no
LENG	Number	yes - length of bridge (Units?)
DK_AREA	Number	no
LSW	Number	no
RSW	Number	no
RDW	Number	no
REFVCU	Text	no
VCU	Number	no
MAINSPANS	Number	yes - number of main spans
DIR	Number	no
PRINC	Number	no
INTERSEC	Text	no
AADT	Number	yes - average annual daily traffic on bridge
PCTTRK	Number	no
DEF	Number	no
NHS	Number	no
FUNCTIONAL	Number	no
DATA_EXTRA	Text	no
PM	Text	no
CITY	Text	yes - city in which bridge is located
RTE	Text	yes - route for the bridge e.g. I80

RECOMMENDATION: Encourage use of the metadata and data dictionary guidance in Appendices A and B by referencing this guidance explicitly in a Caltrans-wide Communications Plan.

RECOMMENDATION: Establish a document management system and data cataloging system or data registry for linking workflows, ontologies, metadata, and data dictionaries to raw data files.

4.2 Establish a Data Sharing Organizational Structure and Culture

4.2.1 Processes for Building a Shared Data Culture

With the focus on data collection as a first step, processes for communicating and cross-checking decisions to collect or purchase data will contribute to building the Shared Data Culture.

Figures 4 through 7 present draft processes for moving toward a Shared Data Culture. These diagrams present five roles in the data acquisition process (represented as the horizontal bars in the diagrams):

- **Business Data Steward.** The business process owner determines the need for the data in their work and makes the decision to purchase data. A detailed description of current responsibilities of the business data steward is in Appendix C.

  Responsibilities: Identify data necessary for the business process. Identify requirements for collection and purchase, e.g., new attributes, temporal refresh, geospatial extension or new type of data required.

- **Enterprise Data Steward.** The enterprise data steward rationalizes the datasets being purchased for their specific domain. A detailed description of current responsibilities of the enterprise data steward is in Appendix C.

  Responsibilities: Manage the data dictionaries in the data catalog, register data, discuss collection and purchases with the cross-organization data resource teams, review ontologies and taxonomies with the data resource team and business process owners. Approve data purchases identified by procurement.

- **Data Resource Team.** The data resource team is a selected group of cross-organizational business process owners who use the specific data in their work. A data resource team member may be a business process owner or simply reuse data purchased by another organization. The team will require a team lead who will represent the organizational value of the data to current business processes.

  Responsibilities: Build and maintain the ontologies and taxonomies, and align data dictionaries for cross-functional use cases. Document use and value of data in Caltrans business processes.

- **Procurement Specialist.** A person from procurement who has been requested to acquire data for the business process owner.

  Responsibilities: Identify data purchases that are considered a part of corporate data. Alert the appropriate enterprise data steward (mechanism for alert depends on technology solution) to any data purchases that have tags associated with corporate data. Wait for approval from the enterprise data steward before purchasing requested data.

- **Data Generator.** The data generator could be an external vendor or an internal Caltrans team.

  Responsibilities: Collect data. Provide detailed descriptions of collection process and any analytics that have been applied to the raw data. Descriptions must adhere to the metadata and data dictionary requirements included in Appendices A and B.

The top horizontal bar in figures 4 through 7 indicates the types of data structures and that would support the data acquisition process. Digital tools can potentially provide mechanisms to capture these structures and automatically initiate some of the actions indicated in the processes. For example, notifications and report generation can likely be automated via a data cataloging tool. The data ontologies and data dictionaries could be maintained in a web environment (e.g., a wiki such as Confluence) or a shared document repository (e.g., Microsoft SharePoint), provided that there is a linkage to the original datasets via the metadata.

Purchasing data begins with a Business Process Owner determining the need for a dataset for their specific purpose. We will assume the existence of a data catalog because this is a core recommendation of the strategic plan. At a high level, there are two basic cases: (1) the data type is known and has been purchased before, or (2) its a new data type, thus requiring additional reporting and communication requirements.

Known Data Type Acquisitions. Figures 4 and 6 describe the simple purchase of a known data type. If the data is known and has been purchased before, the goal of the process should be to quickly facilitate the purchase and support the business process owner's need. There are few known data type categories that might be considered: same type of data, temporally refreshed; same type of data, new geospatial region; same type of data, with additional attributes. Following the initial identification of the need for the data, the business data steward should consult the data catalog to determine if the data already exists. The business data steward should define the specifications of this new data. The specification should not change the core metadata. It should just specify the parameters associated with the collection process, e.g., new spatial region. If additional attributes are requested, specifications for those attributes must be included in the request to the enterprise data steward. If the data does not appear in the data catalog, the enterprise data steward is notified of the need for the data, and a procurement or internal collection is initiated. The business data steward then registers the new dataset for others to use.

The procurement specialist's role is to be aware of data purchases and notify the appropriate enterprise data steward. The procurement specialist uses the keyword tags found in the data dictionaries to filter all data purchases. If the data purchase hits on a data tag, then matches to the data catalog are generated and the associated enterprise data stewards are notified. The specific category of purchase should be highlighted in the report so that the correctness of the data registry can be validated by the enterprise data stewards. The procurement specialist should also ensure that the external data generator provides documentation of the full provenance of the data and that it is included in the purchase agreement. Purchases greater than a certain threshold should require approval by the Data Steward.

In parallel, the enterprise data steward is notified of the plan to purchase, and of any actual purchases. This reporting provides the enterprise data steward a current understanding of all data purchasing in their domain with documentation of temporal updates, additional geospatial coverage, and extended datasets. If a data registry or catalog is in place, it can notify all current data users of refreshed data.

New Data Type Acquisitions. Figures 5 and 7 describe a high-level approach to managing new datasets. This process may include the data resource team that maintains the domain ontology and the associated data dictionaries. The data resource team's role is described in Section 4.1.3 and is primarily focused on building a comprehensive view of the domain and any data that provides insights into the domain and the associated business processes that use the data. As the organization builds a shared data culture, new business process owners may emerge that generate an entirely new view of a domain and its data resources, for example, if new data associated with new sensing technologies emerges for monitoring mobility.

In this case, the enterprise data steward plays an important communication role in the process. Notification from the business data steward of a purchase that doesn't match any data catalog entries will be evaluated by the enterprise data steward to determine if the ontology and associated data dictionaries will need to be adjusted to include this new data type. The enterprise data steward decides if the data resource team needs to be engaged in the discussion before the purchase. If not the enterprise data steward approves the new data structure, and the procurement is initiated.

Procurement registers a new data type if the data purchase does not match any existing data tags. The data purchase is highlighted to an enterprise data steward that is associated with the content category. This process creates a cross-check to ensure that hidden data purchases do not slip through the process. Purchases greater than a certain threshold should require approval by the enterprise data steward.

As mentioned in Section 4.1.3, the enterprise data steward meets with the data resource team as needed and the reports of these purchases are reviewed. Ontology and data dictionaries are modified

Figure 4: Process Diagram for Internal Acquisition of Known Data Type

Figure 5: Process Diagram for Internal Acquisition of Unspecified Data Type

Figure 6: Process Diagram for External Acquisition of a Known Data Type

Figure 7: Process Diagram for External Acquisition of Unspecified Data Type

as necessary to create a comprehensive view of the domain content and its associated datasets. It is not expected that ontologies change often, however, emerging technologies and their associated datasets may extend the measurability of the domain's attributes. This meeting should also address any validation requirements for the new data.

These processes in combination provide a simple, high-level draft process to embed repeatable steps for data collection with inter-organizational cross-checks to avoid hidden data purchases.

RECOMMENDATION: Caltrans should require that data identified as corporate data be curated in a data repository within a reasonable amount of time.

4.2.2 Technology to Identify External Data Collection Activities

Procurement Process. Adjusting the procurement process to identify data purchases will be important to the governance process. While data purchase processes may be in place for users to follow, adherence can sometimes lapse or new employees may be unaware of the processes. With metadata keyword tags, automated processing can be integrated to scan purchase orders and identify data collection activities and purchases.

RECOMMENDATION: Invest in automated technology for identifying data collection contracts or data purchases and provide feedback to the enterprise data stewards for review.

Identifying Hidden Data Users. Contract and project files can contain information related to historical data purchases. This can provide a valuable opportunity to identify data users in the organization. A digital approach to assessing past data procurement activities is possible with current technology. Technologies that support this include text analysis and natural language processing of digital documents and scanning of printed text.

RECOMMENDATION: Consider reviewing historical data collection contracts or data purchases with feedback to the data resources team for the specific domains.

4.2.3 Proposed Organizational Extensions

In the previous section, we recognize the need for the assembly of data resource teams that represent the use and value of the data in current business processes. While these meetings could be ad hoc, recognition of these teams and their importance to generating and maintaining a cohesive approach to data management is best created by an organizational extension that explicitly identifies the teams. **This further strengthens the visibility of "data as an asset that must be resourced and managed like all other physical assets."**.

In addition to the data resource teams, we also recommend assembling an additional team to provide oversight to the development of the data catalog(s) and , as detailed by the data resource teams. Figure 8 identifies the scope of this team as presiding over the metadata development of all topical areas. Responsibilities of the individual data resource teams are shown in Figure 9. We suggest that all data resource team leads are members of the data resource oversight committee.

Data Resource Oversight Committee. The data resource oversight committee is a selected group of cross-organizational business process owners who are considered power users.

Responsibilities: Assist in building and maintaining the ontologies and taxonomies and aligning data dictionaries for cross-functional use cases. Identify skill sets and recommend training necessary for organizations who wish to use, collect or procure the data. Identify data integration opportunities across topical areas that will increase the value and reuse of datasets. Document use and value of data in Caltrans business processes.

Given the survey results, the respondents were ordered by the number of data touch points associated with their work efforts. These respondents were *arbitrarily partitioned* into three tables to provide groups of subject matter experts for developing the metadata. Table 6 represents a group that uses data associated with many of the topical areas across all of the functional areas. As such, they are

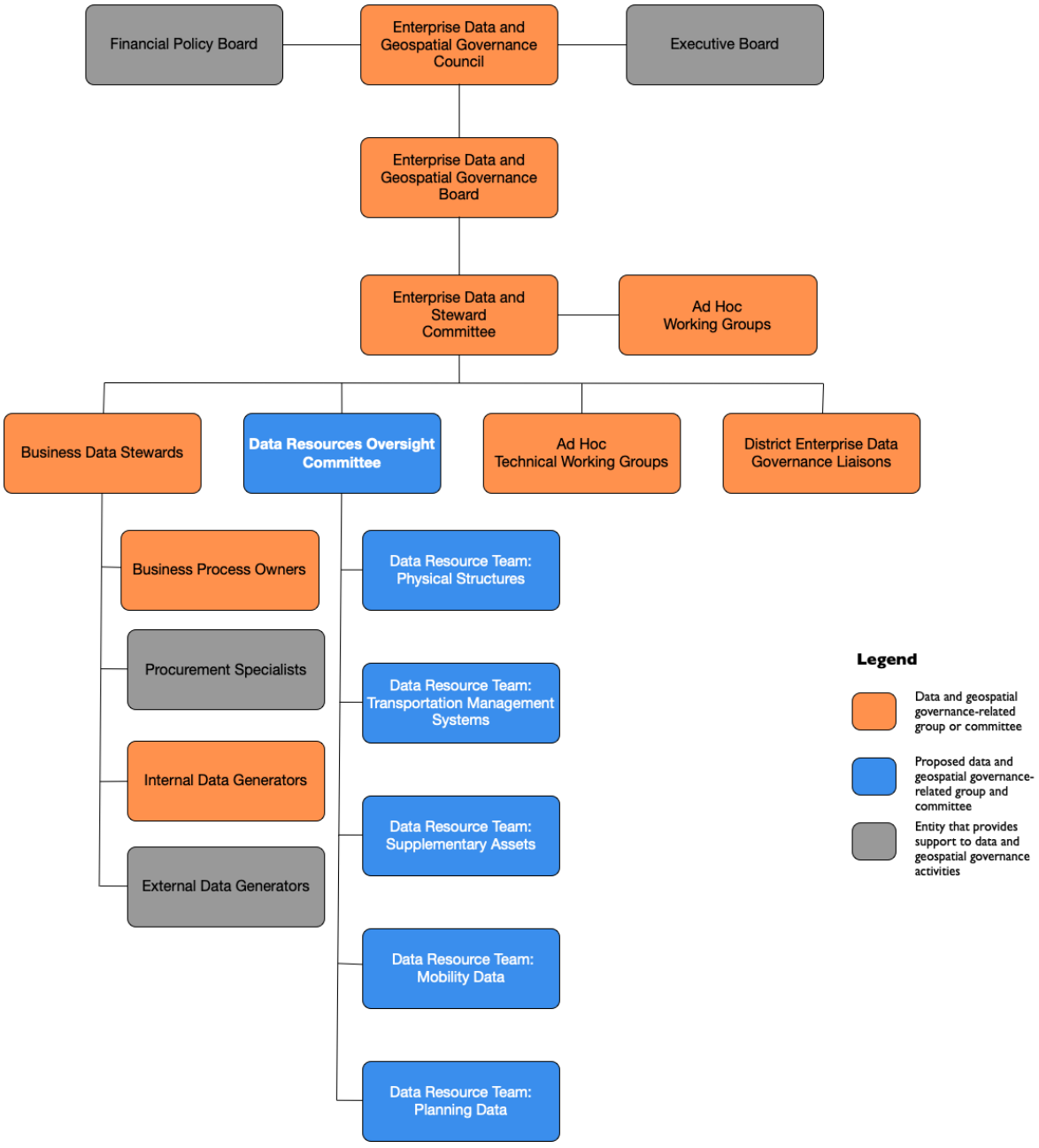


Figure 8: Data Resource Committee and Teams

Figure 9: Data Resource Team Responsibilities

potential recruits as data resource team Leads. Table 7 identifies people who also use a considerable amount of data and could potentially serve as topical area experts and members of data resource teams. And finally, Table 8 could serve on specific topical area teams.

Building out the teams will require outreach and support to Caltrans management that emphasizes "data as an asset that must be resourced and developed as any other physical asset". There must be a commitment to the effort and an agreement that the respondents are allotted time for the activity. This support must come from top management in Caltrans.

Table 6: Data Resource Oversight Team Candidates and Topical Area Specialists

Name	**Data Touch Points**	**Data Needed**
Laura Rose	39	All physical highway structures, all transportation management systems, all supplementary assets, all mobility data except GPS, all planning data
Dick Fahey	39	All physical highway structures except pavement markings, all transportation management systems, all supplementary assets, all mobility data, all planning data
Mihai Giurgiulescu	36	Pavement, pavement markings, barriers, bridges, culverts, guardrail, static signs, highway lighting, changeable message signs, closed circuit cameras, traffic census stations, roadway weather information systems, traffic monitoring detection stations, traffic signals, control cabinets, all supplementary assets, all mobility data, all planning data
Loren Turner	35	All physical highway structures, all transportation management systems, all supplementary assets, all mobility data except intersection turning counts, safety data and injuries, other GIS
Lynn O'Connor	31	Pavement, pavement markings, barriers, sound walls, bridges, static signs, changeable message signs, closed circuit cameras, traffic census stations, traffic monitoring detection stations, freeway ramp meters, traffic signals, ADA infrastructure, roadside rest facilities, park&ride infrastructure, weigh-in-motion scales, sidewalks, bicycle and pedestrian facilities, all mobility data, all planning data
Subu Nujella	30	All physical highway structures, all transportation management systems, irrigation and drainage, ADA infrastructure, roadside rest facilities, park&ride infrastructure,weigh-in-motion scales, sidewalks, bicycle and pedestrian facilities, speed, VHT and VMT, intersection turning counts, freight and goods movement
Alexander Walton	27	Pavement, barriers, bridges, culverts, guardrail, static signs, highway lighting, changeable message signs, closed circuit cameras, traffic census stations, roadway weather information systems, traffic monitoring detection stations, highway advisory radios, traffic signals, control cabinets, Vegetation, ADA infrastructure, roadside rest facilities, park&ride infrastructure, weigh-in-motion scales, sidewalks, bicycle and pedestrian facilities, all planning data

Table 7: Data Resource Topical Area Team Candidates

Name	**Data Touch Points**	**Data Needed**
Deniz Ozakcay	27	Pavement, sound walls, bridges, culverts, guardrail, crash cushions, static signs, all transportation management systems, all supplementary assets, land use and parcels, safety data and injuries, other GIS
Kevin Tucker	23	Pavement, pavement markings, static signs, highway lighting, traffic census stations, traffic monitoring detection stations, traffic signals, ADA infrastructure, park&ride infrastructure, sidewalks, bicycle and pedestrian facilities, flow and volume, speed, VHT and VMT, trip and mode choices, origin-destination, intersection turning counts, GPS traces, all planning data
Diane Jacobs	16	Traffic census stations, traffic monitoring detection stations, freeway ramp meters, traffic signals, park&ride infrastructure, bicycle and pedestrian facilities, flow and volume, speed, VHT and VMT, trip and mode choices, origin-destination, intersection turning counts, GPS traces, freight and goods movement, safety data and injuries, transit data
Joseph Watkins	16	All physical highway structures, changeable message signs, all planning data
Jon Bevan	16	All physical highway structures, changeable message signs,closed circuit cameras, roadway weather information systems, highway advisory radios, traffic signals, control cabinets
Jesus Mora	15	Pavement, barriers, sound walls, bridges, culverts, guardrail, crash cushions, static signs, changeable message signs, highway lighting, irrigation and drainage, ADA infrastructure, weigh-in-motion scales, sidewalks, bicycle and pedestrian facilities
Patrick Lo	15	Pavement, bridges, culverts, changeable message signs, closed circuit cameras, traffic census stations, roadway weather information systems, traffic monitoring detection stations, highway advisory radios, freeway ramp meters, traffic signals, control cabinets, ADA infrastructure, sidewalks, bicycle and pedestrian facilities
Christopher Nicholas	12	ADA infrastructure, roadside rest facilities, park&ride Infrastructure, weigh-in-motion Scales, sidewalks, bicycle and pedestrian facilities, VHT and VMT, trip and mode choices, origin-destination, freight and goods movement, land use and parcels, other GIS

Table 8: Data Resource Team Support Candidates

Name	Data Touch Points	Data Needed
Julia	11	Vegetation, irrigation and drainage, ADA infrastructure, roadside rest facilities, park&ride infrastructure, weigh-in-motion scales, sidewalks, bicycle and pedestrian facilities, census tracts, land use and parcels, transit data
Christopher Dennis	11	Pavement markings, bridges, culverts, guardrail, Vegetation, irrigation and drainage, ADA infrastructure, roadside rest facilities, park&ride Infrastructure, sidewalks, bicycle and pedestrian facilities
Luz Quinnell	11	Pavement, pavement markings, barriers, sound walls, bridges, culverts, guardrail, crash cushions, static signs, changeable message signs, highway lighting
Oscar Aguilar	11	Pavement, barriers, bridges, culverts, guardrail, crash cushions, changeable message signs, closed circuit cameras, freeway ramp meters, traffic signals, ADA infrastructure
Clint Peace	11	All physical highway structures, changeable message signs
Stefan Sutton	10	barriers, sound walls, bridges, culverts, freight and goods movement, census tracts, land use and parcels, safety data and injuries, transit data, other GIS
Tim Hart	9	park&ride infrastructure,weigh-in-motion Scales, flow and volume, speed, VHT and VMT, trip and mode choices, origin-destination, safety data and injuries, transit data
Daisy Laurino	9	Pavement, sound walls, bridges, traffic census stations, flow and volume, speed, trip and mode choices, origin-destination, intersection turning counts
Brian Pecus	8	Closed circuit cameras, traffic census stations, roadway weather information systems, traffic monitoring detection stations, highway advisory radios, freeway ramp meters, traffic signals, control cabinets
Mohammad Iraki	8	Closed circuit cameras, traffic census stations, roadway weather information systems, traffic monitoring detection stations, highway advisory radios, freeway ramp meters, traffic signals, control cabinets
Kien Le	7	Changeable message signs, closed circuit cameras, roadway weather information systems, highway advisory radios, roadside rest facilities, flow and volume, VHT and VMT
William Pan	4	Vegetation, irrigation and drainage, roadside rest facilities, Park & Ride Infrastructure
Kyle Singh	4	Pavement, pavement markings, bridges, culvert
Bhaskar Joshi	3	Pavement, sound walls, culverts
Varsha Kotla	3	Sidewalks, bicycle and pedestrian facilities, other GIS
Jimmy Walth	2	bridges, Culverts
George Anzo	2	Static signs, changeable message signs
Bruce Rymer	1	Sound walls

4.3 Workforce Development

Workforce development efforts in data collection will be an ongoing activity for the organization because as new technologies are continuously emerging. Workforce development will be a large-scale effort within the organization and will require a significant investment by Caltrans. The following sections provide guidelines for establishing the development effort. A full-scale workforce development program is beyond the scope of this strategic plan for data collection.

To best facilitate the transformation, the impacted business processes must be identified. Digital capabilities will likely improve efficiencies for asset management, maintenance and real-time sensing

and provide higher level control. The technology development effort must separate and recognize the difference between using digital resources and capabilities for supporting employee productivity and workforce development efforts and the integration of emerging technologies that can contribute to the core business processes of Caltrans. Separate budgets and action plans should be developed for both such efforts.

4.3.1 Identify the Technical Foundations for the Future Workforce

For workforce development efforts, the first major step is to identify the technical foundations for the future workforce. Example foundation technologies will likely include:

- Data science and communications technology
- IoT sensing
- Data collection methods
- Big data analytics
- Integrated data systems – data systems management and analytics design
- IT security
- Data privacy

Design the Workforce Skill and Position Mix To design the workforce skill and position mix, the focus should be on the high-level priorities of the organization. Decisions regarding collecting data with internal resources or buying data collected by an outside vendor will be key to the determining the skills needed in the organization. These decisions should be made by the business process owners associated with the topical area. With these decisions made, the next step is to identify the core competencies that will be necessary to manage the collection or purchase. For internal collection, there might be a significant data conditioning and data analytics effort, depending on the topical area being addressed. For example, LIDAR collection requires significant data analytics to extract the features of interest. If the data collection effort is outsourced, a process should be established by the ontology teams for finding and evaluating complementary partners with a description of the expertise and experience necessary for the topical area. The metadata descriptions created by the ontology teams should reflect the features and specifications of data quality necessary for the topical area.

Organizational Design Challenges. New job classifications will be necessary to allow the organization to competitively recruit the necessary skill sets. There will be several steps to building these new skill sets, including determining the industry-recognized credentials for the new job classifications, establishing programs for attracting and recruiting the talent, and identifying likely challenges for recruitment and some associated mitigation strategies.

In addition to recruiting the new talent into the organization, a parallel effort focused on identifying the management training necessary to align the newly recruited staff with the foundational knowledge of the organization. This will focus primarily on understanding the gained business process value-add from the introduction of data analysts and software engineers. Support for this transition should be handled by an organizational change management specialist.

4.3.2 Identify Organizational Design Modifications to Accelerate the Transformation

A proposed approach is to establish a special projects team for a specific topic of interest. The ontology team that is made up of business process owners will develop the new job classifications or training classes for their topical area of interest. A special projects team of data analysts and software engineers could be created from internal talent and recruited talent to work specifically on one topical area of interest. By beginning the process as a special projects team, acquiring talent may be easier by acknowledging the importance of this new skill set and could aid in developing untethered innovative solutions. However, it will require sensitivity and understanding of how best to recruit and integrate this team with an existing workforce. In particular, a parallel effort focused on training and development of the existing workforce should be implemented and communicated widely across the

organization so that existing employees have opportunities to grow into the new job classifications. The special projects team would then recruit and build out the team(s) as more topical areas of interest are addressed. The team responsible for organizational change would determine where in the organization these teams should reside. The key sensitivity to the organizational design is to create well-integrated data teams so that analysis teams are not siloed in topical areas.

The parallel effort in training should be designed with outside organizations. Example approaches would be to develop apprenticeship programs and internships. This will create two-way learning potentials where the business process owner can bring in fresh minds and solutions while building excitement for new recruits to learn the business process and return to the organization for career opportunities. In addition, coordination with universities and other regional or national transportation agencies should focus on developing specific courses for workforce development as well as pre-employment training to aid in the recruitment of new graduates.

Establish Budgets and Timeframes for Workforce Development. In this new era of digitization, training and investment in the development of the work force will be a continuous activity. As such, the approach must be intentional – meaning that it must have a well-defined budget, expected outcomes with Key Performance Indicators and support from upper management and a process owner. Their position should be defined by the organizational change management team.

Proposed Initial Steps. Define a short-scoped work effort focused on the following items:

- What should the next-generation work force look like? What additional types of personnel are needed and what should be outsourced to external vendors?
- What are the trends in skill mix? How many software development engineers, software project managers, and data scientists should be resident on a team?
- What are other DOT's doing in this area? What are their successes in recruitment programs? What kind of technology profiles are they staffing to? What percent of their data collection activities are outsourced? What are considered best of breed in organizational design approaches?

4.4 Summary of Initial Steps

RECOMMENDATIONS:

- Initiate a cross-organizational workforce development effort and a communications program associated with data literacy.
- Choose and implement a cross-organizational data cataloging tool. This could be an existing tool already used within Caltrans, a purchased tool, or an open source tool.
- Implement a data registry.
- Establish and initial data resource team for the topical area of bridges. A suggested data resource team has been identified based on the initial survey.
- Define, with the aid of a facilitator, the bridge ontology and work with the enterprise data steward to define the metadata, according to the Caltrans metadata guidance. This group then adds this to the data catalog.
- Harvest existing bridge data, consolidate and align to metadata and add to data catalog.
- Review successes and difficulties with the processes, including the chosen data cataloging tool. Adjust process to Caltrans cultural norms.
- Select more topical areas in the physical highway structures area and repeat these steps with additional data resources teams, if new teams would be useful.
- Expand efforts across all functional areas, creating at least the five identified data resource teams.
- Report out successes and difficulties through the communications plan associated with the workforce development effort.

The steps identified above will be a mechanism to demonstrate the value of existing data assets that are being used to manage Caltrans business processes through an

- expanded understanding of how data is used and how it will improve Caltrans business processes,
- expanded awareness of cross organizational use of data,
- initial development of shared data culture.

All existing topical datasets will be curated and integrated into data catalog.

Through workforce development efforts and a communications plan an effort should be made to communicate across organizations the existence of these datasets and data catalogs and registries. Efforts should be made to develop the skill sets for discovering and using data and the processes for requesting data to be collected or purchased, and to build an organization-wide awareness of the value of a shared data culture.

4.5 Beyond Data Collection

Databases and data lakes are more than just data storage locations; they provide back-end tools to empower the publication of quality-controlled data. We have assumed a level of curation is being accomplished at collection time that must be documented as part of the data delivery process. However, additional analytics can be very valuable in post processing of data, particularly as novel uses of the data emerge and cross-organizational interest grows.

To assist in understanding current analytics capabilities, an evaluation of Caltrans in-house data analytic capabilities should be initiated. This evaluation can be included as a deliverable from the Data Resource teams. As they document the business processes that use the data, the analytics applied to the data should be documented. This documentation will allow reuse and in-house open-sourcing of code, which should lead to less duplication and more robust data processing. Pointers to code repositories that are associated with the data can be included in the data dictionaries, and new derived datasets can be tracked by the data catalog.

RECOMMENDATION: Implement an annual review and approval process for evaluating published data and its provenance. Establish a mechanism for assessing the uncertainties in data in context of its entire life cycle from raw data collection to published data. This evaluation should include metadata and ancillary data (e.g., weather, special events, sensor health) that led up to the final published dataset, including the raw sensor data, validation data, calibration parameters and notes on the data collection, QC, and approval, thus providing a a record to understand why and how the published data was developed. Uncertainty measures may be derived based on post processing or cross-validation processes. A tag to guide a prospective data user could be developed, similar to the USGS Streamgage Program that uses a good, fair, poor uncertainty rating system with defined percentage error for each category.

In parallel, and also beyond the scope of this report, training and mentoring in data use and evaluating data collection vendors should also be initiated. An initial effort should focus on data literacy. "Data literacy is the ability to read, write and communicate data in context. This includes an understanding of data sources and constructs, analytical methods and techniques applies, and the ability to describe the use case, the application and the resulting value [RDL19]". Data literacy is of particular importance when procuring data that has already been processed by the vendor. Asking the right questions about the processed data and assessing the derived data in context of the raw data is extremely important when making a data purchase.

RECOMMENDATION: Support and encourage current data literacy workforce development program activities.

5 Technologies

Given the current state of data maturity, two key recommendations regarding important fundamental infrastructure requirements are:

RECOMMENDATION: Caltrans should establish a centralized or federated registry listing key data repositories, data catalogs, and other data resources with reuse potential outside of their originating business unit.

RECOMMENDATION: Caltrans should establish central or federated data repositories tied to either business areas (i.e., physical structure management, transportation system management) or core business units (i.e., Districts). Data within these repositories should be discoverable through data catalogs using standards-based metadata. Repository implementations should plan to handle structured and unstructured data.

Like other procurements, technology selection has several dimensions, including cost, locality (cloud or on-premises operation), vendor support, integration with existing resources, and configurability or potential customization. For data collection, there are two dimensions that should be highlighted that are more unique to this area: the level of data specificity and the amount of centralization.

Data specificity is the tendency of solutions to provide a rich set of features for a limited set of data. A relevant example are repositories for GIS data that can enable searching by spatial coordinates and perhaps providing visualization of the data discovered. Examples of less data specific repositories are Microsoft SharePoint or Google Drive. These services will store basic metadata for files, like owners, creation dates, and often extracted text; these fields enable very useful discovery across a broad range of file types. However, without customization, they tend to lack the capability to do more than provide a file or present its contents. Further, some types of data may require a specific class of solutions, which exclude other data categories. In particular, databases and streaming data (i.e., sensor or other real-time data feeds) require optimized solutions. Deciding how specific a solution should be is determined by estimating how much data of particular types is being actively used and the criticality of the use cases.

The degree of centralization is how much of the solution is controlled, managed, or operated by a single unit with an organization. A high degree of centralization is often a positive quality, especially for solutions that support critical business functions that are uniform throughout the organization (e.g., payroll). Drawbacks to centralization may arise when individual business units have unique needs that require distinct modifications, e.g., to interact with legacy equipment or vendor-defined processes. Resiliency and availability may be improved through more distributed solutions by ensuring that individual business units are not impacted by a single failure point. However, other disadvantages arise for highly distributed or non-centralized solutions, especially ones that handle similar business needs without any level of interoperability. This fragmentation can lead to a significant amount of human effort to find and extract data.

Between these two extremes of centralized and distributed solutions, federated solutions can provide a level of local control while ensuring that a necessary level of interoperability exists. Federation is the ability to join solutions, like data repositories, to act as a single resource. Features including authentication, data discovery, and data access are common goals of federated solutions. Depending on the level of standardization or interoperability, solutions from multiple vendors can be federated. If the level of customization needed by several business units is significant, federating distributed resources can provide for the necessary independence while still supporting the higher business goals (e.g., global data discovery).

5.1 Current Technology Solutions

As in any technology solutions, vendor lock-in issues like licensing costs, interoperability restrictions and proprietary formats, must be considered. Ideally, open-source systems can provide adequate, if not better, solutions that improve as underlying technologies improve and get incorporated into the software with community contributions and support.

Data catalogs. A data catalog provides a comprehensive view of an organization's data assets. It utilizes metadata to enable an organization's users to quickly search and access data regardless of its location (central or distributed).

Data catalogs implemented using metadata standards enable rapid discovery of data and interoperability and promote reuse and trust of the data. Popular metadata standards include Dublin core, ISO 19115 (Geographic Information), the Federal Geographic Data Committee (FGDC) metadata standard, and Project Open Data Metadata Schema.

Data catalogs must be dynamic and need to adapt and grow with an organization as it addresses the digital transformation that many organizations are experiencing. As data becomes more available and increasingly complex, without a data strategy the result is a fragmented collection of data sources that are siloed, not well documented, and difficult to find, leading to duplicative efforts to collect data, distrust in data that has been collected, and misuse of the data.

We provide some sample technology solutions that highlight how many companies and open-source efforts are addressing the requirements of next-generation data management. It is not in the scope of this project to recommend data management support tools.

5.2 Data Cataloging Solutions

Commercial data catalogs include Snowflake, Databricks, and the traditional vendors like IBM. Two examples of open-source data management solutions are CKAN and Esri Geoportal.

CKAN - Comprehensive Knowledge Archive Network (CKAN) is an open-source data management solution for publishing and sharing data. It provides a rich set of features for free and is extendable via modules. It is the technology behind several government open-data portals, including data.gov (US), open.canada.ca, and opendata.swiss.

Esri Geoportal - The open-source Geoportal Server from Esri enables search and discovery of geospatial resources. Organizations that produce data can publish their resources, including data, tools and web services, by registering the resource's metadata with the geoportal's catalog, an Open Geospatial Consortium (OGC) compliant CS-W 2.0.2 service. Several earth science data facilities like, OpenTopography, use Esri Geoportal for publishing their data inventory.

5.3 Machine Learning Data Catalogs

Data catalog success is driven by the use of metadata standards and depends on the adherence of the organization to guidelines. Machine learning is providing new capabilities for data cataloging tools that enable automated discovery and streamlined ingestion that supports metadata creation. Some claim to generate knowledge graphs from the data. Scaling collaboration, providing detailed data lineage, and an ability to automatically infer metadata may allow the tools to provide contextualized, trusted, accessible data to their customers. These kinds of tools could streamline the activities of the Data Resource teams, however, the domain knowledge and expertise of the Data Resource teams cannot be replaced by technology. The original work to align data with the business processes will be key to successful value creation in the long term.

The Q4 2020 Forrester Wave document evaluated a variety data catalogs that provide machine learning techniques for managing data going forward. The top 10 providers were Alation, Alex Solutions, Collibra, data.world, erwin, Hitachi Vantara, IBM, Infogix, Informatica, and Io-Tahoe. The results presented in [Goe20] identified Alation, Collibra, Alex Solutions, and IBM as leaders.

Future data requirements are difficult to predict given the constantly changing technology solutions both in hardware and software. New sensors that stream location data might require different software solutions and significant post processing to leverage. As such, it would be useful to expand the existing survey or develop a new survey to predict future data requirements that may be beyond current in-house solutions for technology planning and development.

6 Summary of Recommended Actions

This final section presents a list of actions for moving forward. These are mostly near-term recommendations because a robust data collection process is foundational to supporting a data-informed organization and building a full data management process. It is a first step in a multi-step effort to achieve a shared data culture in which data assets are well integrated into the business process and fully leveraged.

We recommend three distinct and equally important efforts:

- Begin a focused outreach program that will communicate the importance of leveraging data as a resource and engages the business process owners in the development of this plan.
- Build a foundation of technology to support the data management processes identified in this plan.
- Create and resource the organization to support the development and maintenance of core data resources.

6.1 Outreach and Workforce Development

Maintain the digital approach that was initiated using the survey. Digital surveys are scalable and excellent for quickly processing the collected information into actionable solutions.

Assess the survey respondents to determine if an expanded survey group would provide improved insights into the current state. If not, launch an expanded survey to extend reach into the organization.

Continue the organizational outreach with additional surveys or workshops to gain a greater understanding of the full data management process. Include upcoming data collection plans into the topical areas, such as including pedestrian, bicycle and micro-mobility data in mobility data.

Develop a Communications Plan to assist in creating a shared data culture. Reinforce the existing work associated with roles and responsibilities of enterprise data stewards and metadata and data dictionary guidance. Communicate the need and value of a well-defined process for data collection, and provide Figures 4 through 7 as candidate processes.

Develop a Workforce Development Plan to increase data literacy across the organization.

Encourage the use of the metadata and data dictionary guidance described in Appendix A and B.

6.2 Data Management Tool Investments

Evaluate data cataloging tools for managing metadata, make recommendations for establishing the data cataloging tools, establishing a data registry, and improving organizational access to existing data.

Choose and implement a cross-organizational data cataloging tool. This could be a tool already used within Caltrans, a purchased tool, or an open-source tool.

Convene a future data requirements team to delineate the data diversity that will be expected with future technology improvements so that technology solutions (e.g., mechanisms for storing data such as data lakes and databases) can be predicted and planned.

Implement a data registry that will be used to assemble existing date resources.

Establish a document management system and data cataloging system and data registry for linking workflows, ontologies, metadata, and data dictionaries to raw data files. This task is dependent on the capabilities of the chosen data cataloging system.

Invest in automated technology for identifying data collection contracts or data purchases with feedback to the Data Stewards for review, as suggested in Figures 6 and 7.

Develop existing or new surveys to predict future data requirements that may be beyond current in-house solutions for technology planning and development.

6.3 Development and Maintenance of Core Data Resources

Begin with one domain-specific topic area related directly to Caltrans' strategic goals, and focus energy on building the ontology and metadata for that topic. A suggested topical area is bridges. A suggested data resource team has been identified based on the initial survey. The data resource team, with the aid of a facilitator, defines the ontology and taxonomy and works with the enterprise data steward to define the metadata according to the Caltrans metadata guidance. This group then adds this to the data catalog.

Assign a data steward for each ontology team. The data steward manages the domain ontology over time. The data steward for an ontology team may be a business data steward or an enterprise data steward. The data steward's responsibilities includes arranging meetings of the data resource team at a frequency that is most relevant for the data resource; facilitating the mapping of the onotology to metadata items that can be represented in a data catalog; ensuring the freshness of the representation by extending the ontology and associated metadata as the needs of the organization change; and attending cross-organizational ontology integration meetings.

Contract a third-party facilitator to attend the initial ontology meetings and provide leveling expertise for the discussions such that the discussions generate concise and comprehensive ontologies. The facilitator should review changes to the resource ontologies over time with the enterprise data steward; attend cross-organization ontology integration meetings.

Harvest existing Bridge data, consolidate and align it to metadata, and add it to a data catalog.

Review successes and difficulties with the processes, including the chosen data cataloging tool. Adjust process to Caltrans cultural norms.

Select more topical areas in the physical highway structures area and repeat these steps with additional Data Resources Teams, if new teams would be useful.

Expand efforts across all functional areas, creating at least five data resource teams focused on each topical asset group.

Implement an annual review and approval process for evaluating published data and its provenance to examine progress toward data quality goals.

References

[BNAoSM15] Transportation Research Board, Engineering National Academies of Sciences, and Medicine. *Data to Support Transportation Agency Business Needs: A Self-Assessment Guide*. The National Academies Press, Washington, DC, 2015.

[Goe20] Michele Goetz. The Forrester Wave ^{TM}: Machine Learning Data Catalogs, Q4 2020. 2020.

[KFM+20] S. Khan, N. Fournier, M. Mauch, A. D. Patire, and A. Skabardonis. Hybrid Data Implementation: Final Report for Task Number 3643. Technical report, UC Berkeley, California Partners for Advanced Transportation Technology, 2020.

[NIS] Niso.org.

[RDL19] Mike Rolling, Alan D. Duncan, and Valerie Logan. 10 Ways CDOs Can Succeed in Forging a Data-Driven Organization. Technical report, Gartner, 2019.

Appendices

A DRAFT Caltrans Guidance for Metadata

Table 9: Descriptive Core Metadata - Required

Column	Required	Description
DATASET_NAME	Yes	Commonly used name for the dataset. Should be the same as DATA_ASSET_NAME in the Data Catalog. Don't use CA or California in the name. Where there are multiple datasets with the same name, add the geographic extent as a suffix separated by an underscore. Examples: Culvert Inventory_District 9 or Culvert Inventory_San Luis Obispo County
TAGS	Yes	Comma-separated list of five or more descriptive keywords or phrases that users will search for to find this dataset. Include applicable ISO 19115 topic categories. See https://www.fgdc.gov/metadata/documents/MetadataQuickGuide.pdf.
DESCRIPTION	Yes	Brief description of what the dataset is about and its intended purpose. Identify what each record represents, including temporal and spatial granularity if applicable. For example: "condition ratings by 0.1-mile pavement section", "travel speeds by TMC for 5-minute intervals", "monthly revenues by organization code".
TOPIC	Yes	California Open Data Topic. Choose the most appropriate topic from https://data.ca.gov/.
DATA_DICTIONARY	Yes	Link to the data dictionary for this dataset.
DATA_DICTIONARY_TYPE	Optional	File format for the data dictionary if other than HTML (such as XML, CSV, PDF, DOC). GIS data dictionary metadata made available as part of a standard metadata file is in XML format.
DATA_LIFE_SPAN	Yes	Estimated useful life span of this dataset – end-of-life year (YYYY) or leave blank if data has no end of life.
CALTRANS_LINK	If Applicable	Link (URL) for access to data and/or useful information about the data.
DATA_STANDARD	Optional	URI or URL documenting a standardized specification to which the dataset conforms.
NOTES	No	Other notes about the dataset not covered by other metadata items. This item can be used to describe dataset limitations.

Table 10: Descriptive Additional GIS Metadata - If Applicable

Column	Required	Description
GIS_THEME	Yes	National Spatial Data Infrastructure (NSDI) Data Theme. See https://www.fgdc.gov/resources/whitepapers-reports/annual%20reports/2008/web-version/AppendixC.html
GIS_HORIZ_ACCURACY	If Applicable	Horizontal accuracy of collected data
GIS_VERT_ACCURACY	If Applicable	Vertical accuracy of collected data
GIS_COORDINATE_SYSTEM_EPSG	If Applicable	EPSG code from: https://spatialreference.org/ref/epsg/
GIS_VERT_DATUM_EPSG	If Applicable	EPSG code from: https://spatialreference.org/ref/epsg/

Table 11: **Administrative/Technical: Location and Temporal - If Applicable**

Column	Required	Description
METHODOLOGY	Yes	Description of how the data was collected, produced, and processed. Note any assumptions made or data issues to be aware of. If data is collected from multiple sources, list who those sources are which specific fields come from each source.
PUBLISHER_ORGANIZATION	Yes	"Caltrans", if published by Caltrans, or other agency, group, department, board, or commission that publishes the data resource. (FGDC definition is "party responsible for the dataset").
PLACE	If Applicable	Geographic locations characterized by the dataset, e.g., California, "Sacramento County".
TEMPORAL_COVERAGE_BEGIN	If Applicable	Start date for which the dataset is applicable (YYYY-MM-DD).
TEMPORAL_COVERAGE_END	If Applicable	End date for which the dataset is applicable (YYYY-MM-DD).

Table 12: Administrative and Technical Contact Metadata - Required

Column	Required	Description
CONTACT_ORGANIZATION	Yes	Contact person's organization (typically Caltrans, unless this is an external dataset maintained by Caltrans).
CONTACT_POSITION	Yes	Position of individual or name of organizational unit that can best answer questions about the data.
CONTACT_NAME	Yes	Name of individual that can best answer questions about the data. This is the Caltrans business data steward.
CONTACT_EMAIL	Yes	Contact's email address.

Table 13: Preservation Currency Information - Required

Column	Required	Description
FREQUENCY	Yes	Frequency with which changes and additions are made to the resource after the initial resource is completed. For example annually, quarterly, monthly, weekly, daily, hourly, none planned.
NEXT_UPDATE	If Applicable	Date when the next dataset update is planned to be published.
CREATION_DATE	If Applicable	Date that the dataset was originally created.
LAST_UPDATE	If Applicable	The last date that the dataset was updated (YYYY-MM-DD).
STATUS	Yes	Current status of this dataset – Complete, In Work or Planned.

Table 14: Administrative Access Rights - Required

Column	Required	Description
PUBLIC_ACCESS_LEVEL	Yes	Public – Data can be made publicly available without restrictions Restricted – Data can be made available with certain restrictions Non-Public – data cannot be made available to the public.
ACCESS_CONSTRAINTS	Yes	Place the boilerplate disclaimer language in this field. Also enter the description of reasons why data is restricted or non-public – e.g., inclusion of sensitive or confidential information. Note: This item should reflect designation of Personally Identifiable Information (PII), Payment Card Industry (PCI), Sensitive, and Confidential classifications for specific data elements in the data dictionary.
USE_CONSTRAINTS	Yes	Description of restrictions or legal prerequisites to using the dataset – e.g., required license or permissions. Enter the default license for Caltrans' datasets: "Creative Commons 4.0 Attribution", unless other license or permission requirements for this dataset have been established.

B DRAFT Caltrans Guidance for Creating a Data Dictionary

B.1 Purpose

A data dictionary provides information about the different data elements or fields in a dataset. It serves several purposes:

- Capture and preserve information about data elements to maintain and sustain a common understanding of the data
- Help future data users understand the meaning of each data element
- Articulate business requirements to be used by database developers to design or modify databases
- Communicate current database design characteristics to application and report developers

If a database already exists, some data dictionary items can be extracted from the database schema and used as a starting point for documenting the complete set of items listed below.

B.2 Data Dictionary Items

The basic set of data dictionary items to include is listed in Table 15.

Table 15: **Data Dictionary**

Column	Required	Description
SYSTEM_NAME	If Applicable	The name of the system or database containing this data element.
TABLE_NAME	Yes	The name of the table or dataset containing this data element. This matches the DATASET_NAME used for the dataset metadata.
FIELD_NAME	Yes	The name of the data element exactly as it appears in the table (or dataset column name).
FIELD_ALIAS	Yes	An easily understood title to be used for the data element in reports or data entry screens. Avoid abbreviations if possible.
FIELD_DESCRIPTION	Yes	A description of the data element's meaning and contents. Avoid descriptions that don't add new information beyond the Field Title.
FIELD_DESCRIPTION_AUTHORITY	If Applicable	Source for the field description if not Caltrans. Include name of organization and applicable document. For example, "FHWA HPMS Field Guide".
CONFIDENTIAL	Yes	Enter the word Confidential if this data element is exempt from disclosure under the provisions of the California Public Records Act (Government Code Sections 6250-6265) or has restrictions on disclosure in accordance with other applicable state or federal laws. Otherwise, enter NA.
SENSITIVE	Yes	Enter the word Sensitive if this data element requires a higher than normal assurance of accuracy and completeness. Typically, sensitive information includes records of state entity financial transactions and regulatory actions. Otherwise, enter NA.
PII	Yes	Enter PII if this data element identifies or describes an individual and must be protected from inappropriate access, use, or disclosure and must be made accessible to data subjects upon request. Examples include name and social security number, date of birth, driver's license/California identification card number, and financial account number. Otherwise, enter NA.
PCI	Yes	Enter PCI if this data element identifies or describes credit cardholder data and must be protected from inappropriate access, use, or disclosure and must be made accessible to data subjects upon request. Otherwise, enter NA.
FIELD_TYPE	Yes	Choose: • Text with Formatting • Date* • Object ID** • Short • DateTime • Boolean (Yes/No) • Long • Geometry** • GUID** • Float • Double • Binary or BLOB See Note 1.
FIELD_LENGTH	Yes	Number of characters or digits needed for this data element.
FIELD_PRECISION	If Applicable	If FIELD_TYPE is Float or Double, enter the number of digits to the right of the decimal point.
UNITS	If Applicable	Choose: Units of measurement (e.g., inches, miles, feet, tons, miles per hour, inches per mile), Currency (e.g., USD), or Percent. See Note 2.
DOMAIN_TYPE	Yes	Select one of the following options to describe the set of allowable values for this data element: • Range* – Domain can be specified by a min and max value (number or date) , • Enumerated – Domain can be specified by a list of values (which should be supplied or referenced) Ex: Good, Fair, Poor, • Coded* – Domain can be specified by a list of codes with associated meanings (which should be referenced) Ex: list of county codes and names, • Unrepresentable – Domain cannot be specified. E.g., free form text. See Note 3.
ALLOWABLE_MIN_VALUE	If Applicable	Minimum allowable value to accept. If ALLOWABLE_MIN_VALUE is used, the DOMAIN_TYPE of Range, and UNITS and ALLOWABLE_MAX_VALUE must be completed.
ALLOWABLE_MAX_VALUE	If Applicable	Maximum allowable value to accept. If ALLOWABLE_MAX_VALUE is used, DOMAIN_TYPE of Range must be used and UNITS and ALLOWABLE_MIN_VALUE must be completed.
USAGE_NOTES	No	Provide additional technical information needed to interpret and validate data. For example: (1) describe usage for "overloaded" data elements used for different purposes in different contexts; (2) document computations for calculated data elements; (3) describe the components of an intelligent ID which has multiple embedded meanings. Reference a related business rule providing further details.

Note 1: *For datasets created in Esri's GIS software, Date is used for DateTime. **Only applies to data stored in a GIS. Source for GIS data types: https://pro.arcgis.com/en/pro-app/latest/help/data/geodatabases/overview/arcgis-field-data-types.htm

Note 2: If UNITS is used, DOMAIN_TYPE must be set to Range and ALLOWABLE_MIN_VALUE and ALLOWABLE_MAX_VALUE must be completed.

Note 3: If the DOMAIN_TYPE of Range is used, UNITS, ALLOWABLE_MIN_VALUE and ALLOWABLE_MAX_VALUE must be completed. Items marked with an asterisk are the only domain types supported by Esri for us in GIS datasets. Source for GIS: https://pro.arcgis.com/en/pro-app/latest/help/data/geodatabases/overview/an-overview-of-attribute-domains.htm#Domain%20type

B.2.1 Mappings to Caltrans ISO 19139 Geospatial Metadata Elements

Mappings to Caltrans ISO 19139 geospatial metadata elements and the California Open Data elements are provided in Table 16. Items marked with an asterisk are managed from within the GIS software and should not be edited using a metadata editor.

Table 16: **Mappings to Caltrans ISO 19139**

Column	CA Open Data	ISO
SYSTEM_NAME		Resource > Lineage > Data Source > Source Citation > Titles > Title
TABLE_NAME*		Resource > Fields > Entity and Attribute Information > Details > Label
FIELD_NAME*	Field Name	Resource > Fields > Entity and Attribute Information > Details > Attribute > Label
FIELD_ALIAS*		Field Title Resource > Fields > Entity and Attribute Information > Details > Attribute > Alias
FIELD_DESCRIPTION		Description Resource > Fields > Entity and Attribute Information > Details > Attribute > Definition
FIELD_DESCRIPTION_ AUTHORITY		Resource > Fields > Entity and Attribute Information > Details > Attribute > Definition Source
CONFIDENTIAL		Resource > Constraints > Security Constraints > Classification = Confidential
SENSITIVE		Resource > Constraints > Security Constraints > Classification = Restricted Resource > Constraints > Security Constraints > User Note = Sensitive Where multiple entries exist, separate by comma (e.g., Sensitive, PII, PCI)
PII		Resource > Constraints > Security Constraints > Classification = Restricted Resource > Constraints > Security Constraints > User Note = PII Where multiple entries exist, separate by comma (e.g., Sensitive, PII, PCI)
PCI		Resource > Constraints > Security Constraints > Classification = Restricted Resource > Constraints > Security Constraints > User Note = PCI Where multiple entries exist, separate by comma (e.g., Sensitive, PII, PCI)
FIELD_TYPE*	Data Type	Resource > Fields > Entity and Attribute Information > Details > Attribute > Type
FIELD_LENGTH*		Resource > Fields > Entity and Attribute Information > Details > Attribute > Width
FIELD_PRECISION*		Resource > Fields > Entity and Attribute Information > Details > Attribute > Precision
UNITS		Resource > Fields > Entity and Attribute Information > Details > Attribute > Range Domain > Units
DOMAIN_TYPE		Select one of the following options to describe the set of allowable values for this data element: **Range:** Resource > Fields > Entity and Attribute Information > Details > Attribute > Range Domain **Codeset:** Resource > Fields > Entity and Attribute Information > Details > Attribute > Codeset Domain > Name Resource > Fields > Entity and Attribute Information > Details > Attribute > Codeset Domain > Source **Enumerated:** Resource > Fields > Entity and Attribute Information > Details > Attribute > Enumerated Domain > Definition Source
ALLOWABLE_ MIN_VALUE		Resource > Fields > Entity and Attribute Information > Details > Attribute > Range > Domain > Minimum
ALLOWABLE_ MAX_VALUE		Resource > Fields > Entity and Attribute Information > Details > Attribute > Range Domain > Maximum
USAGE_NOTES		Resource > Fields > Overview Description > Summary

C ENTERPRISE DATA AND GEOSPATIAL GOVERNANCE ROLES AND RESPONSIBILITIES

C.1 Enterprise Data Steward

An individual with accountability for corporate and shared data within an established domain or program area (e.g., capital projects, roadway, traffic operations, design). The enterprise data stewards are members of the Enterprise Data Steward Committee that reports to the Enterprise Data and Geospatial Governance Board (Board). Their role is to represent the interests of the enterprise-wide use of data within their domain or program area. Primary responsibilities include:

- Work to achieve the mission, vision and core data principles adopted by the Enterprise Data and Geospatial Governance Board.
- Support adoption and implementation of data governance policies, practices, and guidelines.
- Identify business data stewards, and manage the list of individuals assigned to those roles.
- Attend enterprise data steward committee meetings and work to achieve its goals.
- Communicate regularly with business data stewards to make sure they are familiar with data governance policies, practices, guidelines, and resources.
- Inform business data stewards of proposed and implemented changes to systems within their business area that may impact them.
- Obtain feedback from business data stewards regarding enterprise data and geospatial governance related needs.
- Assess business data steward performance and arrange for mentoring and training as needed to help them to successfully carry out their responsibilities.

Additional Responsibilities (Time-Permitting)

- Develop strategies for data and geospatial improvement, with an enterprise-wide perspective.
- Advocate for and lead improvements to how data is defined, collected, maintained, classified, documented, harmonized, shared, and used.
- Resolve cross-business unit issues related to sharing or integrating data.
- Obtain agreement on data standards for adoption by the Board.
- Be aware of and track activities in progress and communicate with all enterprise data stewards about opportunities for synergy or potential duplication.
- Actively collaborate across Caltrans on current and future data governance activities.
- Collect and report metrics regarding adoption of enterprise data governance practices.
- Review information technology projects, as required, to provide feedback regarding adherence to data governance and data management policies, practices, and guidance.
- Conduct activities to support data governance such as regular surveys of corporate datasets and data products (e.g., reports, visualizations, web maps) for both public and internal audiences.

C.2 District Enterprise Data Governance Liaison

An individual with responsibility for coordinating communication related to data and geospatial governance between their district and the Enterprise Data and Geospatial Governance Program staff, governing bodies and data governance working groups. Their role is to represent the interests of their district as it relates to the enterprise-wide implementation of data governance practices, to coordinate requests for information and feedback, and communicate with their executive management team and staff on the status of the enterprise data governance program.

Primary responsibilities include:

- Work to achieve the mission, vision and core data principles adopted by the Board.
- Support adoption and implementation of data governance policies, practices, and guidelines.
- Represent district interests on data governance working groups.
- Coordinate district responses to requests for information and document reviews.
- Share information about enterprise data and geospatial governance efforts with district staff.

C.3 Business Data Steward

An individual with accountability for data within a defined business area or scope. The business data stewards are located in Headquarters and District offices and provide assistance and support to the enterprise data steward assigned to their business area for matters related to data governance. Some business data stewards might also be enterprise data stewards.

Primary responsibilities include:

- Work to achieve the mission, vision and core data principles adopted by the Board.
- Serve as the primary authority (subject matter expert) for data within a particular business area orscope – understand meaning, derivation, quality requirements, and uses.
- Make decisions/sign off on data definitions, business rules, standards, and data management processes.
- Work to enhance data quality and value to the organization:
  - Establish and maintain relationships with data consumers and other stakeholders.
  - Ensure data is entered or made available for reporting in a timely fashion.
  - Ensure that quality management processes are established and applied including:
    - Creation of data quality standards consistent with business needs
    - Definition and documentation of business rules for quality checks (valid ranges, cross-field checks)
    - Tracking and remediation of data quality defects
    - Communication about data limitations to data consumers
  - Identify and advocate for actions to improve data quality and value.
  - Provide business expertise and assess business impact(s) for proposed data improvement initiatives.
- Attend regular meetings called by the Enterprise Data Stewards.

Additional Responsibilities (Time-Permitting):

- Coordinate development and maintenance of documentation (metadata) about databases, datasets, standard reports and other data assets. This documentation may include:
  - A sound description of the data for business (non-IT) data consumers covering sources, derivation, and intended uses
  - Enterprise data glossary entries representing key data entities and attributes
  - Data element definitions (description, type, other data dictionary items)
  - Data flow and lineage diagrams mapping movement of data from original sources to repositories used for analysis, visualization, query and reporting
  - Work flow diagrams indicating steps in the data production process
  - Ensure that data is classified, shared and used appropriately: Classify data elements for confidentiality and security requirements based on established policies and guidance.

- Ensure that data is shared based on established policies.
- Assist in designation of authoritative data sources and systems of record.
- Understand and follow security protocols related to data.
- Support implementation/adoption of data governance and management policies, practices and guidelines:
- Measure and report compliance with standards/processes; data quality, progress in implementing improvements, etc.
- Actively collaborate across Caltrans on current and future data governance activities.
- Ensure systems, datasets, and data products are documented, governed, and managed by staff according to documented best practices.
- Attend knowledge sharing sessions to learn and use new concepts, tools and processes.
- Actively participate on data governance related activities such as regular surveys of corporate datasets and data products (e.g., reports, visualizations, web maps) for both public and internal audiences.

C.4 Data Custodian

An individual with physical custody of the data. Responsible for the technical infrastructure (hardware, software, networking), database administration and backup of a particular set of data. These individuals are typically (but not always) information technology staff.

- Work to achieve the mission, vision and core data principles adopted by the Board.
- Perform or assist with data loading and transfers.
- Serve as technical resource for data integration efforts.
- Perform database administrator functions: capacity planning, hardware and software installation, configuration, database design, data and software migration, performance monitoring, security, troubleshooting, data backup and data recovery.

C.5 OTHER DATA-RELATED RESPONSIBILITIES

While data stewards have accountability for data, operational responsibilities for production, documentation, and sharing of data will typically be distributed across multiple additional staff. Key data-related responsibilities are listed below:

- Data Quality Management
  - Produce or support preparation of a data quality management plan.
  - Propose and gain agreement on data quality standards.
  - Create business rules for quality checks and/or data validation.
  - Develop and/or manage data processes for defect tracking and reporting.
  - Support development of data cleansing processes and review results of those processes to ensure that they are functioning as intended.
- Data Collection and Entry
  - Ensure that data entered or loaded into agency systems adheres to established business rules for timeliness and accuracy.
  - Ensure that data entered into agency systems is consistent with field definitions and other standards assigned to the data items by the responsible business data steward.
  - Keep the business data steward informed about data quality issues and potential solutions to these issues.

  - Adhere to security protocols for managing and protecting sensitive or confidential data.

- Data Documentation
  - Create and maintain descriptive documentation for systems, databases, datasets, reports, and other data assets.
  - Create and maintain data dictionary information.
  - Create and maintain data flow and lineage diagrams.
  - Create and maintain workflow diagrams.
- Data Governance Support
  - Support business data stewards on data governance activities.
  - Serve as the technical functional expert responsible for supporting and implementing data. governance and best practices for a particular set of data assets.
- Data Sharing and Reporting
  - Provide technical expertise, and assess the technical impact of proposed data initiatives.
  - Create business requirements/specifications for datamarts, reports, maps and query tools.
  - Create reports, maps, and other data visualizations.
- Data Integration and Application Development (typically an IT function)
  - Create views of the data tailored to specific audiences or business needs.
  - Write and test Extract-Transform-Load (ETL) scripts.
  - Create or configure applications for data access.
  - Serve as a technical resource to data consumers or other stakeholders seeking to obtain or integrate data.

D Survey Questions

Introduction

Why do we need a questionnaire/survey? The Caltrans Enterprise Data and Geospatial Governance Program is planning a unified and coordinated statewide approach to field asset and mobility data collection/procurement. The Program would like to capitalize on the efficiencies gained through a "collect data once, use numerous times" strategy. A key question being asked is: "What are the existing barriers that reduce the ability for roadside asset data to be discovered, accessed, used and shared across the organization?" The focus of this survey is to assess the current state of data collection/purchasing and sharing both within the organization, as well as delivering data to external organizations.

Who will receive the questionnaire/survey? Caltrans staff that regularly provide or consume data as part of their business process are invited to participate in the questionnaire/survey.

What is the scope of assets and data covered by this questionnaire/survey? This survey intends to elicit information about the lifecycles of physical assets and the data lifecycles that are crucial for supporting their management. This includes primary physical structures (e.g., pavement, bridges), transportation management systems (e.g.,signals, ramp meters, cameras, signs), and supplementary asset classes (e.g.,WIM, bicycle and pedestrian facilities). It also includes data about mobility, (e.g., GPS traces, inductive loop sensors).

How will the results of the study be used? The information derived from this study will inform a strategic plan for a unified and coordinated statewide approach to the collection/purchase of data to support Caltrans' business needs.

Who is conducting the study? In coordination with the Enterprise Data and Geospatial Governance Program, researchers at the University of California at Berkeley and San Diego with expertise in data management, metadata design, and transportation systems created this study.

What are the next steps following the questionnaire/survey? If the respondent is receptive to further discussions, a phone call or in-person interview may be scheduled. All inputs from the questionnaire and interviews will be categorized and consolidated into a summary report. It should be noted that not all recommendations or suggestions from the survey will or can be fully implemented.

We estimate the survey will take about 15 minutes to complete. In each question, there may be multiple levels of information that are relevant to the question. The questions asked during the survey may be personalized depending on answers to prior questions. If you need further explanation or clarification regarding the survey questions or process, please email `<janemacfarlane@berkeley.edu>`.

We look forward to building a data management culture that improves the effectiveness of our business processes.

Enterprise Data and Geospatial Governance Program

Basic Demographic Info

Q 2.1: Please provide your contact information

- Email
- Name
- Title

Q 2.2: Please provide your program and division information

- Select Headquarters or District

- Program / District Number
- Division / Unit

Matrix Questions

Q 3.1: Please select the types of asset or mobility data that you use. Select all that apply.

- Physical Structures (e.g., pavement, bridges, signs)
- Transportation Management Systems (e.g., cameras, signals, ramp meters)
- Supplementary Assets (e.g.,WIM, bike/ped, ADA)
- Mobility Data (e.g., GPS traces, flow, speed, freight)
- Planning Data (e.g., Census, safety, transit)

Q 3.2: Physical Highway Structures

Please indicate your experience with the asset data you need to perform your work.

If the data are needed, please rate the level of digital accessibility using the following scale:

- ☐ Don't know
- ☐ Data do not exist
- ☐ Findable via co-worker (not digitally findable on a centralized system)
- ☐ Digitally Findable (by me on a centralized system) without contacting a data steward
- ☐ Digitally Findable. Accessible by contacting the data source or steward (no direct download)
- ☐ Digitally Findable and Digitally Accessible by direct download from a centralized system
- ☐ Digitally Findable, Accessible, and Usable: Usable immediately in a convenient format without reformating, data wrangling, or re-entering the data into a separate software tool

- Pavement
- Pavement Markings
- Barriers
- Sound Walls
- Bridges
- Culverts
- Guardrail
- Crash Cushions
- Static Signs
- Changeable Message Signs
- Highway Lighting

Q 3.3: Transportation Management Systems

Please indicate your experience with the asset data you need to perform your work.

If the data are needed, please rate the level of digital accessibility using the following scale:

- ☐ Don't know
- ☐ Data do not exist
- ☐ Findable via co-worker (not digitally findable on a centralized system)
- ☐ Digitally Findable (by me on a centralized system) without contacting a data steward
- ☐ Digitally Findable. Accessible by contacting the data source or steward (no direct download)
- ☐ Digitally Findable and Digitally Accessible by direct download from a centralized system
- ☐ Digitally Findable, Accessible, and Usable: Usable immediately in a convenient format without reformating, data wrangling, or re-entering the data into a separate software tool

- Closed circuit cameras
- Traffic census stations
- Roadway weather information systems
- Traffic monitoring detection stations
- Highway advisory radios
- Freeway ramp meters
- Traffic signals
- Control cabinets

Q 3.4: Supplementary Assets

Please indicate your experience with the asset data you need to perform your work.

If the data are needed, please rate the level of digital accessibility using the following scale:

- ☐ Don't know
- ☐ Data do not exist
- ☐ Findable via co-worker (not digitally findable on a centralized system)
- ☐ Digitally Findable (by me on a centralized system) without contacting a data steward
- ☐ Digitally Findable. Accessible by contacting the data source or steward (no direct download)
- ☐ Digitally Findable and Digitally Accessible by direct download from a centralized system
- ☐ Digitally Findable, Accessible, and Usable: Usable immediately in a convenient format without reformating, data wrangling, or re-entering the data into a separate software tool

- Vegetation
- Irrigation / Drainage
- ADA Infrastructure
- Roadside Rest Facilities
- Park & Ride Infrastructure
- Weigh-In-Motion Scales
- Sidewalks
- Bicycle / Pedestrian Facilities

Q 3.5: Mobility Data

Please indicate your experience with the data you need to perform your work.

If the data are needed, please rate the level of digital accessibility using the following scale:

- ☐ Don't know
- ☐ Data do not exist
- ☐ Findable via co-worker (not digitally findable on a centralized system)
- ☐ Digitally Findable (by me on a centralized system) without contacting a data steward
- ☐ Digitally Findable. Accessible by contacting the data source or steward (no direct download)
- ☐ Digitally Findable and Digitally Accessible by direct download from a centralized system
- ☐ Digitally Findable, Accessible, and Usable: Usable immediately in a convenient format without reformating, data wrangling, or re-entering the data into a separate software tool

- Flow / Volume
- Speed
- VHT / VMT
- Trip and Mode Choice
- Origin-Destination
- Intersection Turning Counts
- GPS Traces
- Freight / Goods Movement

Q 3.6: Planning Data

Please indicate your experience with the data you need to perform your work.

If the data are needed, please rate the level of digital accessibility using the following scale:

- ☐ Don't know
- ☐ Data do not exist
- ☐ Findable via co-worker (not digitally findable on a centralized system)
- ☐ Digitally Findable (by me on a centralized system) without contacting a data steward
- ☐ Digitally Findable. Accessible by contacting the data source or steward (no direct download)
- ☐ Digitally Findable and Digitally Accessible by direct download from a centralized system
- ☐ Digitally Findable, Accessible, and Usable: Usable immediately in a convenient format without reformating, data wrangling, or re-entering the data into a separate software tool

- Census Tracts
- Land Use / Parcels
- Safety Data / Injuries
- Transit Data
- Other GIS

Q 3.7: Are there any additional data types you need or use that were not mentioned above?

Q 3.8: What additional kinds of data will be needed in the future?

Role as a Data Producer

Q 4.1: Do you frequently provide data content to other programs/divisions or external providers?

- yes
- no

Q 4.2: In the following questions, please select two Caltrans programs/divisions who rely on you for vital data content

Q 4.3: To whom (which programs/divisions) do you provide data content?

- Select Headquarters or District
- Program / District Number
- Division / Unit

Q 4.4: To whom (which programs/divisions) do you provide data content?

- Select Headquarters or District
- Program / District Number
- Division / Unit

Q 4.5: If your most relevant data consumers do not appear in the drop-down lists above, please write in your response below:

Q 4.6: In which formats do you provide data to others in the organization? Please click all that apply.

- spreadsheet (xlsx, csv)
- text (txt, yml)
- image format (jpg, png)
- webpage (html)
- document such as Word or PDF
- sql query from database
- spatial formats (shapefiles, geojson)
- other, please fill in

Standards

Q 5.1: Do you have a data dictionary to relate (join) multiple datasets across stakeholders / programs / divisions (e.g., are the same units / conventions used across data sets)?

- yes
- no

Q 5.2: If no, why not?

Q 5.3: Do you follow standards when generating metadata for your data assets?

- yes
- no

Q 5.4: If yes, what metadata standards do you use? Please select all that apply or fill in.

- ISO 19115/19139
- Dublin Core
- GTFS
- other, please fill in

Q 5.5: Do you have automatic tools or templates to help generate metadata for you?

- yes
- no

Q 5.6: If yes, what tools or templates do you use to generate metadata?

Role as a Data Consumer

Q 6.1: Do you frequently obtain (consume) data content from other programs/divisions or external providers?

- yes
- no

Q 6.2: In the following questions, please select two Caltrans programs/divisions from whom you obtain the most vital data content

Q 6.3: From whom (which programs/divisions) do you obtain data content?

- Select Headquarters or District
- Program / District Number
- Division / Unit

Q 6.4: From whom (which programs/divisions) do you obtain data content?

- Select Headquarters or District
- Program / District Number
- Division / Unit

Q 6.5: If your most relevant data content suppliers do not appear in the drop-down lists above, please write in your response below:

Q 6.6: In which formats do you obtain data from others in the organization? Please click all that apply.

- spreadsheet (xlsx, csv)
- text (txt, yml)
- image format (jpg, png)
- webpage (html)
- document such as Word or PDF
- sql query from database
- spatial formats (shapefiles, geojson)
- other, please fill in

Q 6.7: Do you know what data conditioning or analytics have been applied to generate the data you obtain from others in the organization?

- yes
- no
- does not apply

Structural Metadata

Q 7.1: What are the highest value datasets that you or your division generates?

Q 7.2: How do you discover and find information that you need? (click all that apply)

- Ask a data steward
- Consult a data catalog
- External web search
- onramp / intranet
- Other, please fill in

Administrative Metadata

Q 8.1: Do you have a mechanism to confirm you have the correct version of the data?

- yes
- no

    Q 8.2: If yes, what mechanism do you use to confirm you have the correct version of the data?

Q 8.3: Do you have a mechanism to confirm that you are using the authoritative copy of the data?

- yes
- no

    Q 8.4: If yes, what mechanism do you use to confirm you have the authoritative copy of the data?

Q 8.5: Do you have a mechanism to know the data is current?

- yes
- no

    Q 8.6: If yes, what mechanism do you use to know the data is current?

Q 8.7: Do you have a mechanism to know if the data is valid and reliable?

- yes
- no

    Q 8.8: If yes, what mechanism do you use to know the data is valid and reliable?

Licensing and Access

Q 9.1: Do you have to manage licensing or access rights to obtain the data you need?

- yes
- no

    Q 9.2: If yes, what kinds of data require this?

    Q 9.3: If yes, how do you manage access to the data?

End of Survey

We thank you for your time spent taking this survey.

Your response has been recorded.

E Survey Results

Caltrans Strategic Plan for Data Collection

Survey Version:2021-11-03 21:03:29 UTC

Survey

Responders

There were a total of 49 responses.

The composition was: 31 from Headquarters and 18 from the Districts.

The composition of the respondents was: at least one representative from Districts 1,2,3,4,6,8,9,10,11; None from Districts 5 and 7; 17 from HQ Project Delivery!; 5 from HQ Planning and Modal, 4 from HQ Maintenance and Operations, 4 from Finance, and 1 from Administration.

Structure of this Report

Sections include:

- The data itself - how needed and accessible they are to the organization and how much users know about the currency and completeness of the data,
- Datasets that are generated by the organizations,
- Formats used to exchange the data,
- Metadata - existance of data dictionaries and the need for data dictionaries for specific areas,
- Provenance and quality of the datasets,
- Data Exchange within the organization,
- Discovery of data within the organization,
- Licensing and access management, and finally
- Future data needs

Importance of Asset and Mobility Data in Daily Business Processes

This area focuses on the type of asset or mobility data that the responders use as a part of their business processes.

Data were divided into several broad categories: physical structures (e.g., pavement, bridges, signs); transportation management systems (e.g., cameras, signals, ramp meters); supplementary assets (e.g.WIM, bike/ped, ADA); mobility data (e.g. GPS traces, flow, speed, freight); and planning data (e.g. census, safety, transit). For each category, respondents were asked to rate several attributes about the data and the ease with which the data can be discovered, accessed, and used. The three key attributes were: Currency (whether the data are known to be up-to-date); Necessity (whether the data are required fulfill a job function); and Completeness (whether the data are comprehensive and exhaustive).

The bar chart below reflects the data usage for these data types, physical structures being the highest but all appearing as important to the organization.

For each of these key asset groups the respondents were asked about whether the data they found about these assets were **Up-to-Date**, if the data was **Needed** for their work and finally, if the data was **Complete and Comprehensive**.

They were then asked to: Indicate your experience with the asset data you need to perform your work. If the data are needed, please rate the level of digital accessibility using the following scale:

- Don't know
- Data do not exist
- Findable via co-worker (not digitally findable on a centralized system)
- Digitally Findable (by me on a centralized system) without contacting a data steward
- Digitally Findable. Accessible by contacting the data source or steward (no direct download)
- Digitally Findable and Digitally Accessible by direct download from a centralized system
- Digitally Findable, Accessible, and Usable: Usable immediately in a convenient format without reformating, data wrangling, or re-entering the data into a separate software tool

Physical Structures

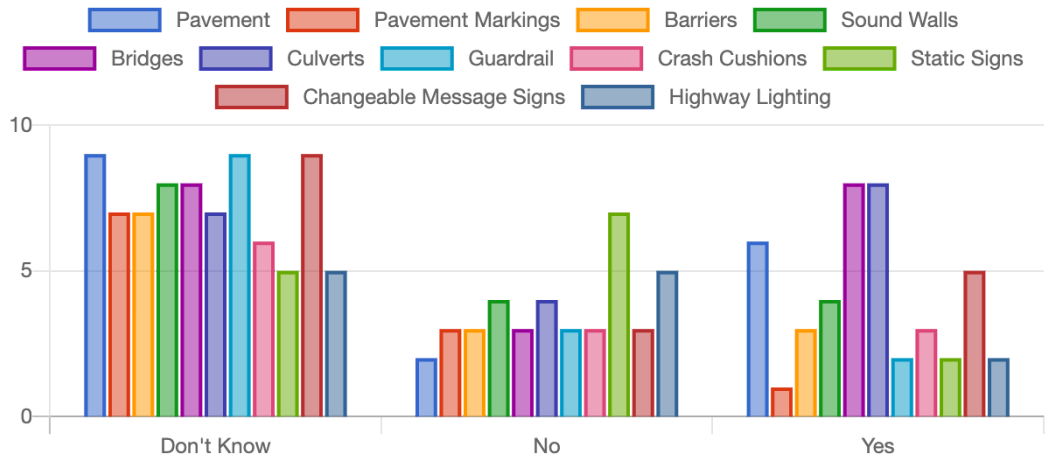
Physical structures considered where: Pavement, Pavement Markings, Barriers, Sound Walls, Bridges, Culverts, GuardRails, Crash Cushions, Static Signs, Changeable Message Signs and Highway Lighting. Respondents were asked to indicate their experience with the data that they need to perform their work.

Summary:

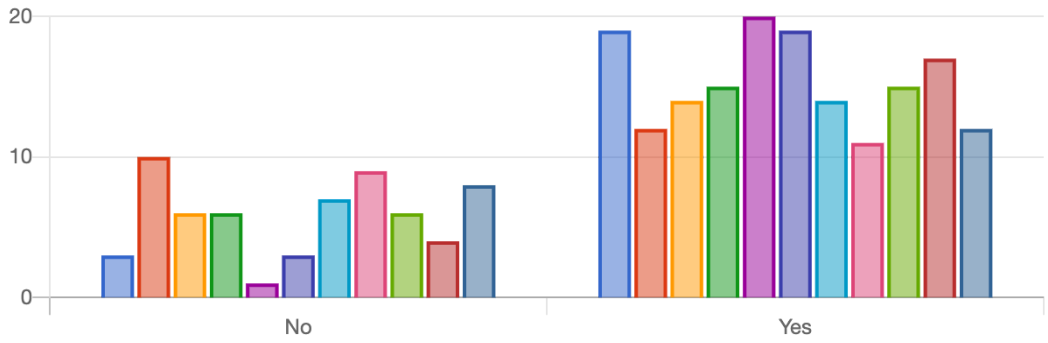
- **Currency**
  - **Many respondents did not know if the data were up to date**
  - **Static signs were most often described as out of date**
  - **Bridge and culvert data were described by several as current**
- **Necessity:**
  - **Most agreed that physical structure asset data is needed**
- **Completeness**
  - **Pavement, bridge, and culvert data were considered the most complete and comprehensive**
  - **Static signs were considered the least complete**
- **Discovery, access, and use**
  - **Respondents had very different opinions about the data**
  - **Bridge data was commonly regarded as findable and accessible by download**
  - **Existence of asset data about guardrail, pavement markings, barriers, and crash cushions was commonly unknown or in dispute**

Pavement, bridges and culverts and perhaps changeable message signs appear to be the best managed datasets. Awareness of datasets is low.

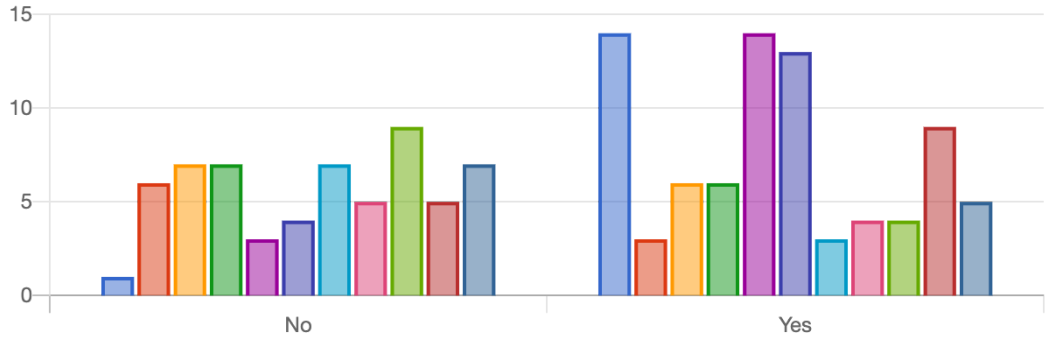
Number of Unique Respondents:23



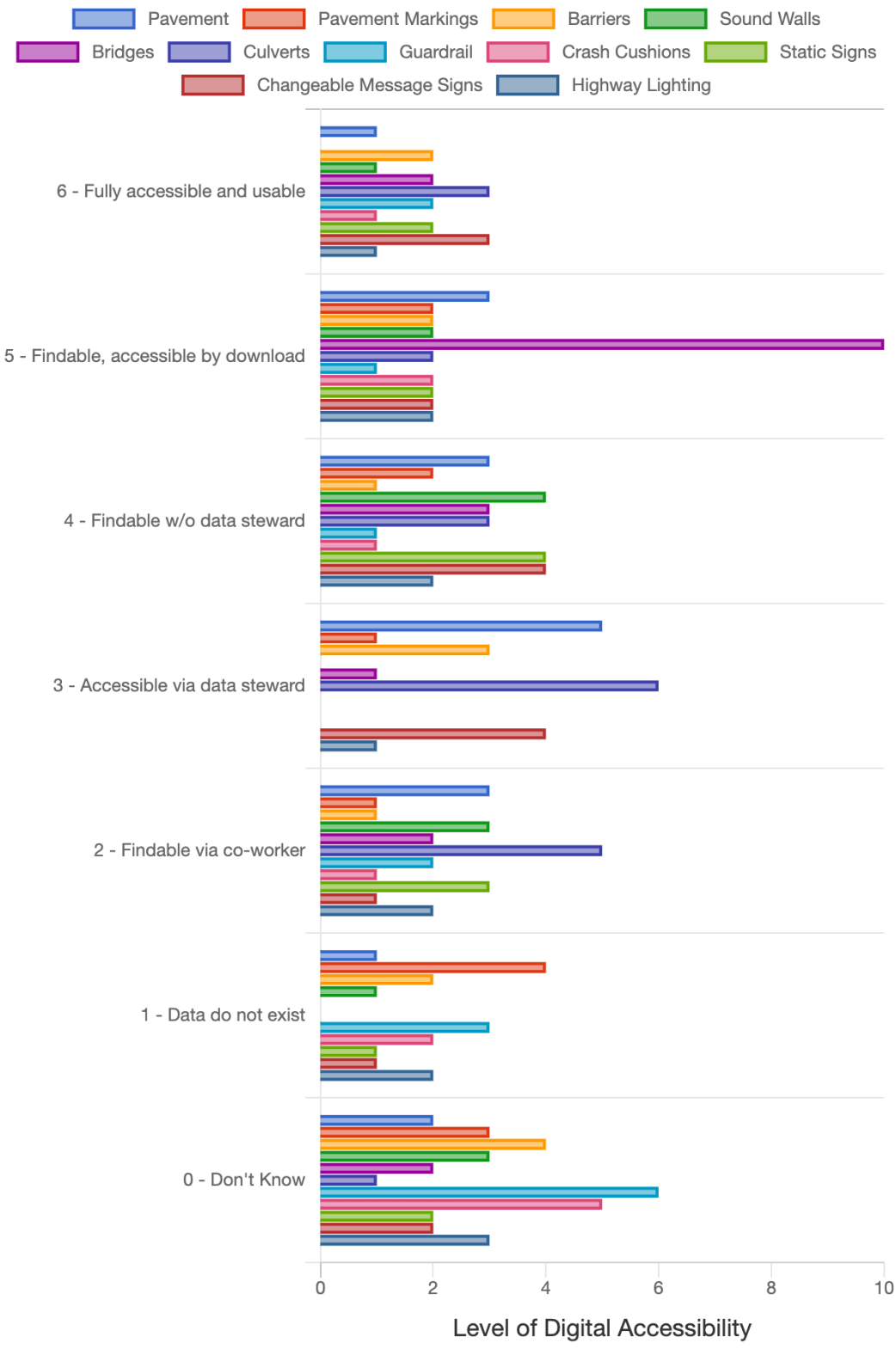
Up to Date



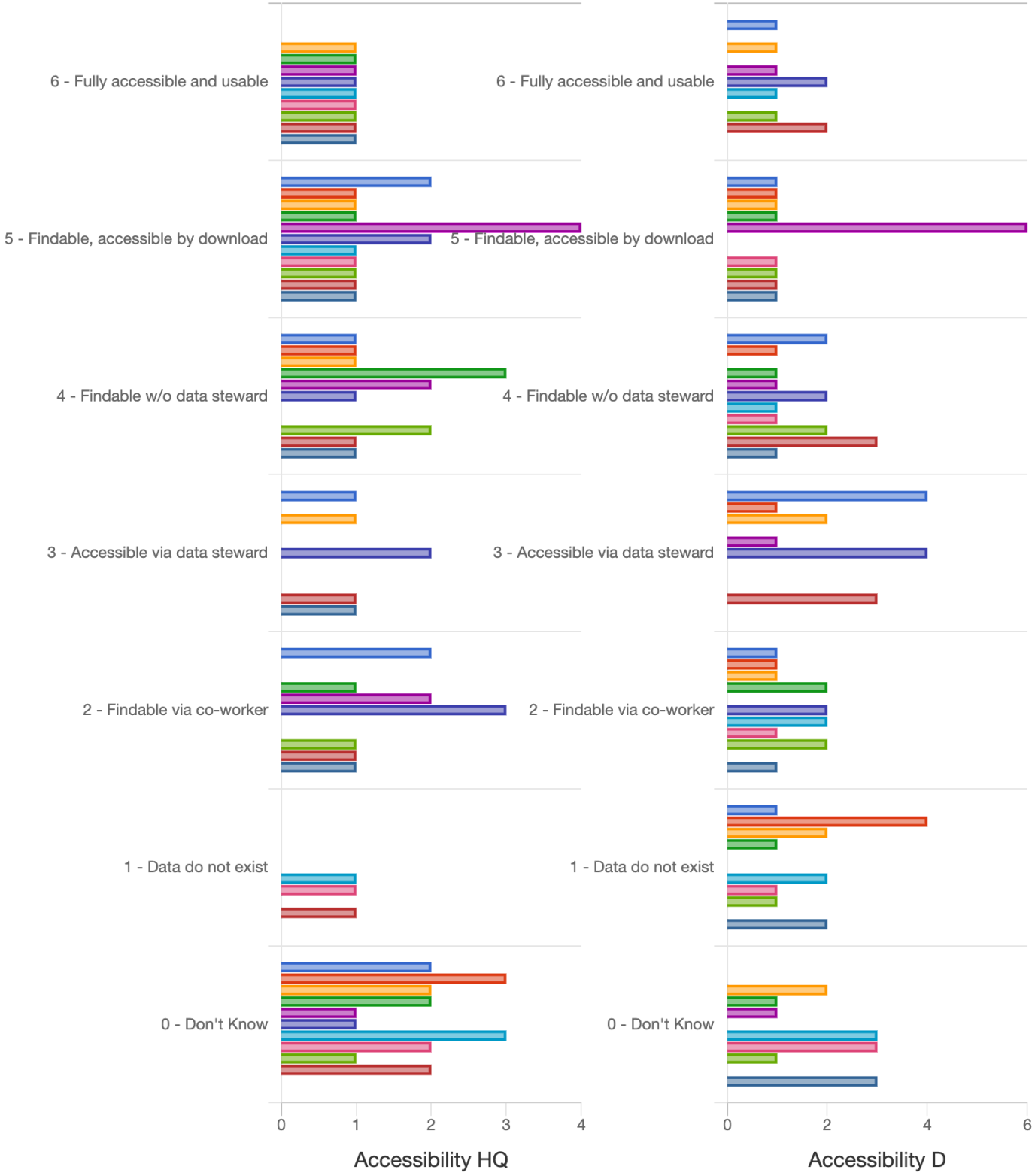
Needed



Complete and Comprehensive


Pavement
Pavement Markings
Barriers
Sound Walls
Bridges
Culverts
Guardrail
Crash Cushions
Static Signs
Changeable Message Signs
Highway Lighting
6 - Fully accessible and usable
5 - Findable, accessible by download
4 - Findable w/o data steward
3 - Accessible via data steward
2 - Findable via co-worker
1 - Data do not exist
0 - Don't Know
0
2
4
6
8
10
Level of Digital Accessibility

Headquarters vs District



Transportation Management Systems

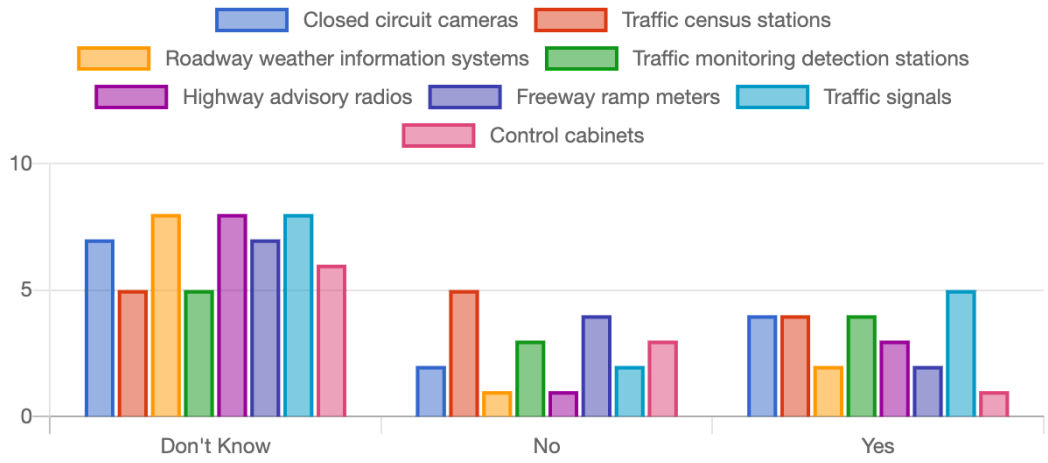
Transportation management systems considered were: closed circuit cameras, traffic census stations, roadway weather information systems, traffic monitoring detection systems, highway advisory radios, freeway ramp meters, traffic signals, and control cabinets.Respondents were asked to indicate their experience with the data that they need to perform their work.

Summary:

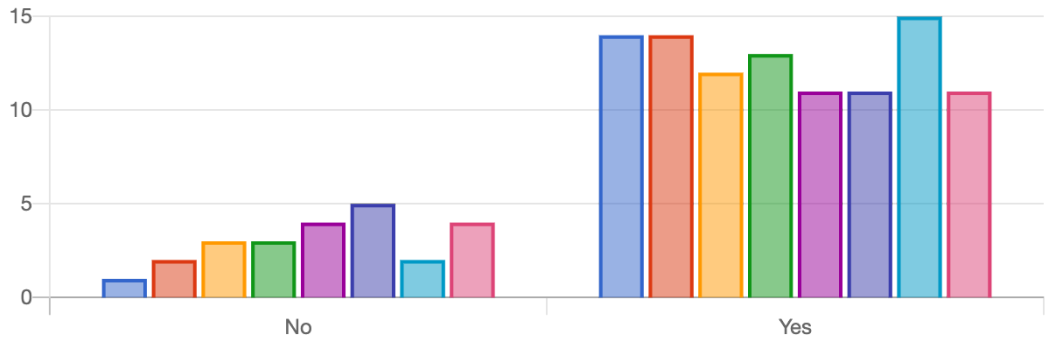
- **Currency**
  - **Many respondents did not know if the data were up to date.**
- **Necessity:**
  - **Most agreed that TMS data is needed.**
- **Completeness**
  - **Many were comfortable with the completeness and comprehensiveness of the data.**
- **Discovery, access, and use**
  - **Many respondents reported that the data were accessible through a data steward.**
  - **Very few respondents reported that the data were accessible and useable by themselves.**

Awareness of the data was low. Rely on data steward.

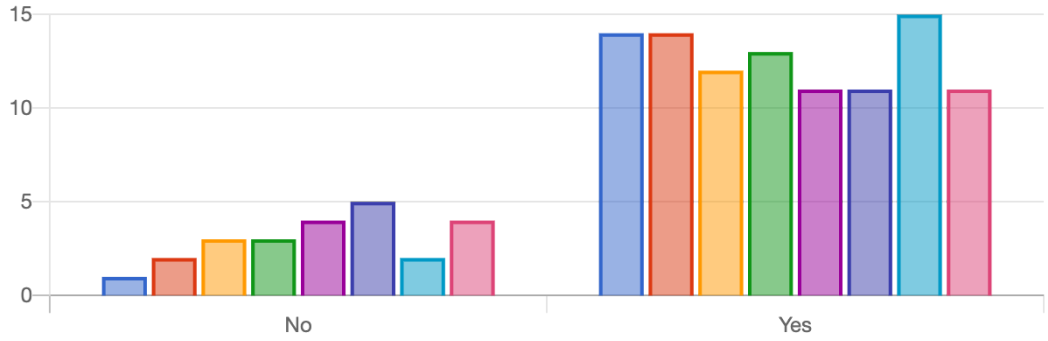
Number of Unique Respondents:16



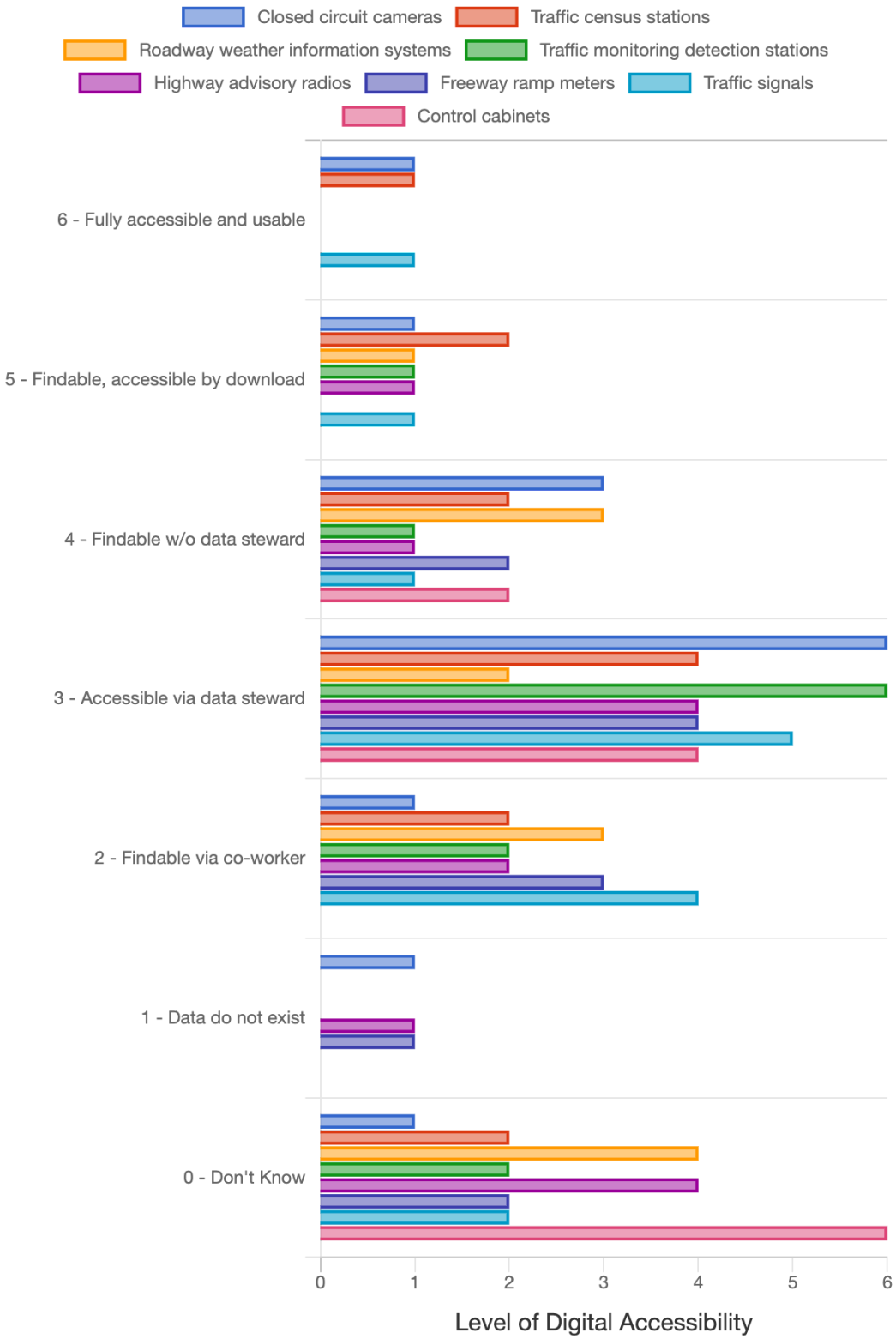
Up to Date



Needed



Complete and Comprehensive


Closed circuit cameras
Traffic census stations
Roadway weather information systems
Traffic monitoring detection stations
Highway advisory radios
Freeway ramp meters
Traffic signals
Control cabinets
6 - Fully accessible and usable
5 - Findable, accessible by download
4 - Findable w/o data steward
3 - Accessible via data steward
2 - Findable via co-worker
1 - Data do not exist
0 - Don't Know
0
1
2
3
4
5
6
Level of Digital Accessibility

Headquarters vs District

Supplementary Assets

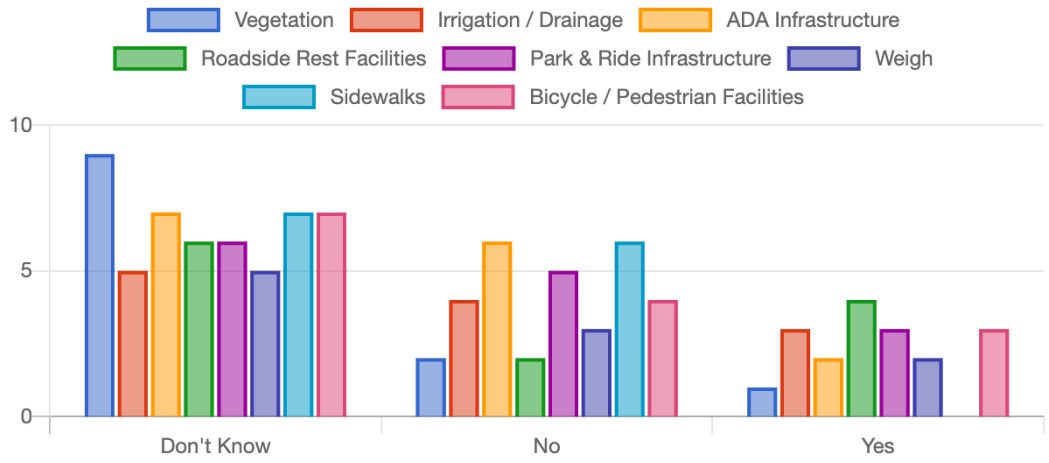
Supplementary assets considered were: vegetation, irrigation/drainage, ADA infrastructure, roadside rest facilities, park and ride infrastructure, weight stations, sidewalks and bicycle/pedestrain facilities.Respondents were asked to indicate their experience with the data that they need to perform their work.

Summary:

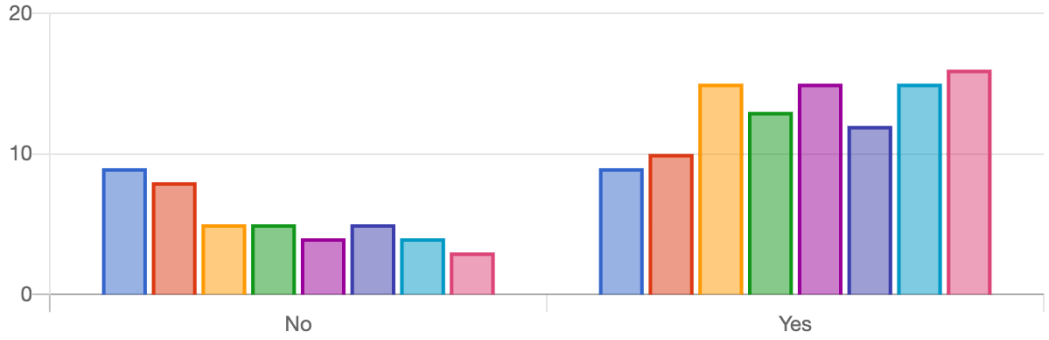
- **Currency**
  - **Many respondents did not know if the data were up to date.**
- **Necessity:**
  - **Most agreed that supplementary data is needed.**
- **Completeness**
  - **Roadside rest facility data were considered the most complete in this category.**
  - **Data on sidewalks and bike/ped facilities were considered the most incomplete.**
- **Discovery, access, and use**
  - **Respondents had very different opinions about the data.**
  - **Some using the data easily and others relying on data stewards and co-workers.**
  - **Existence and accessibility of vegetation and ADA infrastructure information was unclear.**

Awareness of the data was low.

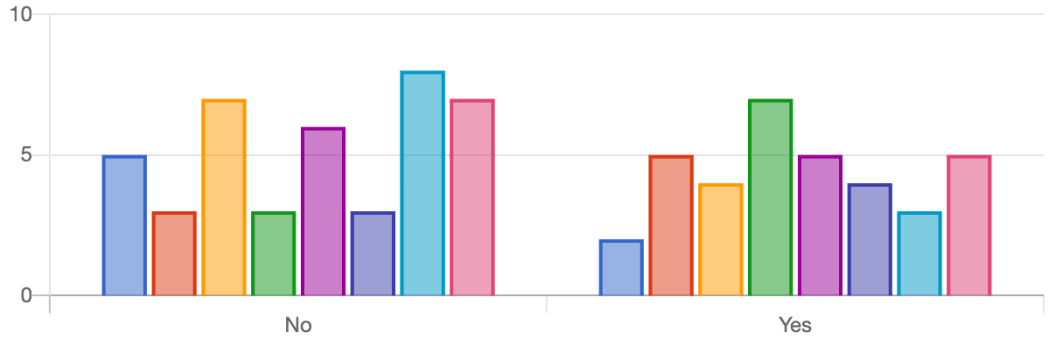
Number of Unique Respondents:19



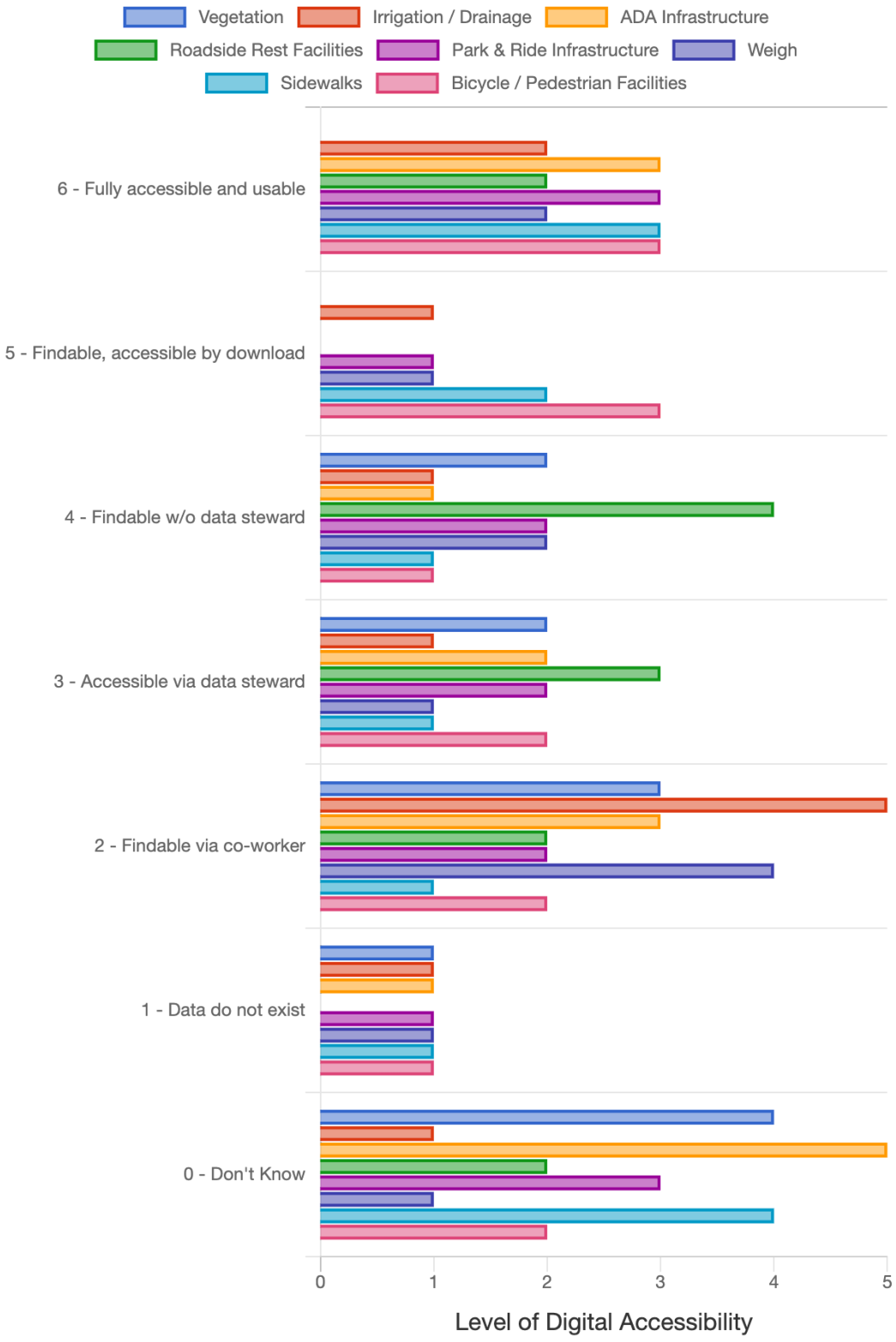
Up to Date



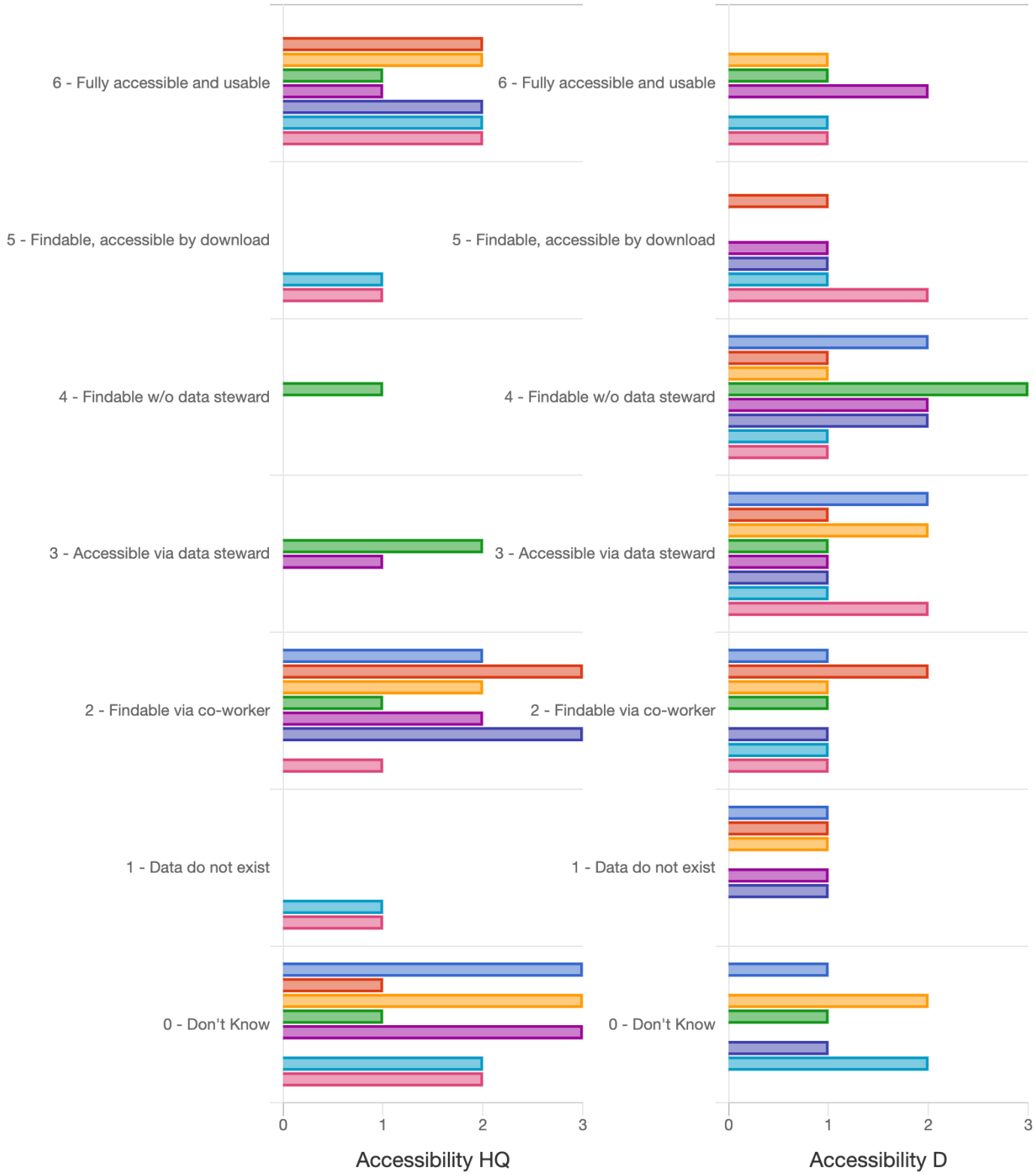
Needed



Complete and Comprehensive


Vegetation
Irrigation / Drainage
ADA Infrastructure
Roadside Rest Facilities
Park & Ride Infrastructure
Weigh
Sidewalks
Bicycle / Pedestrian Facilities
6 - Fully accessible and usable
5 - Findable, accessible by download
4 - Findable w/o data steward
3 - Accessible via data steward
2 - Findable via co-worker
1 - Data do not exist
0 - Don't Know
0
1
2
3
4
5
Level of Digital Accessibility

Headquarters vs District



Mobility Data

Moblity data considered were: flow/volume data, speed data, vehicle hours travelled/vehicle miles travelled, trip and mode choice, origin, intersection turning counts, GPS traces and freight/good movement.Respondents were asked to indicate their experience with the data that they need to perform their work.

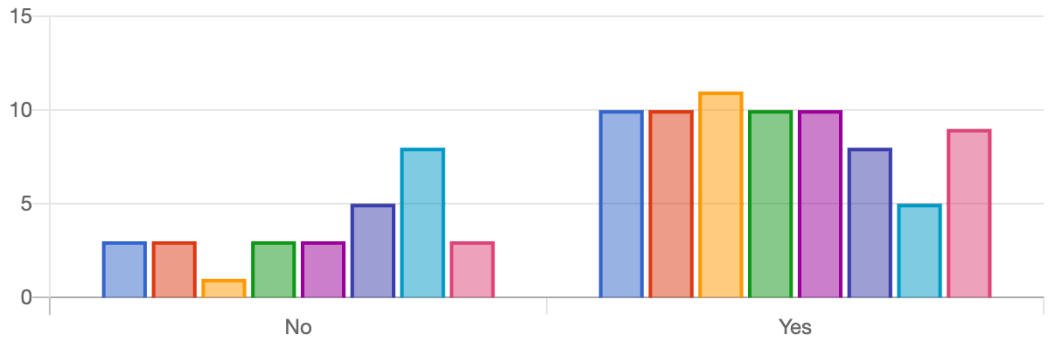
Summary:

- **Currency**
  - **Very few respondents thought that mobility data were up to date**
- **Necessity:**
  - **Most agreed that mobility data is needed.**
  - **GPS traces were ranked the lowest.**
- **Completeness**
  - **Flow, speed, and VHT/VMT data were considered the most complete in this category**
  - **Few considered freight, OD and GPS traces to be complete**
- **Discovery, access, and use**
  - **Freight data were considered the most findable, accessible and usable, with flow/wolume and VHT/VMT not quite there but most digitally accessible.**
  - **Respondents had very different opinions about the other data types or did not know.**

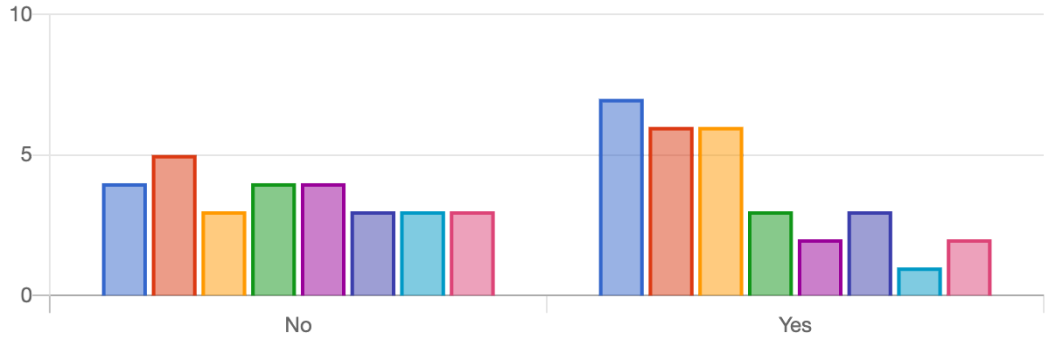
Awareness of the data was low.

Number of Unique Respondents:12

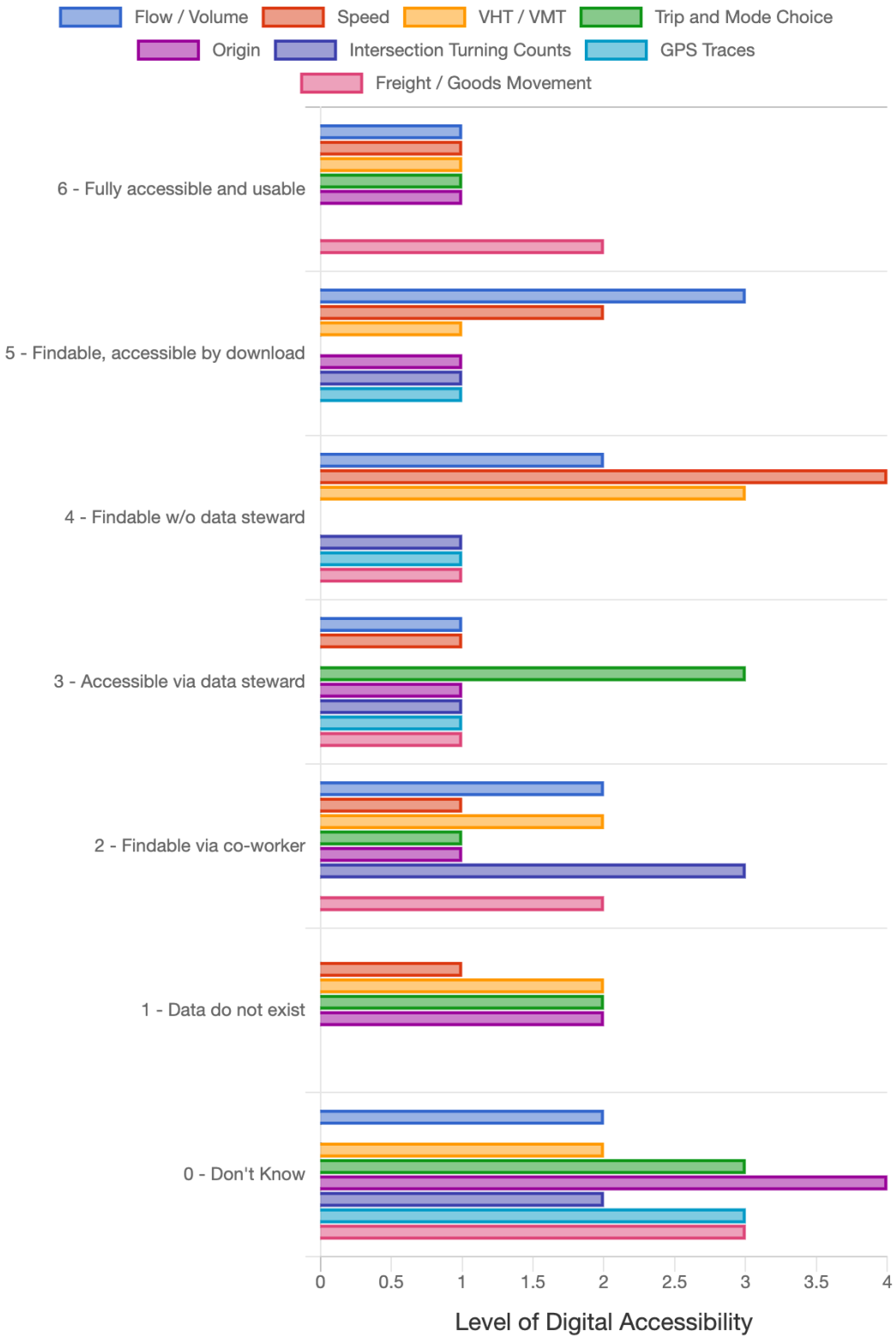
Up to Date



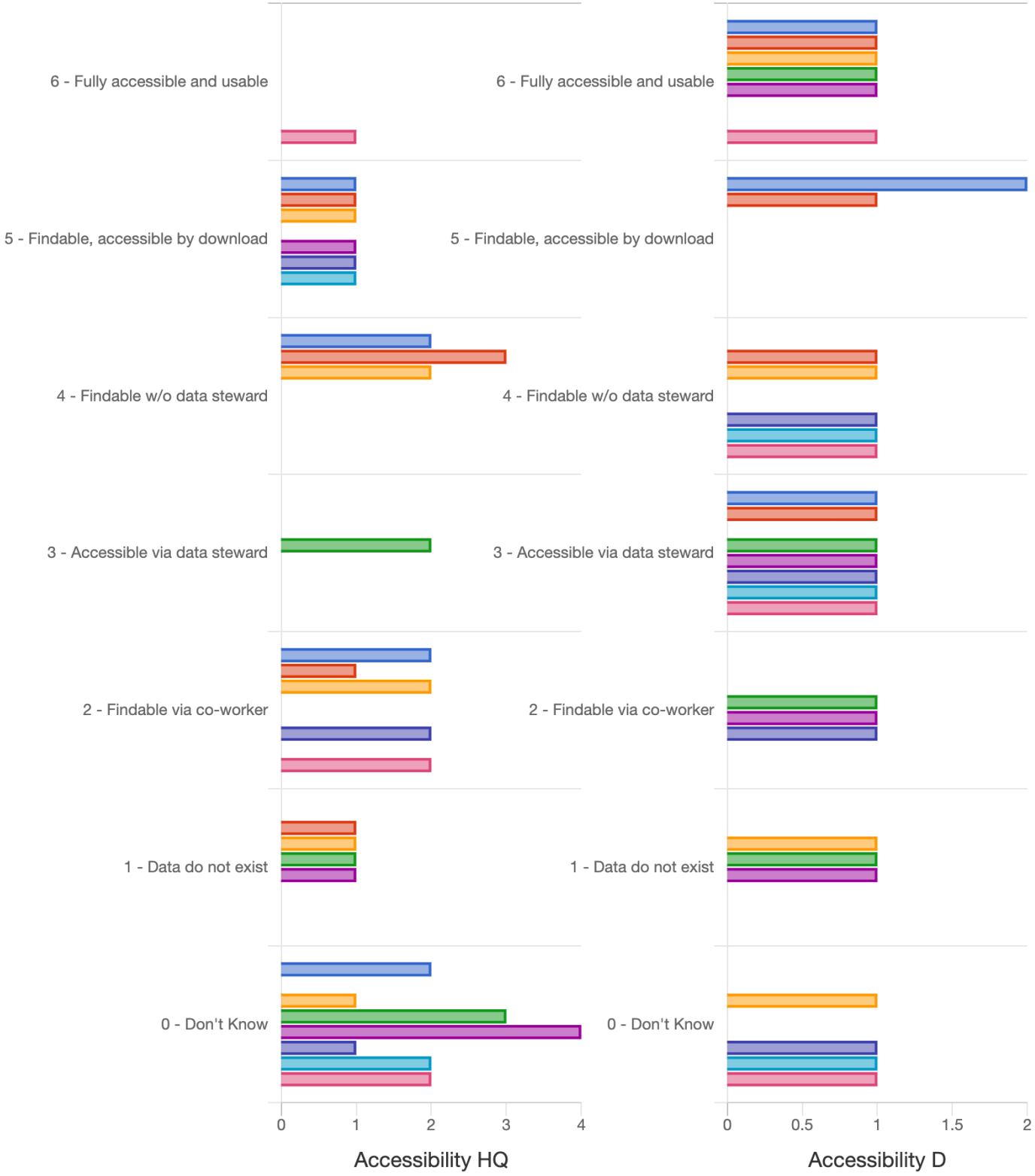
Needed



Complete and Comprehensive


Flow / Volume
Speed
VHT / VMT
Trip and Mode Choice
Origin
Intersection Turning Counts
GPS Traces
Freight / Goods Movement
6 - Fully accessible and usable
5 - Findable, accessible by download
4 - Findable w/o data steward
3 - Accessible via data steward
2 - Findable via co-worker
1 - Data do not exist
0 - Don't Know
0
0.5
1
1.5
2
2.5
3
3.5
4
Level of Digital Accessibility

Headquarters vs District



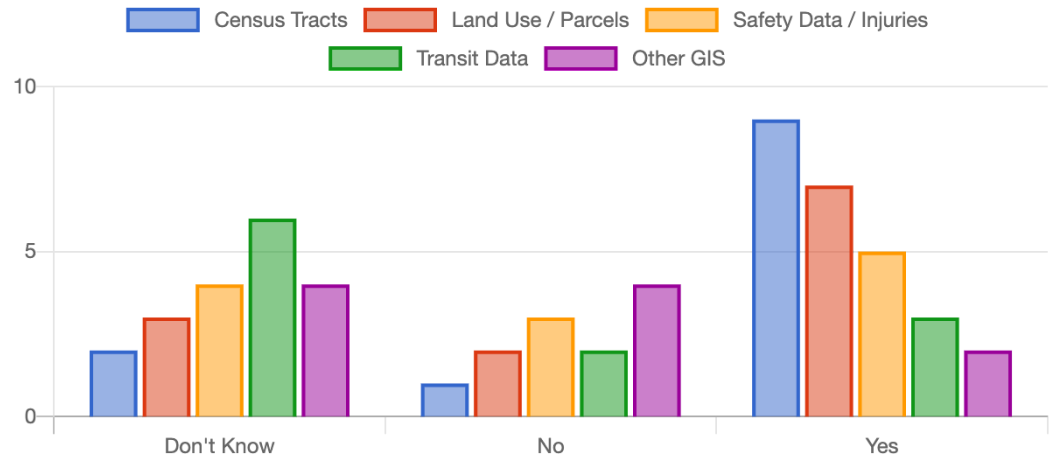
Planning

Planning data considered were: census tracts, land use/parcels, safety data/injuries, transit data and other GIS data.

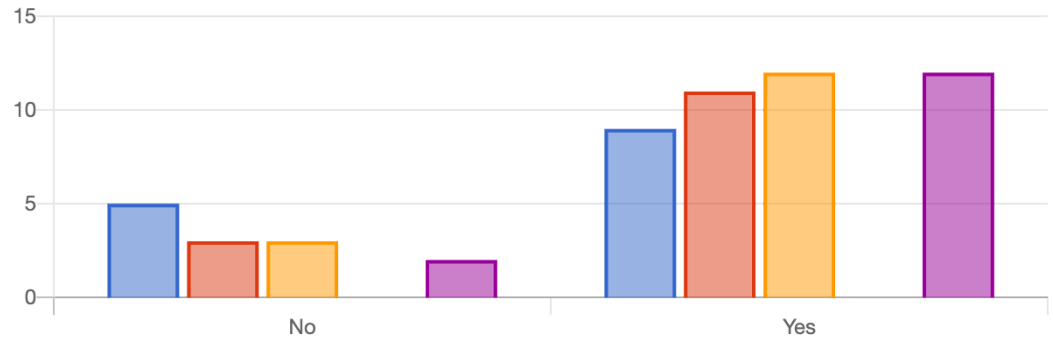
Summary:

- **Currency**
  - **Census tracts and land use / parcel data were considered the most up to date**
- **Necessity:**
  - **Most agreed that planning data is needed**
- **Completeness**
  - **Census tracts and land use / parcel data were largely considered to be complete**
  - **Respondents had split opinion about completeness of transit data**
- **Discovery, access, and use**
  - **Census tracts and land use / parcel data were considered the most findable and accessible**
  - **Respondents had very different opinions about transit data**

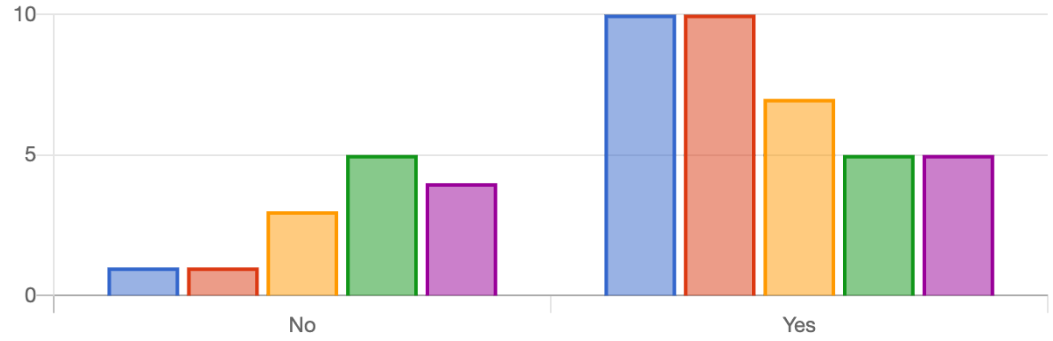
Number of Unique Respondents:14



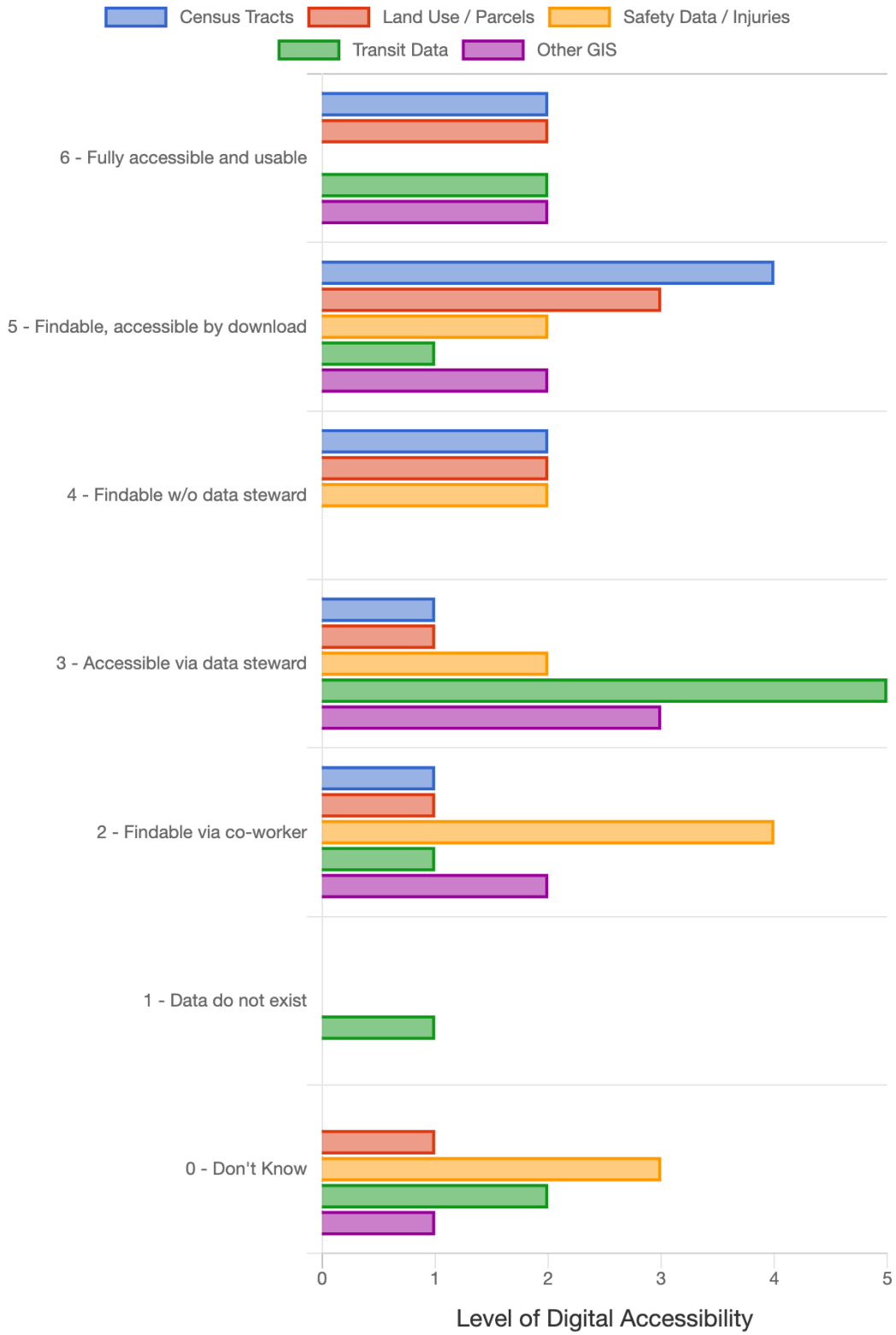
Up to Date



Needed



Complete and Comprehensive


Census Tracts
Land Use / Parcels
Safety Data / Injuries
Transit Data
Other GIS
6 - Fully accessible and usable
5 - Findable, accessible by download
4 - Findable w/o data steward
3 - Accessible via data steward
2 - Findable via co-worker
1 - Data do not exist
0 - Don't Know
0
1
2
3
4
5
Level of Digital Accessibility

Headquarters vs District

Additional Data Types, Suppliers and Consumers

Summary

Management of the Organization
Employee Project Tracking
Financial Data
Project Management
SB-1 post mile locations
Project Status
Construction Contracts
Prevailing Wage and Labor Compliance
Personel Information
Cost of Equipment
Change Orders

Physical World Data
Right of Way
Freight
Utility
Land Survey
Monument Mapping
Species, Habitat, Water
LiDAR, Aerial Imagery
CADD
Climate Change
Pump Stations
Road Classifications
Road Characteristics
Elevation
Curb Ramps
Extinguishable Message Signs
Changeable Message Signs
Fiber Optic Cable
V-2-1 Stations
Communications and Data Hubs
WIM Stations
Wildlife Movements Around Bridges and Culverts
Travel Time Between Modes for Different Corridors
Board/Alight Counts for Transit
Stormwater Treatment
Trash Collection, Sediment, Drainage

IOT Device Monitoring: Exposure Risk, Pollution, Water, Air Quality
Fleet Mix Beyond Truck/No Truck, Fuel Type

Detailed by Repondent Organization

Q: Are there any additional data types you need or use that were not mentioned above?

From	To (responder)
Right of way lines	District, 6, Environmental Analysis
Employee data, employee tracking, project management, Project tracking, workload tracking, archival records, Consultant tracking	District, 6, Construction
Project, financial data	District, 11, Project Management
need more freight data (all the good stuff is proprietary) need better, more accurate way to get O/D, VMT, occupancy data	District, 10, Transportation Planning
It wasn't clear how to interpret the last two column categories. I treated them as if any data we use should all be as complete & comprehensive as possible. And whether they are up to date, I answered I don't know. I feel that data we do use, should be as current as possible or at least as current as feasible. I will point out as an example, that having to use 2019 traffic census data when it is already 2021 is a bit behind the times. As new processes and systems for census data collection continue to evolve, hopefully there will be less latency for future volume updates. Essentially as a Planner we don't manage the data we use accessible data. We can get questions from agencies or the general public about all of the physical structures and TMS along a highway, so having direct access to what is already out there is useful. We go to subject matter experts when we need more detail or clarification. Some of the information may not be utilized when preparing a long range plan for instance; however it is a time saver when there are conglomerated- "go-to" or "one-stop-shop" places to find our standard roadway asset data along a given highway (if and when we need it).	District, 2, Transportation Planning
Utility data, project specific data, land survey data	District, 11, Transportation Planning
species occurrence data, habitat assessment data, jurisdictional waters data. This data is represented as point, line, or polygon features in GIS. Currently captured as a geodatabase for D8 Biology.	District, 8, Environmental Analysis
Yes. See list of assets in the asset management tool	District, 11, Project Management
LiDAR, aerial imagery, CADD (Design drawings/plans), climate change/sea level rise, project data (STIP/SHOPP/SB-1, etc.), pump stations, road classification, right-of-way boundaries, road characteristics (edge of pavement, number of lanes, lane width, etc.), elevation, curb ramps.	District, 4, Transportation Planning
Project Management, Financial data	District, 11, Project Management
Yes	District, 9, Project Management
Right of Way and monumentation mapping	District, 6, Right of Way and Land Surveys
Extinguishable Message Signs, Changeable Message Signs, fiber optic cable, V-2-I stations, Border Wait Time locations, Communications and Data Hubs and Weigh-In-Motion stations.	District, 11, Maintenance
for bridges and culverts, an inventory of wildlife movement through these structures would be beneficial in determining if these assets can be improved	Headquarters, Project Delivery,

	Environmental Analysis
SB 1 Project data including post mile locations, project cost (programmed and allocated), project status, milestones, etc. in a centralized database that can be accessed with a direct download and is usable. We would like to see more accurate and consistent recording of this type of information.	Headquarters, Finance, SB1 Program
n/a	Headquarters, Planning and Modal, Transportation Planning
No	Headquarters, Planning and Modal, Transportation Planning
Financial data from CTIPS, PRSM, AMS	Headquarters, Finance, Transportation Asset Management
Construction contract information (project identifying information, contractor, subcontractors, bid items, payments, schedule milestones, specifications). Prevailing wage and labor compliance information.	Headquarters, Project Delivery, Construction
Travel time between modes for different corridors. Boardings/alightings for transit stops.	Headquarters, Planning and Modal, Transportation Planning
Project Delivery - Construction Data	Headquarters, Project Delivery, Construction
N/A	Headquarters, Maintenance and Operations, Traffic Operations
We use Accounting data.	Headquarters, Finance, Accounting
Fiber Optic Plant infrastructure. And broadband locations and types.	Headquarters, Maintenance and Operations, Traffic Operations
Personnel Data, including supervisor subordinate relationships, work location, training records, performance management records, pay and benefits	Headquarters, Administration, Human Resources
Utilities, above and below ground. This included our own facilities, such as traffic conduit.	Headquarters, Project Delivery, Design

State owned/privately owned underground utility information.	Headquarters, Project Delivery, Design
1. STORMWATER TREATMENT INFRASTRUCTURE inventory, maintenance history 2. Pavement type inventory and maintenance history data by pavement category 3. Trash collection, sediment removal, drain cleaning quantitative data and status of various infrastructure elements so that alerts can be transmitted. Likely we will need IOT enabled infrastructure and a scada network to monitor our infrastructure so that human exposure to risky environments is minimized 4. Continuous water and air quality monitoring infrastructure and data collection, management	Headquarters, Project Delivery, Environmental Analysis
Data related to the cost of equipment (Equipment Watch provides this type of data) used on construction projects	Headquarters, Project Delivery, Construction
No	Headquarters, Project Delivery, Environmental Analysis
Pollution control/monitoring devices	Headquarters, Project Delivery, Environmental Analysis
no	Headquarters, Project Delivery, Design
Fleet mix - CT census only provides trucks/non-trucks. What about type of fuel (e,g, electric-powered cars, etc).	Headquarters, Project Delivery, Environmental Analysis
Stormwater BMPs.	Headquarters, Project Delivery, Environmental Analysis
Office Engineer Data for Processing of Contract documents Construction Change Order Data	Headquarters, Project Delivery, Design

High Value Data Generated by the Organization

Summary

Management of the Organization
Schedules
Financials
Reporting
Construction / Contractors

Physical World Data
Civil3D Topographic Drawings/Terrains
TMS Related Data
Bike/Ped Counts
ADL Burial Locations
Paleontology Sensitivity Maps
Hazardous Waste Site Maps
Contaminated Property Acquisitions
Water Quality and Hazardous Waste
Stormwater Quality
Project Emmissions and Noise Levels

Detailed by Repondent Organization

Q: What are the highest value datasets that you or your division generates?

Answer	Responder
Civil3d topographic drawings and digital terrain models	District, 6, Right of Way and Land Surveys
We typically create spatially-enabled data (maps) from tabular format (Excel) provided by other functional units such as projects, highway assets, traffic data, and many others.	District, 4, Transportation Planning
Project scheduling and financial data	District, 11, Project Management
Project Reporting tool that provide for statewide employees to track and monitor their projects from project information, financial, mapping, and workplan status.	District, 6, Transportation Asset Management
TMS element related data	District, 11, Maintenance
bike and ped counts	District, 10, Transportation Planning
Transportation Planning related data. We mostly try and wrangle other peoples data and try and serve it.	District, 11, Transportation Planning
Projects financials Projects scheduling	District, 11, Project Management
Project Data	District, 11, Project Management
Program/Project Management information	District, 11, Project Management
Report quarterly on the status of district projects (planned, programmed, in construction) via interactive web applications Situational awareness web applications for maintenance dispatch Maps of annual construction Management of all district assets	District, 9, Project Management
Planning is a consumer of data sets consumed by others primarily traffic census and roadside asset inventory information is our starting base.	District, 2, Transportation Planning
GIS shape files and Excel spreadsheets (worksheet).	Headquarters, Project Delivery, Environmental Analysis
I believe CTIPS data has a lot of valuable information for the SB 1 program. QMRS also has a lot of data, but I don't know how accurate that database is compared to others. I've heard of some data errors when pulling reports from QMRS	Headquarters, Finance, SB1 Program
Asset performance data	Headquarters, Finance, Transportation Asset Management
Contractor payment information. Contract administration related dashboards.	Headquarters, Project Delivery, Construction
Active Transportation Inventory	Headquarters, Planning and Modal,

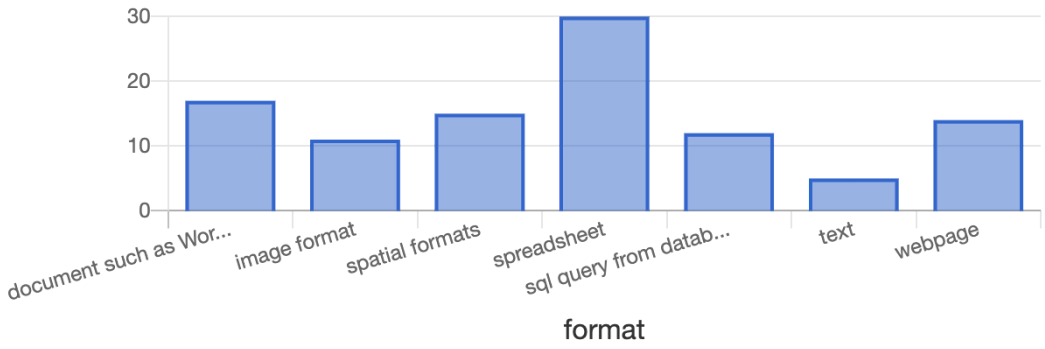
	Transportation Planning
PID Workload in PRSM database consisting of multiple projects statewide including project location, description, type, delivery date and so on.	Headquarters, Planning and Modal, Transportation Planning
I don't know what highest value means?	Headquarters, Maintenance and Operations, Traffic Operations
Personnel data	Headquarters, Administration, Human Resources
Depends on how "highest value" is defined. We generate many data sets that when needed, are of high value. For examples: - ADL burial locations - Paleontology sensitivity maps - Hazardous Waste site locations -Hazardous mineral locations -Contaminated Property acquisition locations/projects	Headquarters, Project Delivery, Environmental Analysis
Design data. Speaking on behalf of Districts Design units.	Headquarters, Project Delivery, Design
Water quality and hazardous waste	Headquarters, Project Delivery, Environmental Analysis
sotrmwater quality monitoring	Headquarters, Project Delivery, Environmental Analysis
NA	Headquarters, Project Delivery, Environmental Analysis
unknown	Headquarters, Project Delivery, Environmental Analysis
Over 3 million .pdfs	Headquarters, Project Delivery, Design
Numerical data sets. Bivariate data sets. Multivariate data sets. Categorical data sets. Correlation data sets.	Headquarters, Maintenance and Operations, Traffic Operations
project emissions/noise levels	Headquarters, Project Delivery, Environmental Analysis

Exchange Formats

Summary:

Spreadsheets and documents are used for exchange.

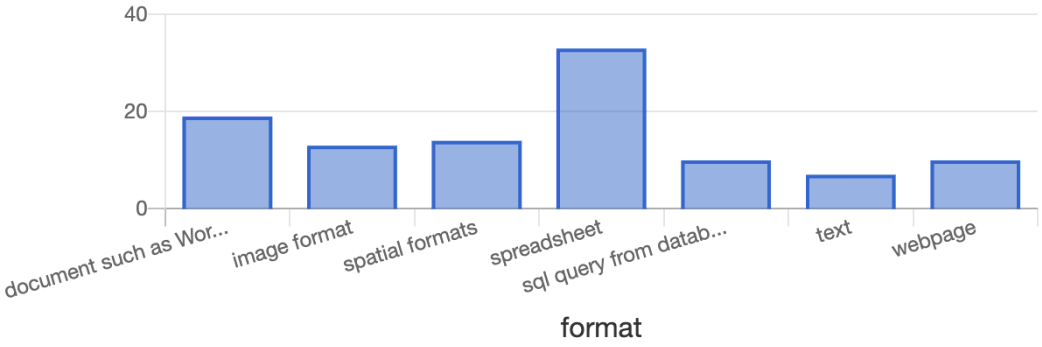
Q: In which formats do you provide data to others in the organization? Please click all that apply. - Selected Choice



Q:In which formats do you provide data to others in the organization? Please click all that apply. - other, please fill in - Text

Answer	Responder
Tableau, SSRS, SSAS	District, 11, Project Management
ODBC tables	District, 6, Transportation Asset Management
C3D DRAWINGS	District, 6, Right of Way and Land Surveys
Land Xml	Headquarters, Project Delivery, Design
Tableau dashboards	Headquarters, Finance, Transportation Asset Management

Q: In which formats do you obtain data from others in the organization? Please click all that apply. - Selected Choice



Q:In which formats do you obtain data from others in the organization? Please click all that apply. - other, please fill in - Text

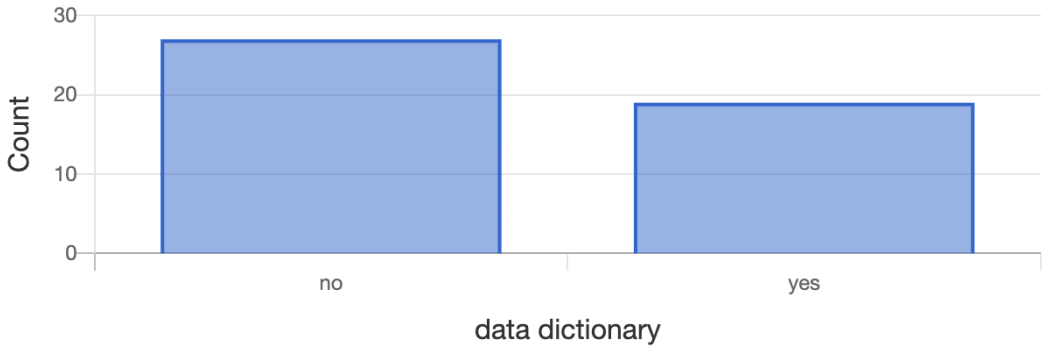
Answer	Responder
Aerials	District, 6, Environmental Analysis
python, Access DB, FileMaker Pro DB,	District, 1, Transportation Planning

MetaData

Summary:

Lack of metadata for relating datasets because metadata is still under development. Lack of standards and automatic tools or templates for generating metadata.

Q: Do you have a data dictionary to relate (join) multiple datasets across stakeholders / programs / divisions (e.g., are the same units / conventions used across data sets)?

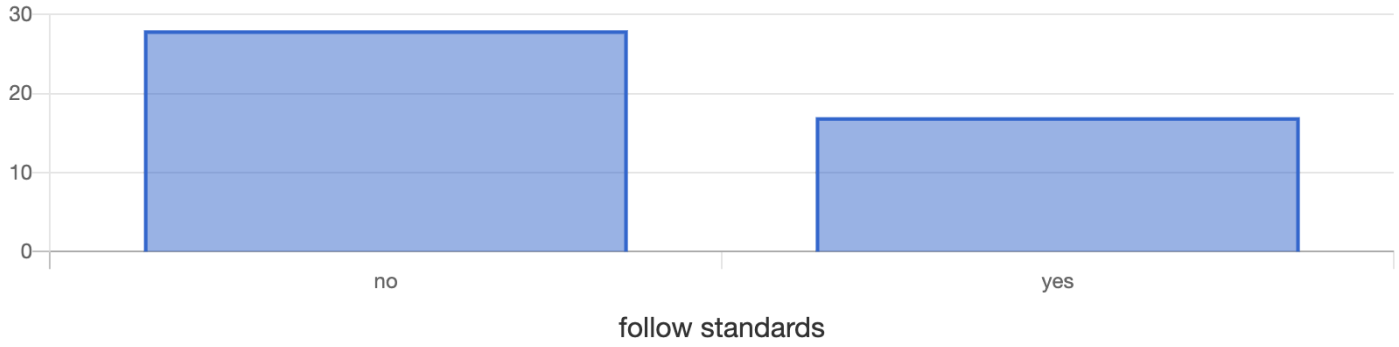


Q: Why not?

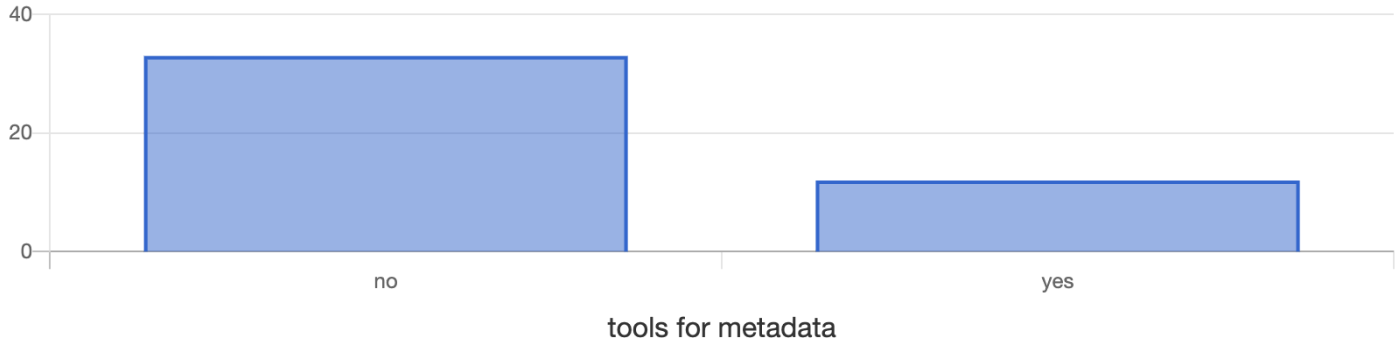
Answer	Responder
Not applicable should be an option, we do not create published data sets, a data dictionary, or metadata in planning, we use existing data sets prepared by others.	District, 2, Transportation Planning
There isn't a formal document in existence yet.	District, 11, Project Management
A data map is not established yet.	District, 11, Project Management
We do not share our data with different divisions	District, 6, Construction
Too many programs. We have basic ones like direction, county, route	District, 11, Transportation Planning
development currently in progress	District, 8, Environmental Analysis
not applicable	Headquarters, Project Delivery, Construction
NA	Headquarters, Project Delivery, Environmental Analysis
In the process of having one.	Headquarters, Planning and Modal, Transportation Planning
We collect data from programs that don't necessarily have a data dictionary in place	Headquarters, Finance, SB1 Program
N/A	Headquarters, Maintenance and Operations, Traffic Operations
Will create data dictionary and associated data governance documents after completion of CAS data governance effort.	Headquarters, Project Delivery, Construction
I dont know	Headquarters, Maintenance and Operations, Traffic Operations

Personnel/position data is consistent accross stakeholders	Headquarters, Administration, Human Resources
Much of our data is not developed to the point where a data dictionary can even be considered.	Headquarters, Project Delivery, Environmental Analysis
Data gets stored in our db which has been already standardized.	Headquarters, Project Delivery, Design
different functions., offices have their databases, gis sites built to different specs and even platforms	Headquarters, Project Delivery, Environmental Analysis

Q: Do you follow standards when generating metadata for your data assets?



Q: Do you have automatic tools or templates to help generate metadata for you?



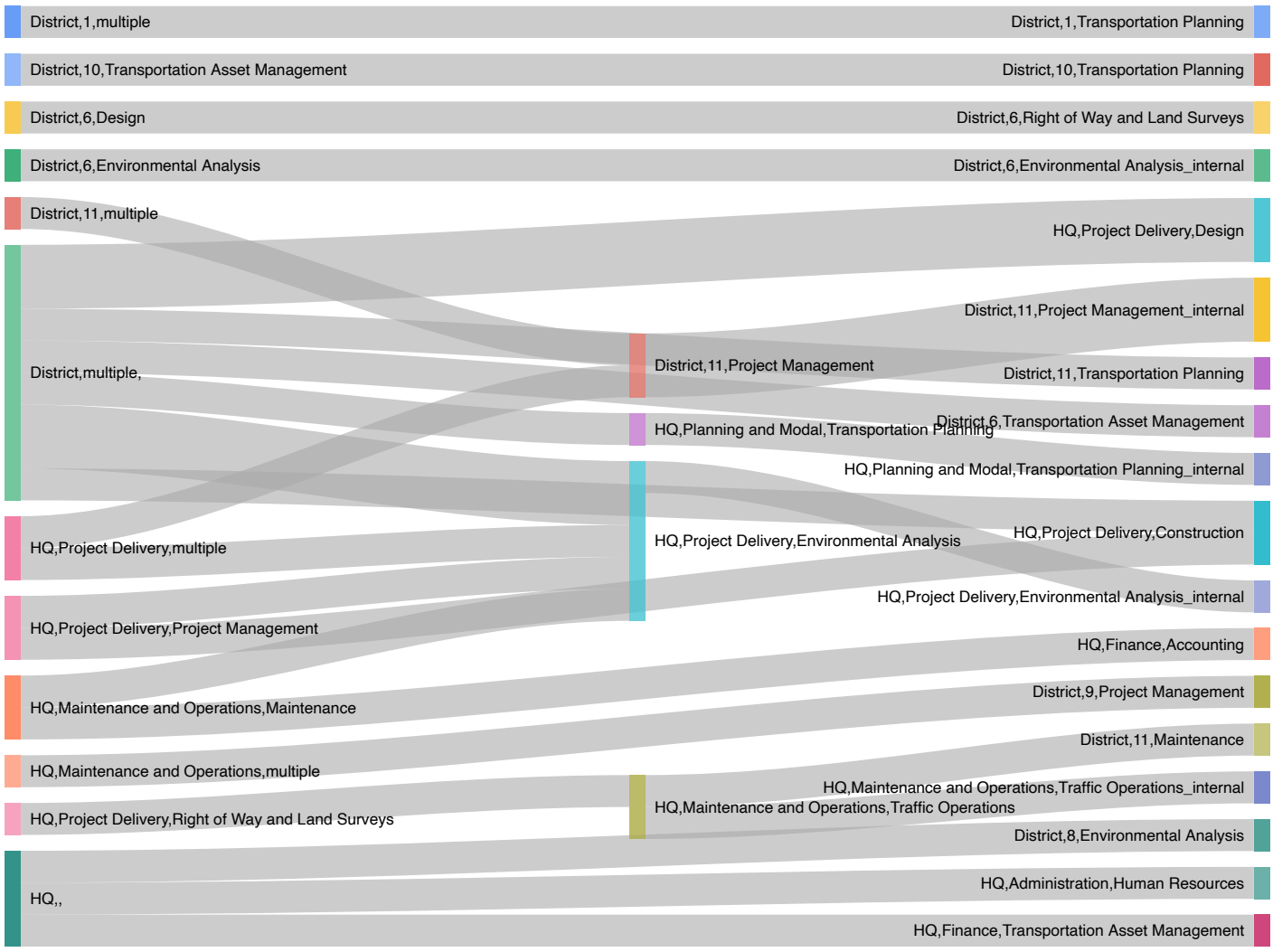
Q:What metadata standards do you use? Please select all that apply or fill in. - other, please fill in - Text

Answer	Responder
CSAC Coding	District, 6, Right of Way and Land Surveys
Caltrans Standards	District, 11, Maintenance
Caltrans Web GIS Guide & NR GIS Guidance 2017	District, 1, Transportation Planning
Per required by model	Headquarters, Project Delivery, Environmental Analysis
Caltrans data governance standards.	Headquarters, Project Delivery, Construction
This information is unknown, selected yes to provide comment	Headquarters, Administration, Human Resources
Subsurface Utility Engineering standards	Headquarters, Project Delivery, Design
defined by 23CFR772	Headquarters, Project Delivery, Environmental Analysis

Organizational Exchanges of Data

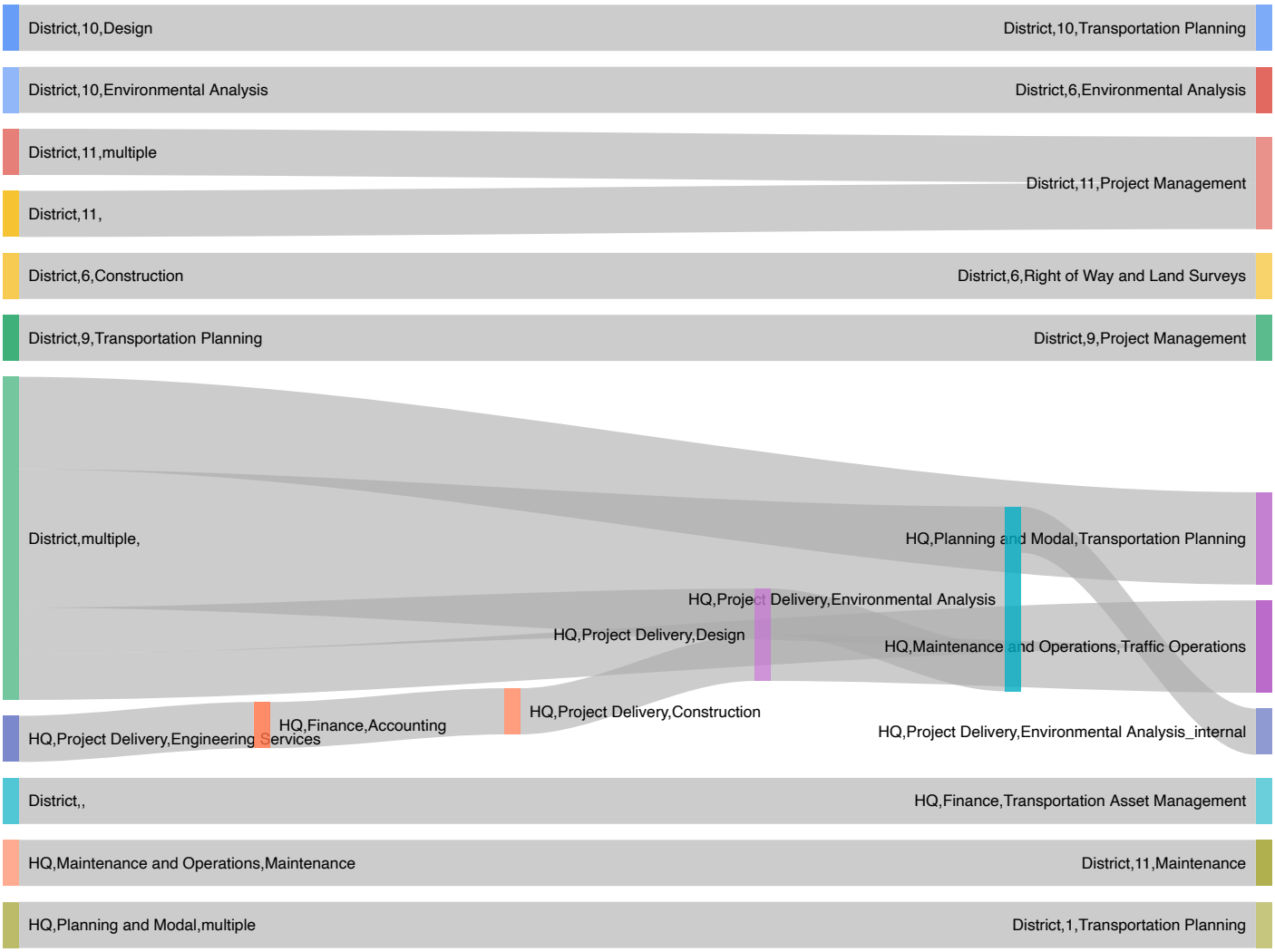
Provide Data To (Respondent is on the right side)

Data Flow <----------



Consume Data From (Respondent is on the right side)

Data Flow ---------->



Q: If your most relevant data content suppliers do not appear in the drop-down lists above, please write in your response below:

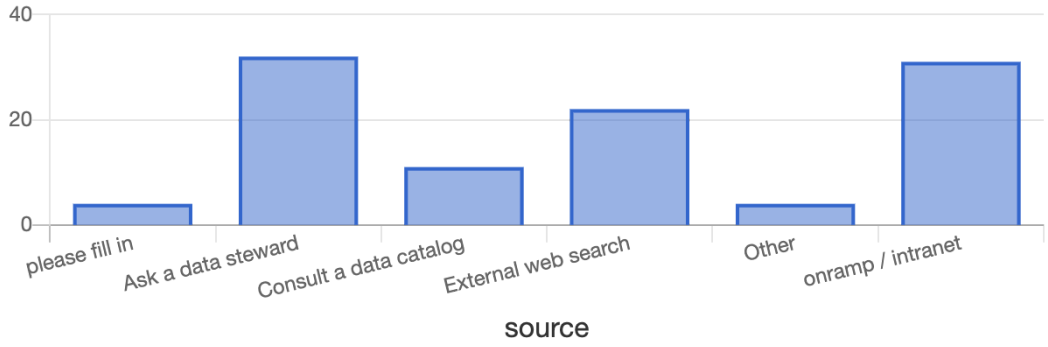
Most Relevant Supplier not on list	To (responder)
Traffic Census, Asset Inventory Information are the most frequently used.	District, 2, Transportation Planning
Engineering	District, 11, Project Management
Obtain data from other State and Federal sources, open source data (OpenStreetMap)	District, 1, Transportation Planning
Program/Project Management, Traffic Operations, and most other functional units	District, 4, Transportation Planning
on call consultant	District, 8, Environmental Analysis
traffic volume, speed, fleet mix	Headquarters, Project Delivery, Environmental Analysis
We obtain asset inventory/condition data from HQ Programs	Headquarters, Finance, Transportation Asset Management
Project Delivery Reports online database	Headquarters, Project Delivery, Construction
Traffic Operations	Headquarters, Maintenance and Operations, Traffic Operations

Lane Closure information	Headquarters, Maintenance and Operations, Traffic Operations
equipment watch (equipment rental rates)	Headquarters, Project Delivery, Construction
Design, Construction	Headquarters, Project Delivery, Design

Q: If your most relevant data consumers do not appear in the drop-down lists above, please write in your response below:

Most Relevant Data Consumer not on list	To (responder)
Engineering	District, 11, Project Management
Planning	District, 10, Transportation Planning
A mix of every division.	District, 11, Transportation Planning
D11 Engineering	District, 11, Project Management
On call consultant, resource agencies	District, 8, Environmental Analysis
Construction, Design, Environmental, Legal, Maintenance, Planning, Right of Way, Traffic Operations	Headquarters, Project Delivery, Design
District NPDES Coordinators	Headquarters, Project Delivery, Environmental Analysis
We collect data from various programs/databases and provide the compiled information on our Rebuildingca.ca.gov website. We use this website to communicate the funding and status of SB 1 projects throughout the state	Headquarters, Finance, SB1 Program
We provide asset management data to HQ and Districts through the AM Tool	Headquarters, Finance, Transportation Asset Management
District labor compliance officers and staff.	Headquarters, Project Delivery, Construction
Traffic Operations	Headquarters, Maintenance and Operations, Traffic Operations
Personnel information is provided to EEO and ODS. Position related information is provided to all districts/divisions on a scheduled basis.	Headquarters, Administration, Human Resources
waterboards	Headquarters, Project Delivery, Environmental Analysis
CDFW	Headquarters, Project Delivery, Environmental Analysis

Q: How do you discover and find information that you need? (click all that apply) - Selected Choice



Q:(click all that apply) - Other, please fill in - Text

Answer	Responder
GIS Support Unit, IT	District, 8, Environmental Analysis
Ask fro help	District, 3, Environmental Analysis
Some data is sent to us from HQ or the Engineers	District, 6, Construction
ArcGIS Online, D3 TMC Portal, HQ Web GIS	District, 1, Transportation Planning
Self generate from GIS	District, 9, Project Management
co-worker	Headquarters, Project Delivery, Environmental Analysis
Various control agency databases including State Controller's Agency, CalPERS, SCIF	Headquarters, Administration, Human Resources
cross reference	Headquarters, Project Delivery, Construction
Every 3 years ask Districts to report data	Headquarters, Project Delivery, Environmental Analysis
Caltrans GIS libraries	Headquarters, Project Delivery, Environmental Analysis

Provenance and Quality

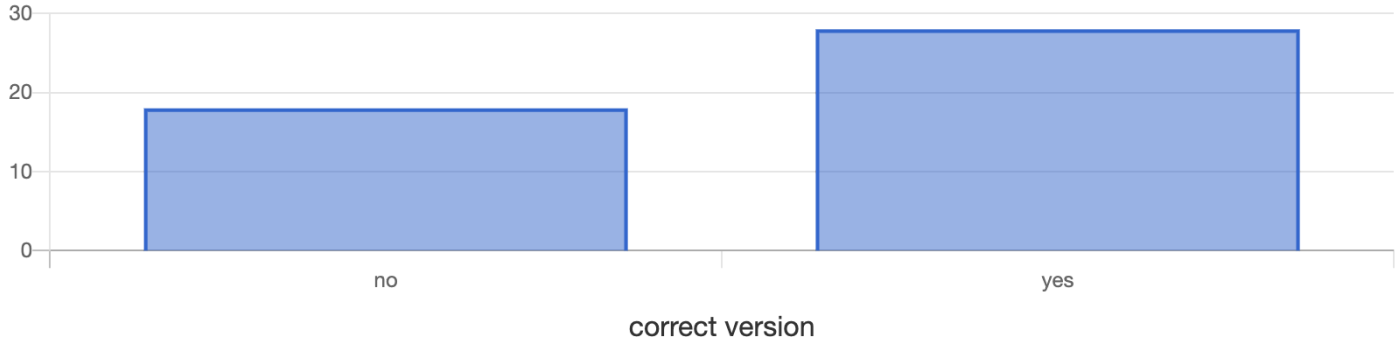
Summary:

Generally feel they have correct data, slightly high no on authoritative, higher on currency, slightly high no on reliable. Very little understanding of any data conditioning or analytics that may have been applied to the data.

	Inference	**Cross Check**	**Data Steward**	**Source**	**Data/Signature**	**Process**
Correct	4	10	2	5	3	0
Authoritative	4	0	2	7	1	3
Current	3	4	2	3	8	2
Reliable	2	8	1	4	3	0

Correctness

Q: Do you have a mechanism to confirm you have the correct version of the data?



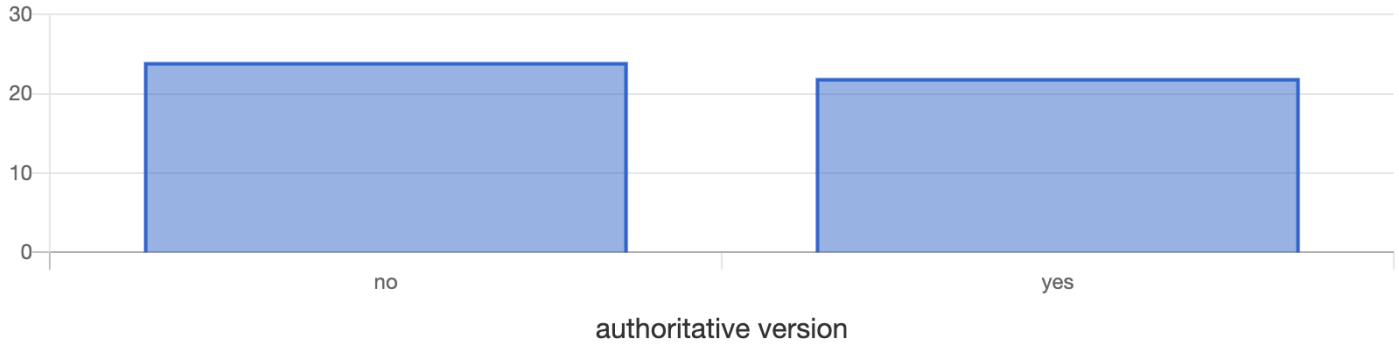
Q: What mechanism do you use to confirm you have the correct version of the data?

Answer	Responder
Use the authoritative published source	District, 2, Transportation Planning
Compare data with HQ main data.	District, 6, Transportation Asset Management
Contact the Data steward for confirmation	District, 11, Maintenance
We have data validated through our internal databases and also have humans review inputs to make sure it was done correctly	District, 6, Construction
ask the data source	District, 10, Transportation Planning
Depends on source, not always available	District, 11, Project Management
We rely on data provider to provide current data sets.	District, 4, Transportation Planning
check data versus other systems i.e, PRSM vs AMS vs CTIPS	District, 11, Project Management
Run the same data from other reporting systems to cross check.	District, 11, Project Management
Data Steward performs checks	District, 6, Right of Way and Land Surveys
Download data from different reporting systems and ensure the results matched	District, 11, Project Management
Check with data source owner	Headquarters, Finance, Transportation Asset Management
Confirm publish date.	Headquarters, Project Delivery, Construction
multiple cross references	Headquarters, Project Delivery, Construction
Reach out to the originator!	Headquarters, Maintenance and Operations, Traffic Operations
IT sends an email that the previous nights processing was complete.	Headquarters, Finance, Accounting
Visual inspect display or go to Google maps	Headquarters, Maintenance and Operations, Traffic Operations
If it is from the Intranet, I must assume it is the latest version. From the Internet, I can check version numbers if published.	Headquarters, Project Delivery, Environmental Analysis
time stamps are added to records when data are created.	Headquarters, Project Delivery, Environmental Analysis
dates	Headquarters, Project Delivery, Environmental Analysis

Google Map and Streetview to help verify	Headquarters, Project Delivery, Environmental Analysis
Caltrans GIS libraries	Headquarters, Project Delivery, Environmental Analysis
Documents containing the As Built Plans date stamp and Engineers signature.	Headquarters, Project Delivery, Design
Mandatory fields/Automated checks	Headquarters, Planning and Modal, Transportation Planning
get from the unit that prepare or will prepare the report or use defaults from CARB	Headquarters, Project Delivery, Environmental Analysis
Personally check	Headquarters, Planning and Modal, Transportation Planning

Authoritative

Q: Do you have a mechanism to confirm that you are using the authoritative copy of the data?

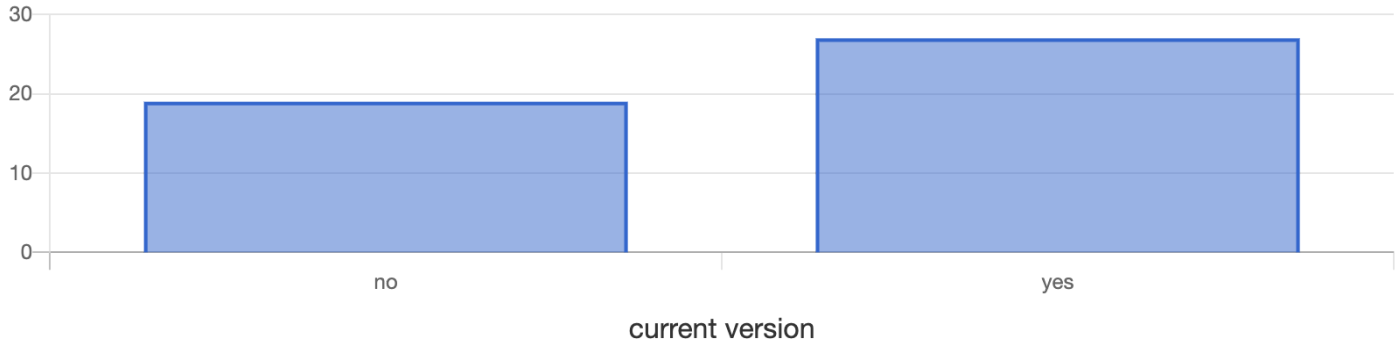


Q: What mechanism do you use to confirm you have the authoritative copy of the data?

From (responder)	To
Download via County websites	District, 6, Environmental Analysis
Contact the Data steward for confirmation	District, 11, Maintenance
ask the data source	District, 10, Transportation Planning
We rely on data provider to provide current data sets.	District, 4, Transportation Planning
In General validate correct authoritative sources and utilize data directly from their website.	District, 2, Transportation Planning
Data steward	District, 6, Right of Way and Land Surveys
Depends on source, not always available or have access to source	District, 11, Project Management
Our data comes from PRSM database	Headquarters, Planning and Modal, Transportation Planning
get from the unit that prepare or will prepare the report or use defaults from CARB	Headquarters, Project Delivery, Environmental Analysis
Check with data source owner	Headquarters, Finance, Transportation Asset Management
go to the source	Headquarters, Project Delivery, Construction
IT sends an email that the previous nights processing was complete.	Headquarters, Finance, Accounting
If it is from the Intranet, I must assume it is the authoritative version. If it is from the Internet, I get data from the source that created it.	Headquarters, Project Delivery, Environmental Analysis
we understand the flow of data between systems.	Headquarters, Project Delivery, Environmental Analysis
The data that we use is gathered from the official publishers of the data (Federal sources)	Headquarters, Project Delivery, Design
laboratory reports	Headquarters, Project Delivery, Environmental Analysis
Caltrans GIS libraries	Headquarters, Project Delivery, Environmental Analysis
Documents containing the As Built Plans date stamp and Engineers signature.	Headquarters, Project Delivery, Design

Current

Q: Do you have a mechanism to know the data is current?



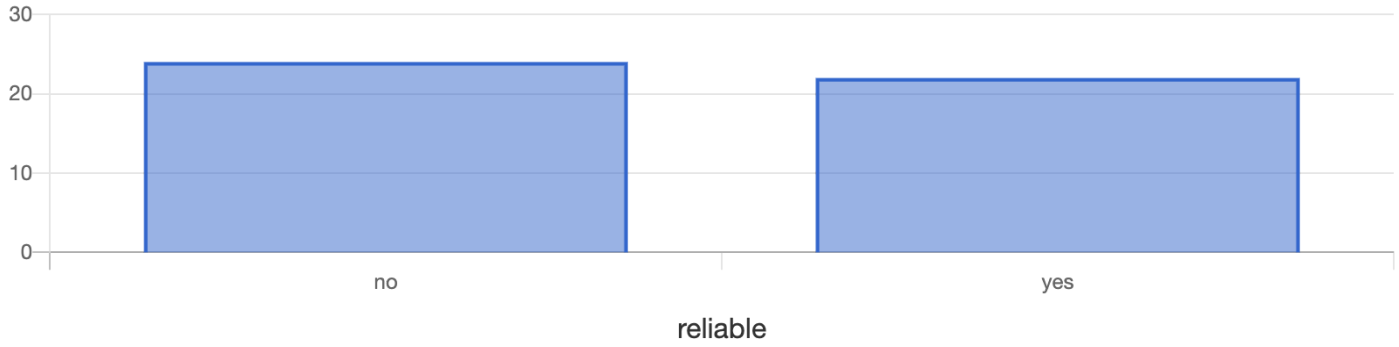
Q: What mechanism do you use to know the data is current?

Answer	Responder
We rely on data provider to provide current data sets.	District, 4, Transportation Planning
Contact the Data steward for confirmation	District, 11, Maintenance
ask the data source	District, 10, Transportation Planning
Depends on source, not always available	District, 11, Project Management
Run the same data from other reporting systems to cross check.	District, 11, Project Management
In General validate correct authoritative sources and utilize data directly from their website.	District, 2, Transportation Planning
Data steward	District, 6, Right of Way and Land Surveys
Download data from different reporting systems and ensure the results matched	District, 11, Project Management
District communication/Data checks with other databases for project information consistency	Headquarters, Planning and Modal, Transportation Planning
ask the unit, check the model what data was used - perhaps not current but latest	Headquarters, Project Delivery, Environmental Analysis
verify the update date in the database	Headquarters, Project Delivery, Design
Check with data source owner	Headquarters, Finance, Transportation Asset Management
Confirm publish date.	Headquarters, Project Delivery, Construction
cross reference	Headquarters, Project Delivery, Construction
Date Stamp	Headquarters, Maintenance and Operations, Traffic Operations
IT sends an email that the previous nights processing was complete.	Headquarters, Finance, Accounting
All Personnel data has dates associated	Headquarters, Administration, Human Resources
we create a time stamp when data is pulled from another system	Headquarters, Project Delivery, Environmental Analysis
The data generally published by year by the Federal sources	Headquarters, Project Delivery, Design
dates	Headquarters, Project Delivery, Environmental Analysis
field survey	Headquarters, Project Delivery, Environmental Analysis
Caltrans GIS libraries	Headquarters, Project Delivery, Environmental

	Analysis
Documents containing the As Built Plans date stamp and Engineers signature.	Headquarters, Project Delivery, Design

Reliable

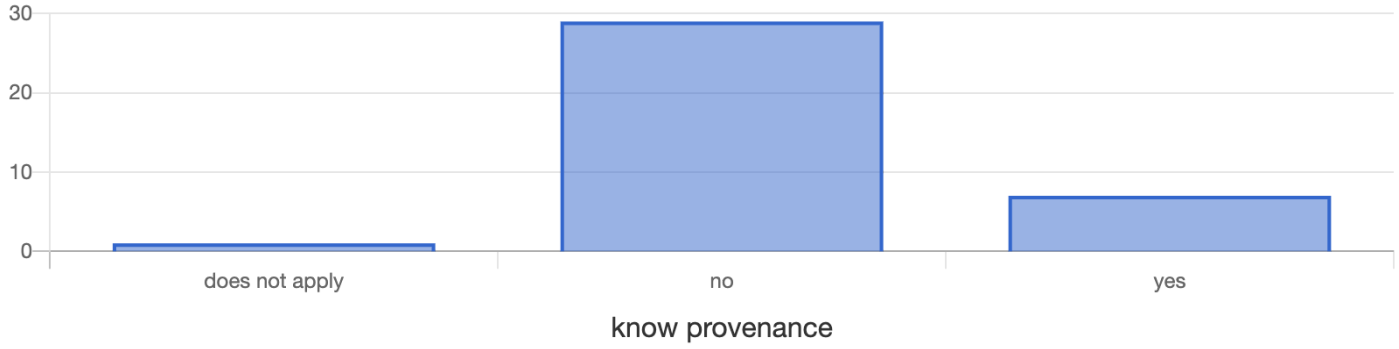
Q: Do you have a mechanism to know if the data is valid and reliable?



Q: What mechanism do you use to know the data is valid and reliable?

Answer	Responder
Download data from different reporting systems and ensure the results matched	District, 11, Project Management
Contact the Data steward for confirmation	District, 11, Maintenance
Run the same data from other reporting systems to cross check.	District, 11, Project Management
We rely on data provider to provide current data sets.	District, 4, Transportation Planning
Field checks	District, 6, Right of Way and Land Surveys
Depends on source, not always available. Various data quality checks	District, 11, Project Management
Documents containing the As Built Plans date stamp and Engineers signature.	Headquarters, Project Delivery, Design
District communication/Data checks with other databases for project information consistency	Headquarters, Planning and Modal, Transportation Planning
ask the unit; use the approved model	Headquarters, Project Delivery, Environmental Analysis
Check with data source owner	Headquarters, Finance, Transportation Asset Management
cross reference	Headquarters, Project Delivery, Construction
Various reconciliations	Headquarters, Finance, Accounting
Data comes from control agency official record sources, data is assumed to be reliable	Headquarters, Administration, Human Resources
we try to always collect data from authoritative sources	Headquarters, Project Delivery, Environmental Analysis
Most of the data that we use is generated by verified data publishers.	Headquarters, Project Delivery, Design
QA/QC OF DATA	Headquarters, Project Delivery, Environmental Analysis
Caltrans GIS libraries	Headquarters, Project Delivery, Environmental Analysis

Q: Do you know what data conditioning or analytics have been applied to generate the data you obtain from others in the organization?



Looking Forward

Summary

Future Data Needs
Safety
Litter Abatement
Bike and Pedestrians
More: GIS, WIM, PeMS
UAV
Imagery
Charging Stations
Point Cloud
CAV Infrastructure
Noise Monitoring
Storm Water Treatment
More Monitoring: water, air quality, pollution
Road Characteristics
Right of Way
CADD
Disadvantaged Businesses, Disabled Veteran Businesses
Derived Data: Reduction GHG, VMT..

Detailed by Respondent

Q: What additional kinds of data will be needed in the future?

Future Data Needs	Responder
Safety, Litter Abatement data	District, 11, Project Management
Possibly more project management tracking	District, 6, Construction
Bike/Ped Census counter location and census data	District, 1, Transportation Planning
Combined GIS mapping of all assets	District, 10, Maintenance
need more WIM and PeMS stations especially in rural areas aggregate data from the sensors on cars	District, 10, Transportation Planning
UAV outputs	District, 11, Transportation Planning
unknown. It is evolving in need.	District, 8, Environmental Analysis
All of the above.	District, 4, Transportation Planning
CAD, Right of Way, Surveys, Imagery (satellite/plane/drone)	District, 9, Project Management
Charging stations, Park and ride lots with "Smart" parking capabilities, EV charging stations, point-to-point radios, wrong way systems and drone docking sites.	District, 11, Maintenance
Point Cloud Data	Headquarters, Project Delivery, Construction
N/A	Headquarters, Maintenance and Operations, Traffic Operations
The same.	Headquarters, Finance, Accounting
Connected Automated Vehicle infrastructure.	Headquarters, Maintenance and Operations, Traffic Operations
Data linked through a platform such as GIS, where all data can be access from one portal rather that in many separate data silos. The data may exists, it is just unknown where.	Headquarters, Project Delivery, Environmental Analysis

1. STORMWATER TREATMENT INFRASTRUCTURE inventory, maintenance history 2. Pavement type inventory and maintenance history data by pavement category 3. Continuous water and air quality monitoring infrastructure and data collection, management	Headquarters, Project Delivery, Environmental Analysis
As per federal code 23CFR772, every three years, newly constructed sound walls or noise barriers are to be reported to FHWA Washington DC.	Headquarters, Project Delivery, Environmental Analysis
Possible new pollution control/monitoring devices	Headquarters, Project Delivery, Environmental Analysis
GIS online library that translates the BRIS data	Headquarters, Project Delivery, Environmental Analysis
Outcomes of completed projects including reduced VMTs, reduction in GHG emissions, incident reduction, increase in transit use, etc. Data we need in the future consist of outcomes from a completed project. Examples of outcomes are, reduction in GHG emission, reduction in VMTs, amount increase/decrease of transit trips, reduction in safety incidents, etc.	Headquarters, Finance, SB1 Program
n/a	Headquarters, Planning and Modal, Transportation Planning
Disadvantaged Business Enterprises and Disabled Veteran Business Enterprises information.	Headquarters, Project Delivery, Construction