# Data Assessment, Governance Scoping and Initiation Guide

## Overview

Caltrans has defined a standard data governance process that can be pursued by any business unit that has an interest in improving data quality, documentation, interoperability and accfess. A set of templates and guidance resources are available to support these efforts and promote consistency in how we document and share information about Caltrans data resources. The process is illustrated and described below. It is best carried out by a working group that includes people who are knowledgeable about how the data of concern are stored and managed and any issues related to data quality, consistency and access.

Step 4

Implement

Step 3

Scope Governance Effort

Step 2

Identify Gaps

Step 1

Select Data Scope

## Step 1. Select Data ScopeX

**Identify what data you would like to include in the effort**.You may want to select an entire database, a set of related databases, or a set of data elements that appear across multiple systems you would like to include in a new data mart or report series. Be careful not to take on too much at one time.

Use the criteria in Table 1 to see if the data you have selected meets criteria for Corporate Data at Caltrans. If it does, meet with the Caltrans Geospatial Data Officer (GDO) to discuss whether to pursue formal designation of the data. Data governance support provided by the Enterprise Data Governance Program (CTDATA) is prioritized for Corporate Data.

If you are considering multiple data governance efforts but don’t have the resources to pursue all of them, work through the questions in Table 2 to help you identify which effort should be pursued first. The more “Yes” responses, the higher the priority.

## Step 2. Identify Gaps

**Assess the data to identify gaps.** Conduct an assessment of the data to identify what issues might need to be addressed within the scope of a data governance effort. Use the assessment and scoping tool in Appendix A.

## Step 3. Scope Governance Effort

**Create a workplan for a data governance effort**. Use the assessment and scoping tool in Appendix A to identify activities to close gaps in the data. This workplan should consider actions in the following categories:

* Roles
* Data Documentation
* Data Value
* Data Quality
* Data Standardization, Consistency, and Interoperability
* Data Classification
* Data Protection and Preservation

Walk through your draft workplan with the GDO. This provides an opportunity for you to clarify what is involved for each activity and get connected with tools or other resources that may be helpful. It also enables the GDO to identify any possible synergies between your effort and other data governance efforts. You may also be requested to provide a brief presentation to the Enterprise Data and Geospatial Governance Board.

Line up resources needed to carry out your workplan. Identify internal staff that can spend time on the effort and (if appropriate) line up supplemental resources.

## Step 4. Implement

If your governance initiative will be making decisions impacting multiple business units across Caltrans, it is a good idea to establish a Technical Advisory Committee (TAC) to ensure coordination and alignment across affected parties.

A charter for this group should clearly identify the scope of the effort, the membership and its roles, and expectations with respect to meetings and other time commitments. A sample TAC charter can be found here (future). You may wish to consult with your Enterprise Data Steward (EDS) on TAC membership, and include them in an advisory role.

Move forward with carrying out your initiative. Please provide quarterly status updates to the GDO/Enterprise Data and Geospatial Governance Board.

As part of the effort, create a Data Documentation Package. The Data Documentation Package (DDP) provides a single “home” to organize and link to all of the data governance deliverables that were created as part of your initiatives.

## Prioritization of Candidate Data Governance Efforts

**Table 1. Corporate Data Criteria**

| **Criteria** | **Question to Ask** | **Response** | **For Further Information, See:** |
| --- | --- | --- | --- |
| a. Master or Reference Data | Does it include master or reference data? | □ Yes  □ No | * [Caltrans Corporate Data Identification Document](about:blank) * Caltrans Data Documentation Package – Appendix B |
| b. Required by Control Agencies | Is it required for reporting to federal or state-level control agencies? | □ Yes  □ No |
| c. Shared across Caltrans | Is it currently shared across multiple Caltrans business units? | □ Yes  □ No |
| d. Shared externally | Is it currently shared with external partners or with the public? | □ Yes  □ No |
| e. Potential for further internal sharing | Is it potentially of value to other Caltrans business units that may not be using it currently? | □ Yes  □ No |
| f. Potential for future external sharing | Is it potentially of value to agency partners and/or the traveling public? | □ Yes  □ No |
| g. Purchased/ Licensed Data | Is it produced as the result of state grants, contracts, or cooperative agreements? | □ Yes  □ No |

**Table 2. Data Governance Initiative Prioritization**

| **#** | **Item** | **Question** | **Response** |
| --- | --- | --- | --- |
|  | Corporate Data | Does this dataset meet criteria for corporate data? | □ Yes  □ No |
|  | Scope | Is the dataset statewide in scope? | □ Yes  □ No |
|  | Emergency Operations | Is the dataset needed for emergency operations? | □ Yes  □ No |
|  | Frequency of Use | Is the data accessed/used on a daily or weekly basis? | □ Yes  □ No |
|  | Frequency of Updating | Is the dataset updated on a regular basis to meet a stated ongoing business need? | □ Yes  □ No |
|  | Suitability for Use | Are there known issues with the dataset that impact its suitability for its intended uses | □ Yes  □ No |
|  | Manual Processes | Are the data in this dataset maintained primarily through manual entry? | □ Yes  □ No |
|  | Data Consistency | If this is a time series dataset, have there been changes to data collection methods or data definitions that have impacted consistency of data over time? | □ Yes  □ No |
|  | Access | Is or should the dataset be broadly shared within Caltrans? | □ Yes  □ No |
|  | External access | Is or should the dataset be made available to non-Caltrans employees? | □ Yes  □ No |
|  | Location referencing | If the dataset incorporates location, does it use a location referencing method not compatible with the official LRS maintained by the Division of Research, Innovation and System Information? | □ Yes  □ No |
|  | Data duplication | Does this dataset fully or partially overlap with the contents of another Caltrans dataset? | □ Yes  □ No |
| Count # of Yes Responses: | | |  |

# Appendix A. Data Assessment and Scoping Tool

Once you have identified the dataset(s) to be included in your effort, use the tables below to scope your data governance effort.

The first step to take in a data assessment is to see whether the data you have selected in your scope meets criteria for Corporate Data. A full gap assessment is recommended for Corporate Data.

Each item asks a Yes/No question that provides an “as-is” assessment of your dataset(s). If the answer to each question is “Yes”, there is no need for further action – roles are clearly defined; the data are well documented, and there is no pressing need for improvements to data quality, access or consistency. Otherwise, a data governance effort may be warranted to bring your data to a better “To Be” governed state. Each “No” answer represents a gap that can be closed through a data governance effort. Actions to consider for closing the gaps and available resources to help in carrying out the action are listed.

Not all of the gaps need be addressed through a single effort – actions should be prioritized based on the business needs. However, the items marked with asterisks are \*strongly recommended\* for all Corporate Datasets.

Note that this tool was developed primarily to support governance of Corporate Data. However, even if your dataset(s) don’t meet criteria for Corporate Data, it is still beneficial to walk through the steps outlined in the table and use the resources provided to improve data stewardship, data quality and documentation.

| **1. Roles and Responsibilities – Items checked “No” are Gaps** | | | **Actions to Consider for Closing Gaps** | **Resources Available** |
| --- | --- | --- | --- | --- |
| a. Enterprise Data Steward | Has an Enterprise Data Steward been identified? | □ Yes  □ No | * \*Work with the Geospatial Data Officer to have stewards designated and management buy-in. | * [Caltrans Data Governance Roles and Responsibilities](about:blank) * Data Steward Listing |
| b. Business Data Stewards | Have Business Data Steward(s) been identified? | □ Yes  □ No | * \*Work with assigned Enterprise Data Steward to identify the steward(s). Provide assignments back to the Geospatial Data Officer. |
| c. Custodians | Have Data Custodian(s) been identified? | □ Yes  □ No | * \*Work with assigned Business Data Steward(s) to identify the custodian(s). Provide assignments back to the Geospatial Data Officer. |
| Comments: | | | | |

| **2. Data Documentation – Items checked “No” are Gaps** | | | **Actions to Consider for Closing Gaps** | **Resources Available** |
| --- | --- | --- | --- | --- |
| a. Data Catalog | If the data meets any of the criteria for Corporate Data, are there one or more entries for this data in the Caltrans Data Catalog? | □ Yes  □ No | * \*Prepare a data catalog entry and provide to the Geospatial Data Officer | * Data Documentation Package – Appendix A, Guidance for Data Catalog * Data Catalog Template |
| b. Dataset Metadata | Is there metadata for the dataset(s) available in the Caltrans dataset metadata catalog? | □ Yes  □ No | * \*Compile any existing dataset-level metadata, and use to populate the Caltrans Dataset Metadata Template. * \*Fill in the remaining items, and provide to the Geospatial Data Officer. | * Data Documentation Package – Appendix E, Guidance for Dataset Metadata * Dataset Metadata Template |
| c. Data Dictionary | Is there a data dictionary describing each of the data elements in the Caltrans data dictionary catalog? | □ Yes  □ No | * \*Compile any existing data dictionary information and use to populate the Caltrans Data Dictionary Template. * \*Fill in the remaining items, and provide to the Geospatial Data Officer. | * Data Documentation Package – Appendix G, Guidance for Creating a Data Dictionary * Data Dictionary Template * Dataedo Tool (work with Enterprise Data and Geospatial Governance Program staff to utilize this tool) |
| d. Corporate Data Element Sources | For data that has been designated as Corporate Data, have the Corporate Data Element Sources been identified and categorized? | □ Yes  □ No | * \*Complete the Corporate Data Element Sources Template for each corporate data element | * Data Documentation Package – Appendix B, Guidance for Documenting Corporate Data Element Sources * Caltrans Corporate Data Element Sources Template |
| e. Entity-Relationship Diagram | Is there an Entity-Relationship (E/R) diagram showing how the different physical database tables are related to one another? | □ Yes  □ No | * Work with the data custodian to create an E/R diagram | * Dataedo Tool (work with Enterprise Data and Geospatial Governance Program staff to utilize this tool) * [Sample E/R diagram 1](about:blank) * [Sample E/R diagram 2](about:blank) |
| f. Context Diagram | Is there a high level context diagram available that illustrates the systems that store the data, how data flows into and out of that system, and the sources and recipients of the data? | □ Yes  □ No | * \*Create a high level context diagram showing data flows. | * Data Documentation Package – Appendix D, Data Diagrams * [Sample context diagram](about:blank) |
| g. Lineage Diagrams | Are there more detailed lineage diagrams showing how given data elements flow across different systems? | □ Yes  □ No | * Consider creating detailed data lineage diagrams. | * Data Documentation Package – Appendix D, Data Diagrams * [Sample lineage diagram 1](about:blank) * [Sample lineage diagram 2](about:blank) |
| h. Business Glossary | Have terms and definitions for key entities and attributes for this dataset been identified for inclusion in the Caltrans business glossary? | □ Yes  □ No | * Scope a task to identify potential glossary terms and gain agreement on their definitions. | * Data Documentation Package – Appendix F, Guidance for Defining Business Glossary Terms * Caltrans Business Glossary Template |
| Comments: | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **3. Understanding of Data Value – Items checked “No” are Gaps** | | | **Actions to Consider for Closing Gaps** | **Resources Available** | |
| a. Users and Uses | Is there any documentation of how this data is being used (e.g. list of primary users, diagram of how data flows to different reports, data value maps, etc.) | □ Yes  □ No | * Consider creating a data value map that links data sources, products and users/uses or conducting a data value assessment | * Data Documentation Package – Appendix D, Data Diagrams * NCHRP Data Value Assessment Tool * Sample data value diagram (future) | |
| b. Business Function Mapping | Have the Caltrans business functions been identified that this data supports? | □ Yes  □ No |
| c. Data Processing Schedule | If there are reporting deadlines – internally or externally imposed that drive the data collection and updating schedule, have these been documented? | □ Yes  □ No | * Consider mapping the data collection and processing activities and responsibilities (if not already completed) | * [Sample data processing diagram](about:blank) | |
| Comments: | | | | | |
| **4. Data Quality – Items checked “No” are Gaps** | | | **Actions to Consider for Closing Gaps** | **Resources Available** |
| a. Business Rules | Have business rules for data validation been documented for each of the data elements? | □ Yes  □ No | * \*Create business rules using the Caltrans business rules template | * Data Documentation Package – Appendix H, Guidance for Specifying Business Rules * Caltrans Business Rules Catalog Template |
| b. Data Quality Management Plan | Is data quality being actively managed for this dataset – i.e. are there well-defined of quality control and assurance steps applied to the data and are these documented in a data quality management plan (or equivalent)? | □ Yes  □ No | * \*Create and implement a Data Quality Management Plan | * Caltrans Data Quality Management Plan Template |
| c. Data Profiling - Completeness | Has data completeness been assessed: e.g. percentage of records with missing values, percentage of target scope covered? | □ Yes  □ No | * Scope a data profiling task to obtain quantitative measures of quality and identify issues such as duplicate records, missing records, or invalid records. | * [Data Profiling Tutorial](file:///C:\_fdh\SPP%20Projects\Caltrans%20CMAS\Tasks%20and%20Deliverables\3.1%20Ad%20Hoc%20Support\3.1.1-Templates\Final%20Versions\di_loreto_data_profiling_tutorial_monday_am.pdf) |
| d. Data Profiling - Validity | Has data validity been assessed: e.g. number or percent of records with invalid values? | □ Yes  □ No |
| e. Data Profiling - Currency | Has data currency been assessed: e.g. date of last update for data records? | □ Yes  □ No |
| f. Data Usability | Has data usability been assessed: e.g. reported user ratings of how easy it is to obtain and transform the data for use? | □ Yes  □ No | * Scope a task to convene a user focus group or conduct a user survey the better understand user perspectives. | * NCHRP Data Value Assessment Tool |
| Comments: | | | | |

| **5. Data Standardization, Consistency, and Interoperability – Items checked “No” are Gaps** | | | **Actions to Consider for Closing Gaps** | **Resources Available** |
| --- | --- | --- | --- | --- |
| a. Location Referencing | If this dataset includes location references, are these consistent with the Caltrans Linear Referencing System (LRS), and are processes in place to enable historical data to be located on the current LRS? | □ Yes  □ No  □ NA | * \*Scope a task to explore options for updating location references in the dataset. | * [Caltrans Linear Referencing System Branch](about:blank) |
| b. Authoritative Sources | Is there a shared understanding or agreement about what the “official” authoritative source system is for the data elements in the dataset? | □ Yes  □ No | * \*Scope a task to gain agreement on and document authoritative sources. |  |
| c. Duplicative Data | Is there reasonable confidence that the data in this dataset are not duplicated across various Caltrans systems – other than to provide an authorized reporting source? | □ Yes  □ No | * \*Scope a task to identify and suggest strategies to reduce data duplication. |  |
| d. Interoperability | If elements from this dataset are found in multiple systems, has there been a review of opportunities to improve consistency or streamline how these elements are shared across systems? | □ Yes  □ No  □ NA | * \*Scope a task to explore improvements to interoperability – e.g. using data integration tools, data warehouse solutions, or data services. |  |
| e. Applicable Standards | Has this dataset been assessed for conformance to existing Caltrans or national data standards? | □ Yes  □ No | * \*Scope a task to identify and assess conformance to data standards. |  |
| f. Future Standardization | Is there reasonable confidence that there would be no benefit to further standardization of definitions, formats, or lists of values of the data elements in this dataset? | □ Yes  □ No | * \*Scope a task to develop Caltrans data standards for one or more elements. |  |
| Comments: | | | | |

| **6. Data Classification – Items checked “No” are Gaps** | | | **Actions to Consider for Closing Gaps** | **Resources Available** |
| --- | --- | --- | --- | --- |
| a. Identification of Sensitive and Confidential Data | Have the data been reviewed to identify any sensitive or confidential data elements (e.g. personally identifiable information or PII)? | □ Yes  □ No | * \*Review and classify data elements based on established California Department of Technology Information Security Program Management Standard * Scope a task for data de-identification | * [California Department of Technology Information Security Program Management Standard SIMM 5305-A](about:blank) * Caltrans information security guidance (future) * [CHHS Deidentification Guidelines](about:blank) |
| b. Documentation of Sensitive and Confidential Data | Have data element classifications been documented in a data dictionary? | □ Yes  □ No | * \*Record data element classifications in the Caltrans Data Dictionary Template | * Data Documentation Package – Appendix G, Guidance for Creating a Data Dictionary * Data Dictionary Template |
| c. Identification of Public, Restricted and Non Public Data | Have the data been assessed to determine whether they are public, restricted or non-public? | □ Yes  □ No | * \*Review and classify datasets based on whether they are public, restricted or non-public | * Data Documentation Package – Appendix A, Guidance for Data Catalog and Appendix E, Guidance for Dataset Metadata * Caltrans Data Catalog Template * Caltrans Dataset Metadata Template |
| d. Documentation of Public, Restricted and Non Public Data | Have dataset classifications of public, restricted, and non-public been documented in dataset metadata? | □ Yes  □ No | * \*Record dataset classifications in the Caltrans Data Catalog and Dataset Metadata Templates | * Data Documentation Package – Appendix A, Guidance for Data Catalog and Appendix E, Guidance for Dataset Metadata * Caltrans Data Catalog Template * Caltrans Dataset Metadata Template |
| Comments: | | | | |

| **7. Data Access, Protection and Preservation – Items checked “No” are Gaps** | | | **Actions to Consider for Closing Gaps** | **Resources Available** |
| --- | --- | --- | --- | --- |
| a. Data Access | Is this dataset generally available and accessible to people who need it or may find it valuable – both internally and externally? | □ Yes  □ No | * Scope a task to improve internal and/or data accessibility | * [California Open Data Handbook](about:blank) |
| b. Sensitive Data Protection | If the data are sensitive, are sufficient protections in place to control access to the data? | □ Yes  □ No | * \*Scope a task to review and revise how data access permissions are granted and revoked. |  |
| c. Backups | Are the data currently being backed up? | □ Yes  □ No | * \*Work with Caltrans Information Technology to implement backup procedures. |  |
| d. Verification of Backups | Have backups been tested and verified by a data steward (or other business user representative) over the past 12 months to ensure that data can be restored properly? | □ Yes  □ No | * Work with Caltrans Information Technology to schedule a data restore test including business representative involvement. |  |
| e. Data Storage Requirements | Is there an understanding of data storage requirements over the next 5 years and a plan in place to meet these requirements? | □ Yes  □ No | * Scope a task to estimate future storage requirements * Scope a task to develop a data storage plan identifying where the data are to be stored, and how storage is to be resourced. |  |
| f. Data Update Plan | Is there an established and resourced data update cycle or approach to meet anticipated business needs? | □ Yes  □ No | * Scope a task to develop a data updating strategy that addresses business needs. |  |
| g. Data Retention | Has a data retention schedule been established and validated by business users? | □ Yes  □ No | * \*Scope a task to identify and revise (as needed) the applicable data retention schedule |  |
| h. Data Archiving Plan | Is there a resourced plan for data archiving? | □ Yes  □ No | * Scope a task to create a data archiving plan to ensure sufficient access to the data to meet business needs through the end of the retention period. |  |
| Comments: | | | | |