

**Data Documentation Package**

**for:**

**<Scope>**

Version XXX

**Document Control**

|  |  |  |
| --- | --- | --- |
| **Version** | **Date** | **Description** |
| 1.0.0 | 11/1/2019 | Baseline |
| 1.0.1 | 12/16/2020 | Updated GIS metadata elements to align with State Geoportal statewide standard |
| 1.0.2 | 1/31/2020 | Updated GIS metadata and data dictionary elements based on staff feedback and further internal review. |
| 1.0.3 | 03/09/2020 | Updated metadata and data dictionary elements based on GIS Strategic Framework Team feedback. |
| 1.0.4 | 12/15/2020 | Minor clean-up work to ISO metadata elements and their descriptions. |
| 1.0.5 | 1/29/2021 | Clean up of Field Type and Domain Type language in the metadata section to align with how Esri (GIS) has implemented those items. |
| 1.0.6 | 2/14/2021 | Additional metadata tag clean-up to such that each data dictionary item has a unique metadata element it maps to. Removed data dictionary items that should be addressed in data quality management plan. |
| 2.0.0 | 8/11/2021 | Version 2 release of document based on adopted Metadata and Data Dictionary practice documents. |
| 2.1.0 | 3/17/2022 | Updated data catalog elements. |
| 2.1.1 | 4/8/2022 | Minor correction to data catalog elements. |
| 3.0.0 | 5/5/2023 | Updated data catalog and business data glossary tables to conform with newly released applications.  Updated metadata to improve usability and clarity |
| 3.1 | 3/11/2024 | Updated metadata to improve usability and clarity in regard to the Access and Use Constraints |

Contents

[1. Introduction 1](#_Toc161037435)

[Contents 1](#_Toc161037436)

[2. Data Catalog 2](#_Toc161037437)

[Guidance 2](#_Toc161037438)

[Process 2](#_Toc161037439)

[Reference 2](#_Toc161037440)

[3. Corporate Data Element List 3](#_Toc161037441)

[Guidance 3](#_Toc161037442)

[Process 3](#_Toc161037443)

[Reference 3](#_Toc161037444)

[4. Data Element Standards 4](#_Toc161037445)

[Guidance 4](#_Toc161037446)

[Process 4](#_Toc161037447)

[Reference 4](#_Toc161037448)

[5. Data Diagrams 5](#_Toc161037449)

[Guidance 5](#_Toc161037450)

[Scope 5](#_Toc161037451)

[Process 5](#_Toc161037452)

[Reference 5](#_Toc161037453)

[6. Dataset Metadata 6](#_Toc161037454)

[Guidance 6](#_Toc161037455)

[Scope 6](#_Toc161037456)

[Process 6](#_Toc161037457)

[Reference 6](#_Toc161037458)

[7. Business Glossary Terms 7](#_Toc161037459)

[Guidance 7](#_Toc161037460)

[Scope 7](#_Toc161037461)

[Process 7](#_Toc161037462)

[Reference 7](#_Toc161037463)

[8. Data Dictionary 8](#_Toc161037464)

[Guidance 8](#_Toc161037465)

[Scope 8](#_Toc161037466)

[Process 8](#_Toc161037467)

[Reference 8](#_Toc161037468)

[9. Business Rules 9](#_Toc161037469)

[Guidance 9](#_Toc161037470)

[Scope 9](#_Toc161037471)

[Process 9](#_Toc161037472)

[Reference 9](#_Toc161037473)

[Business Rule Specification Details 9](#_Toc161037474)

[10. Data Quality Management Plan 10](#_Toc161037475)

[Guidance 10](#_Toc161037476)

[Process 10](#_Toc161037477)

[Reference 10](#_Toc161037478)

[Appendix A: Guidance for Data Catalog 11](#_Toc161037479)

[Purpose 11](#_Toc161037480)

[Types of Catalog Entries 11](#_Toc161037481)

[Data Catalog Elements 12](#_Toc161037482)

[Appendix B: Guidance for Documenting Corporate Data Element Sources 15](#_Toc161037483)

[Corporate Data Element Sources 16](#_Toc161037484)

[Source Type Definitions: 17](#_Toc161037485)

[Appendix C: Guidance for Nominating Data Element Standards 17](#_Toc161037486)

[Appendix D: Data Diagrams 19](#_Toc161037487)

[Overview 19](#_Toc161037488)

[Data Flow Diagrams 19](#_Toc161037489)

[Data Use Diagrams 19](#_Toc161037490)

[Data Value Maps 19](#_Toc161037491)

[Lineage Diagrams 20](#_Toc161037492)

[Appendix E: Guidance for Dataset Metadata 21](#_Toc161037493)

[Purpose 21](#_Toc161037494)

[Metadata Items 21](#_Toc161037495)

[GIS Data 22](#_Toc161037496)

[Appendix F: Guidance for Defining Business Glossary Terms 27](#_Toc161037497)

[Corporate Business Glossary 27](#_Toc161037498)

[Style Guidelines for Writing Definitions 29](#_Toc161037499)

[Appendix G: Guidance for Creating a Data Dictionary 30](#_Toc161037500)

[Purpose 30](#_Toc161037501)

[Data Dictionary Items 30](#_Toc161037502)

[Appendix H: Guidance for Specifying Business Rules 35](#_Toc161037503)

[Purpose 35](#_Toc161037504)

[Types of Business Rules 35](#_Toc161037505)

[Business Rules Catalog Contents 36](#_Toc161037506)

[Examples 37](#_Toc161037507)

# Introduction

## Contents

This document describes and references several standard data documentation deliverables:

* **Data Catalog** –entries for the Caltrans data catalog
* **Corporate Data List** – list of data elements proposed or approved for designation as “corporate data” and designated authoritative sources of these elements
* **Data Element Standards** – existing or proposed data element standards that apply to data within the defined scope
* **Data Flows** – context and lineage diagrams describing how data flows across systems
* **Dataset Metadata** – documentation of dataset(s) packaged for end user consumption
* **Business Glossary Terms** – definitions of business terms represented by data entities or attributes within the defined scope
* **Data Dictionary** – information about each data element
* **Business Rules** – rules that can be used to provide extended documentation of individual data elements or validate data element or dataset contents
* **Data Quality Management Plan** – documentation of quality control and quality assurance measures applied to data as it is created/collected, maintained in systems, and exported for reporting or other purposes.

# Data Catalog

## Guidance

See [Appendix A](#_Appendix_A:_) for guidance on creating or updating data catalog entries covering the scope of this data documentation package. Data catalog entries should be created using the standard Caltrans Data Catalog Template.

## Process

The following process was used to create the data catalog entries:

<Describe who was involved and what sources were used>

The data catalog entries were last reviewed and updated on: <date> by <person>.

## Reference

Provide this information to the Caltrans Geospatial Data Officer (GDO). Include links to any additional related background information below.

Data catalog information is listed in <accompanying spreadsheet link>.

# Corporate Data Element List

## Guidance

See [Appendix B](#_Appendix_B:_) for guidance on identifying corporate data elements, existing sources for those elements, and conformance with adopted field name aliases for analytics and reporting.

## Process

The following process was used to identify corporate data elements and sources:

<Describe who was involved and what sources were used>

## Reference

Caltrans corporate data assets are identified in <link to Caltrans corporate data list>.

# Data Element Standards

## Guidance

See [Appendix C](#_Appendix_C:_) for guidance on nominating data element standards for consideration. Proposed data element standards should be documented using the Caltrans Data Element Standard Nomination Form.

## Process

The following process was used to identify and create the proposed data element standard:

<Describe who was involved and what sources were used>

## Reference

A list of proposed and accepted Caltrans data element standards can be found at <link to web page>

# Data Diagrams

## Guidance

See [Appendix D](#_Appendix_D:_) for guidance on creating data flow and lineage diagrams.

## Scope

Diagrams were created covering following data assets:

* <reference data inventory items >

## Process

The following process was used to create and validate the diagrams:

<Describe who was involved and what sources were used>

The diagrams were last reviewed and updated on: <date> by <person>

## Reference

Diagrams can be found in the following files: <links to files or list of file names>

# Dataset Metadata

## Guidance

See [Appendix E](#_Appendix_E:_) for guidance on creating dataset-level metadata. This is information about the dataset as a whole. For GIS datasets, some required metadata is also in the Data Dictionary [Appendix G](#_Appendix_G:_). Make sure to review the required elements.

## Scope

Metadata was created for the following datasets:

* <reference datasets >

## Process

The following process was used to create the dataset metadata:

<Describe who was involved and what sources were used>

The dataset metadata was last reviewed and updated on: <date> by <person>

## Reference

Dataset metadata can be found in the following files: <links to files or list of file names>

# Business Glossary Terms

## Guidance

See [Appendix F](#_Appendix_F:_) for guidance on creating business glossary terms representing the key business concepts represented within the scope of this data documentation package. Business glossary terms should be created using the standard Caltrans business glossary template or (when available) added directly to the Caltrans Business Glossary.

## Scope

Business glossary terms were completed for the following datasets:

* <reference data inventory items >

## Process

The following process was used to compile business glossary terms.

<Describe who was involved and what sources were used>

The business glossary was last reviewed and updated on: <date> by <person>

## Reference

Business glossary terms are listed in <accompanying spreadsheet link or link to Caltrans Business Glossary>

# Data Dictionary

## Guidance

See [Appendix G](#_Appendix_G:_) for guidance on creating a data dictionary for data assets within the scope of this data documentation package. Data dictionaries should be created using the standard Caltrans data dictionary template. Elements required in GIS Datasets are also present here.

## Scope

Data dictionary items were completed for the following data assets:

* <reference data assets >

## Process

The following process was used to create the data dictionary:

<Describe who was involved and what sources were used>

The data dictionary was last reviewed and updated on: <date> by <person>

## Reference

Data dictionary information is listed in <accompanying spreadsheet link>

# Business Rules

## Guidance

See [Appendix H](#_Appendix_H:_) for guidance on specifying business rules.

## Scope

Business rules were documented for the following datasets:

* <reference data inventory items >

## Process

The following process was used to identify and validate these rules:

<Describe who was involved, what references were consulted and what was done to review and validate the rules.>

The business rules were last reviewed and updated on: <date> by <person>

## Reference

Business rules for identified corporate data elements are listed in <accompanying spreadsheet link> and described further below.

## Business Rule Specification Details

The following sections contains additional detail to supplement what is included in the Business Rules Catalog.

##### Business Rule ID – Business Rule Name

<add supplemental detailed descriptions of business rules here – if business rules are fully documented in the catalog it is not necessary to repeat information here.>

# Data Quality Management Plan

## Guidance

See the Caltrans Data Quality Management Plan (DQMP) Template for guidance on creating a plan that describes data quality objectives, the current level of data quality (if known), current and future planned quality management practices, and quality tracking and communication strategies. The DQMP should be a living documented, updated regularly.

## Process

The following process was used to create the DQMP:

<Describe who was involved, what steps were taken, etc.>

## Reference

The DQMP can be found at <provide link>.

# Appendix A: Guidance for Data Catalog

## Purpose

The Caltrans Data Catalog is a listing of Caltrans’ current and proposed corporate databases and datasets. It is used to meet several needs:

* To facilitate compliance with the California Open Data Policy by providing a repository of datasets that can be considered for posting to the open data portal;
* To enable Caltrans staff to identify current data resources and associated points of contact;
* To provide a resource for Caltrans Enterprise Data Stewards to understand existing data sources as they review proposals for acquisition of new data and systems;
* To serve as the authoritative source of Corporate Data Assets as approved by the Caltrans Enterprise Data Governance Board; and
* To track the Enterprise Data Stewards, Business Data Stewards and Data Custodians responsible for different Caltrans Data Assets.

The catalog will store information about Caltrans databases and datasets that have been formally designated as corporate data, or that meet criteria for corporate data but have not yet been formally designated as such.

The catalog will store information about any dataset being considered for inclusion on an open data portal.

The catalog will *not* store information about individual data elements. Data element information is documented in the Data Dictionary.

## Types of Catalog Entries

The following entries will be distinguished in the Caltrans Data Catalog:

* **Enterprise Source Database** – an enterprise database (i.e. managed by Caltrans IT) containing source data. A source database is where data are actively maintained.
* **Enterprise Reporting Database[[1]](#footnote-2)** – an enterprise database created specifically to provide a source for reporting, analysis or visualization. They store extracted, cleaned and transformed data sourced from other databases. Changes to data are only made through updates from the source databases. Examples include data warehouse tables, data marts, BI reporting universes.
* **File Data Source** – a spreadsheet or desktop-based database management system (DBMS) where data are maintained that is not managed as an enterprise data system by Caltrans IT.
* **Dataset** – a collection of data elements prepared specifically for posting on a data portal or sharing with others. May be provided in various formats including comma-separated-values (CSV) file, spreadsheet, document, KML, shapefile, file geodatabase, and API.
* **Data Access Points** – online applications that provide interactive query capabilities, providing access to data not otherwise available through file downloads (for example, the Rebuilding CA map and tabular download webpage and DataLink for accounting data).

The catalog will store limited information for each entry. Additional metadata for datasets will be provided separately (see Appendix E).

## Data Catalog Elements

|  |  |  |
| --- | --- | --- |
| Data Catalog Element | Required | Description |
| APPLICATION\_ID | If Applicable | If there is a corresponding entry in the SNOW Configuration Management Database, include the asset number here. (Example: “APM0001019”) |
| PARENT\_DATA\_ ASSET\_NAME | Yes | Enter the commonly used name for the data asset – may be an abbreviation. (Example: “QMRS”) |
| PARENT\_DATA\_ASSET\_LONG\_NAME | Yes | Full name of the database or dataset  Don’t just repeat the PARENT\_DATA\_ASSET\_NAME. Spell out any abbreviations. Avoid abbreviations. (Example: “Quality Management Reporting System”) |
| DATA\_ASSET\_ NAME | If Applicable | This should be used to subdivide parent data assets into individual modules or datasets to distinguish individual data subjects or modules of a large data system. Leave this item blank for a stand-alone entry.  Use the naming convention: <Abbreviation for Parent Asset> - <Name of Subsidiary Asset> Example: “AMS-Budgets”) |
| DATA\_ASSET\_DESCRIPTION | *Yes* | Brief description of the data asset – sufficient to inform a user about the content, intended use and format of the data asset. Focus on the content of the data, not the technical aspects of the system that stores the data. |
| DATA\_ASSET\_TYPE | Yes | Select one or more of the following data asset types:  *Enterprise Source Database* (an enterprise database like Oracle, SQLServer where data is entered or updated.  *File Data Source* (a non-enterprise database like Access, Filemaker Pro or Excel where data is entered or updated.)  *Enterprise Reporting Database* (an enterprise database with data compiled for reporting from other sources)  *Dataset* (a data product such as a GIS data layer, spreadsheet, or API providing data compiled for sharing via a portal, website, etc.)  *Data Access Point* (a dashboard, mapping interface or other website providing access to data views, reports, and downloads)  *Other/Transactional* (use this item when the data asset doesn’t fit within any of the other categories and there is a business need to have it in the catalog  *Unknown* (use this item when you are not sure how to classify the data asset) |
| DATA\_ASSET\_STATUS | Yes | Active *(Default) –* (data asset is currently available and in use)  Future – (planned data asset that is not currently in use)  Retired – (data asset is no longer in active use but archived data may still be available) |
| CORPORATE\_DATA\_CATEGORY | If Applicable | Select one or more of the following:  Master Data  Reference Data  Control Agency Required Data  Shared Data  Purchased/Licensed Data  None/Unknown (EDGG staff to review) |
| DATA\_SUBJECT | Yes | Select one or more Data Subjects from the list. If there are none applicable or you aren’t sure which one fits, select Other. |
| DATA\_LINK | No | If available, enter a link providing access to the actual data or information about how to request access |
| METADATA\_LINK | No | If available, enter a link to where further information about the data asset (the dataset metadata or data dictionary) is available. |
| DATA\_AUTHORITY | If Applicable | If this dataset is used to meet an external federal or state requirement, enter the name of governing body responsible for the requirement (e.g. FHWA, CalSTA) and, if applicable, the name of the legislation (e.g. FAST Act, California Senate Bill 1) |
| ENTERPRISE\_DATA\_STEWARD | Yes | Select the name of the designated Enterprise Data Steward assigned to the asset |
| EDS\_BUSINESS\_UNIT | Yes | The business unit associated with the Enterprise Data Steward is automatically populated when the Enterprise Data Steward name is selected. |
| BUSINESS\_DATA\_STEWARD | Yes | Enter the name of the designated or proposed Business Data Steward assigned to the asset. |
| BDS\_BUSINESS\_UNIT | Yes | Select the business unit of the Business Data Steward |
| OPEN\_DATA\_READY | No | Select Yes if this data asset is ready for sharing (in whole or in part) to an open data portal;  Select No otherwise  (CTDATA Staff to manage this item) |
| PUBLIC\_ACCESS\_LEVEL | Yes | Select one:  Public – data could 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 | If Applicable | If *Public Access Level* is *Restricted* or *Non-Public*, enter a description of why data is restricted or non-public – e.g. inclusion of sensitive or confidential information.  Leave blank if *Public Access Level* is *Public*. |
| USE\_CONSTRAINTS | Yes | Enter description of restrictions or legal prerequisites to using the dataset – e.g. required license or permissions.  Enter license that is applied to the data. |
| COMMENTS | No | Enter any comments about the entry to clarify any of the other items or note information which requires verification. |
| ADD\_DATE | Yes | Date that this catalog entry was added. This item is automatically populated. |
| EDIT\_DATE | Yes | Date that this catalog entry was last updated. This item is automatically populated. |
| EDIT\_USER | Yes | Name of person who added or last updated the entry. This item is automatically populated. |

# Appendix B: Guidance for Documenting Corporate Data Element Sources

Corporate Data Elements can be identified in one of two ways:

1. A database is designated as a Corporate Data Asset, and then some or all of the data elements within that database are identified as Corporate Data Elements.
2. A set of Corporate Data Elements are identified first, independent from the database(s) in which they are stored. This second approach may be taken, for example, for design of a new external website or data mart with data items to be sourced from multiple systems; or for a data governance effort seeking to improve the quality of a selected set of important data elements found in multiple systems.

In general, Corporate Data Elements should meet one of the criteria established for Corporate Data Assets:

* They represent master or reference data
* They are required by control agencies
* They are widely shared within or outside of Caltrans
* They are procured or licensed data

Effective governance of Corporate Data requires:

* Understanding where the data are maintained, and what sources should be used for reporting
* Establishing clear and commonly understood data definitions
* Establishing clear and easy to understand data element aliases for analysis and reporting
* Establishing business rules for the data that provide the basis for assessing data validity

This appendix provides guidance for documenting data sources for corporate data. See Appendices F, G and H for guidance on creating glossary terms, data dictionary entries and business rules for corporate data.

The following steps should be followed to document data sources for each Corporate Data Element:

1. Identify the source database(s) – where the data element is entered and maintained.
2. Determine which of these source databases are considered the Source System of Record (SSOR). In general, there should be a single SSOR for each element. However, there will be exceptions. (For example, attributes for a project may have a different source system depending on what phase the project is in.)
3. Identify any reporting databases containing the element.
4. Review the field name and description for the data element as documented in the source system of record. Determine if the alias should be the same as the data element field name and description in the source system of record and needs to be modified for understandability
5. Complete the Corporate Data Element template to associate the source and reporting databases with the Corporate Data Elements. See below for documentation of this template.

## Corporate Data Element Sources

| Corporate Data Element | Required? | Description |
| --- | --- | --- |
| PRESENTATION\_DATA\_ELEMENT\_ALIAS | Yes | The agreed data element alias used for inclusion in the data warehouse and standard reports meant to be consumable by the public and non-technical readers. The is the name that would be used in the Enterprise Data Warehouse. Multiple CORPORATE\_DATA\_ELEMENT\_NAME entries must have the same PRESENTATION\_DATA\_ELEMENT\_ALIAS entry when the corporate data element definitions are the same. |
| PRESENTATION\_DATA\_ELEMENT\_DESCRIPTION | Yes | A description of the data element’s meaning and contents written in plain language and understandable by all data consumers. Avoid obvious descriptions that don’t add new information beyond the Field Title. |
| CORPORATE\_DATA\_ELEMENT\_NAME | Yes | Name of the Corporate Data Element. Should be unique across all Corporate Data Elements. Use a name that conveys the meaning of the element. Consider using the Field Title from the data dictionary for an authoritative source system. |
| CORPORATE\_DATA\_ELEMENT\_DESCRIPTION | Yes | A description of the data element’s meaning and contents as taken from the data dictionary for the data source. Avoid obvious descriptions that don’t add new information beyond the Field Title. |
| DATA\_ASSET\_NAME | Yes | Commonly used name for the data asset – may be an abbreviation – e.g. “QMRS”. 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 |
| SOURCE\_TYPE | Yes | SS - Source System, SSOR - Source System of Record, or ARS - Authoritative Reporting Source. (See definitions in Data Documentation Package) |
| COMMENTS | No | Use to document situations where there are multiple SSORs for the element or to note dynamic situations – e.g. when a new authoritative reporting source is under development. |

## Source Type Definitions:

* **A Source System (SS)** – a database in which the data element is entered or updated but is not the recommended place to obtain the data. For example, a non-authoritative system may be used to support an isolated business process and have no connection to other databases. Data in this system is not necessarily current.
* **A Source System of Record (SSOR)** for the data element. This is a source system where data are entered and updated that you would point to as the most authoritative place to obtain the data element. In some cases, there will be multiple SSORs for a data element – for example, project information may be sourced from one system during planning and another system during scoping and design.
* **An Authoritative Reporting System (ARS)** - this is a data warehouse table or other reporting database (e.g. Tableau data source) that has been designated as an authoritative source for reporting. Data are not modified within this system – data are pulled in from the SSORs and may be transformed to provide consistency and ease of reporting. In cases where there are multiple SSORs for a data element, the Authoritative Reporting System may be the most reliable place to obtain the element.

# Appendix C: Guidance for Nominating Data Element Standards

Lack of consistency across data sources can create problems when there is a need to combine data from multiple systems for reporting. Lack of consistency also presents a barrier to integrating systems to improve efficiencies. Data governance initiatives can propose new Caltrans data element standards to address inconsistencies in data elements across multiple systems. Data element standards for individual data elements should include the following:

1. Definition of the data element – sufficient to avoid varying interpretations of what it means.
2. Value Domain/Reference Tables – including code list or source for the list of possible values if applicable
3. Data type (e.g. numeric, character, date)
4. Scale/units of measure (e.g. Celsius or Fahrenheit for a temperature)

In addition, the following optional items can be considered, as appropriate:

1. Naming convention (for the physical database field name)
2. Positional accuracy (e.g. +-40 feet)
3. Coordinate system
4. Reference(s) that provide further detail on how the data element is to be captured or calculated.

To create a new data element standard:

1. Create a problem statement describing the need for the standard
2. Research existing data element standards to determine if there is an existing related standard that can be modified
3. Form a working group involving individuals who have technical knowledge of how the data element is used within the different systems where it is stored.
4. Collaborate to draft a proposed standard that will address the identified need and can be feasibly implemented.
5. Circulate the draft standard to technical stakeholders (as appropriate) and the Caltrans Enterprise Data Stewards (EDS) for review and comment.
6. Revise the draft standard as needed to respond to comments.
7. Submit the draft standard to the Caltrans Enterprise Data Governance Board for review and approval.
8. The Caltrans Enterprise Data Governance Board may approve the standard or return it with comments to be addressed.
9. Repeat steps 4-8 to modify the draft standard as needed.
10. Help put the standard into practice by communicating with colleagues and advocating for its use. The EDS will ensure that the business data stewards are informed, and the standard is posted online.

# Appendix D: Data Diagrams

## Overview

Several types of data diagrams can be produced to document how data are entered or loaded into different information systems, and how they may be transformed along the way. These diagrams are valuable supplements to tabular metadata – they can be used to support data quality improvement efforts and data change control processes. They also help data stewards, custodians and users gain a common understanding of how data are produced and managed.

It is recommended that each data governance initiative produce (at a minimum) a high-level data flow diagram – also called a context diagram.

## Data Flow Diagrams

Data flow diagrams are a well-established method for documenting how data moves from people/roles into and out of processes and data stores. A process is something that takes data as input and produces data as an output. Examples of processes are permitting, order processing, and data validation. A standard data flow diagram shows:

* External Entities - the people/roles/other systems that generate and/or receive information
* Data Stores – the databases or files that store information
* Processes - the processes that occur in the system to manipulate the information
* Data Flows - the information that enters and leaves the system and flows across processes and data stores within the system

Data flow diagrams can be prepared at different levels of detail. The highest-level diagram is called a “context diagram”. A context diagram is useful for showing what information is flowing in and out of a system, and the external sources and recipients of the information. More detailed data flow diagrams show information flows within a system’s constituent processes and data stores.

A tutorial on preparing data flow diagrams is available [here](https://datagovernance.onramp.dot.ca.gov/downloads/datagovernance/files/Process/Data%20Flow%20Diagram%20Tutorial.doc).

A sample data flow diagram is available [here](https://datagovernance.onramp.dot.ca.gov/downloads/datagovernance/files/Process/Data%20Flow%20Diagram%20Example.docx).

## Data Use Diagrams

A variant of a data flow diagram is a Data Use Diagram. A data use diagram shows how different business units use data from a selected information system.

A sample Caltrans data use diagram is available [here](https://datagovernance.onramp.dot.ca.gov/downloads/datagovernance/files/Process/Data%20Use%20Diagram%20Example.pdf).

## Data Value Maps

A data value map identifies types of users and uses for the data, and traces these to data products and data sources. An example data value map is available [here](https://datagovernance.onramp.dot.ca.gov/downloads/datagovernance/files/Process/Examples/Crash%20Data%20Value%20Map%20Example.pdf).

## Lineage Diagrams

Lineage diagrams trace the path of one or more data elements from their original point of entry to their final destination(s). They may include documentation of transformations performed along this path. Lineage diagrams are helpful for identifying and diagnosing data quality issues that may cascade across multiple databases as data are transferred from one system to another (via manual re-entry or automated processes). They are also helpful for analyzing the potential impacts of a change to a data element’s structure or definition within one data store on downstream data stores or reports. One common way to approach a lineage diagram is to begin with a critical report and then trace the elements on that report backwards to their origin. Because lineage diagrams can be complex, typically specialized tools are used to create and manage them.

# Appendix E: Guidance for Dataset Metadata

## Purpose

A dataset is defined as a collection of data elements prepared specifically for posting on a data portal or sharing with others. Datasets may be provided in various formats including spreadsheet, document, KML, shapefile, file geodatabase, API.[[2]](#footnote-3)

Dataset metadata provides information about a dataset as a whole. It is supplemented by Data Dictionary information for each data element in the dataset. Dataset metadata helps people to search for data of interest because the metadata elements can be used as filter or search criteria in a data catalog or data portal. Metadata also helps people to understand the purpose, derivation, and limitations of a dataset. It provides important information for determining whether a dataset is suitable to meet a particular information need.

Note that this guidance was developed for documenting collections of data elements assembled for distribution or reporting. It is not intended to be used for documenting every physical database table within a system.

## Metadata Items

Caltrans has established the following recommended metadata elements to be provided for all datasets to be designated as Corporate Data Assets. These elements were selected for consistency with the guidelines in the [California Open Data Handbook](https://handbook.data.ca.gov/guidelines/) and the Federal Geographic Data Committee’s (FGDC) recommended metadata standard.

Data classification terminology (e.g., Personal Identity Information (PII), confidential information, etc.) is defined per California Department of Technology [Data Classification Standard](https://datagovernance.onramp.dot.ca.gov/downloads/datagovernance/files/References/Data_Classification_Standard-3113.pdf).

## GIS Data

Descriptions of how to enter metadata are provided in the table below. In ArcGIS Pro, the metadata required style is ArcGIS Pro Metadata Style ISO 19139 Metadata Implementation Specification GML 3.2. Go to the Project settings and click on Options to find Metadata. You can set it as your metadata style there.

Make sure to start editing Metadata by going to the Catalog Pane and right clicking the feature class in the file geodatabase. This will give you access to the metadata in the data source which enables sharing. If you access the metadata by right clicking the layer in the Contents pane of a map, it will not save in the data source and be stored when shared, pasted, or dragged to a new location.

There are three categories – Overview, Metadata, and Resource - with subcategories organizing the information within. When you click on a subcategory the data entry forms are visible.

Make sure to review the Data Dictionary ([Appendix G](#_Appendix_G:_)) as there are GIS metadata requirements listed there for the Fields.

| **Caltrans Metadata Element** | **Required?** | **Description** | **CA Open Data** | **ISO 19139 GML 3.2 (ArcGIS Pro Location)** | **Example Entry** |
| --- | --- | --- | --- | --- | --- |
| **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 | Title | Overview > Item Description > Title | SHN POSTMILES TENTH |
| **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 Topic Categories. Also include the District/Division/Program responsible. See  <https://www.fgdc.gov/metadata/documents/MetadataQuickGuide.pdf> | Tags | Overview > Item Description > Tags | Transportation, Postmiles, State Highway, District 3 - Division of Design, Division of Equipment, Sustainability Program |
| **SUMMARY** | Yes | Write a brief description one or two sentences describing this dataset. This differs from the Description (Abstract) by being less detailed and a simplified version of the Description. | Summary | Overview > Item Description > Summary (Purpose) | State Highway Network (SHN) postimles valid to the tenth (0.1 or 1/10th) of a mile based on the Office of GIS Linear Referencing System (LRS) layer. |
| **DESCRIPTION** | Yes | More detailed 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”. | Description | Overview > Item Description > Description (Abstract) | Postmiles are valid postmile points at 0.1 (1/10th) mile intervals, based on Office of GIS Linear Referencing System layer (2015, September release). Exceptions that are included are the beginning and ends of routes (and other situations) that don't necessarily begin or end on the even 0.1-mile interval. |
| **PUBLIC\_ACCESS\_LEVEL** | Yes | Public – data could 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 | Public Access Level | Overview > Item Description > Use Limitation | Restricted |
| **ACCESS\_CONSTRAINTS** | If Applicable | Can be null if PUBLIC\_ACCESS\_LEVEL is Public. Enter 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. | Rights | Overview > Item Description > Use Limitation | This dataset’s public access level is restricted and can only be shared publicly if the inspection date field is removed. |
| **TOPIC** | Yes | California Open Data Topic – choose the most appropriate topic from: <https://data.ca.gov/>. | Topic & Keywords | Overview > Topics and Keywords > Topic Categories | Transportation |
| **GIS\_THEME** | Yes – GIS Datasets only | National Spatial Data Infrastructure (NSDI) Data Theme. See <https://www.fgdc.gov/resources/whitepapers-reports/annual%20reports/2008/web-version/AppendixC.html> |  | Overview > Topics & Keywords > Theme Keywords (if none present, click + New Theme Keywords) | Transportation |
| **CREATED\_DATE** | Yes | Date that the dataset was originally created (YYYY-MM-DD) |  | Overview > Citation > Dates > Created | 2015-01-01 |
| **REVISED\_DATE** | Yes | The last date the data set was updated (YYYY-MM-DD). Set to creation date until first revision of dataset. |  | Overview > Citation > Dates > Revised | 2015-01-01 |
| **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”). | Groups | Overview > Citation Contacts > Load a contact or + New Contact –   * Organization * Role | *Organization* = Caltrans  *Role* = Publisher |
| **CONTACT\_NAME** | Yes | Name of individual that can best answer questions about the data. This is the Caltrans Business Data Steward. | Program Contact Name | \*The CONTACT is one section and this is the first box in the section.  Metadata> Contacts > Load a contact or + New Contact –   * Name | Cal T. Rans |
| **CONTACT\_ORGANIZATION** | Yes | Contact person organization (typically Caltrans unless this is an external dataset maintained by Caltrans) | Program Contact Name | Metadata> Contacts > \*Continue entering Contact info here after Name   * Organization | Caltrans |
| **CONTACT\_POSITION** | Yes | Position of individual or name of organizational unit that can best answer questions about the data. Also include the District/Division/Program responsible. |  | Metadata>Contacts> \*Continue entering Contact info here after Organization   * Position – Division (if Caltrans) | Caltrans LRS GIS Data Contact – District 3 - Division of Design, Division of Equipment, Sustainability Program |
| **CONTACT\_EMAIL** | Yes | Contact Email address | Program Contact Email | Metadata>Contacts> Contact Information or + New Contact Information \*Continue entering Contact information after Position   * Email | [cal.t.rans@dot.ca.gov](mailto:cal.t.rans@dot.ca.gov) |
| **STATUS** | Yes | Current status of this dataset – Completed, On Going, or Planned |  | Resource > Details > Status | Complete |
| **FREQUENCY** | Yes | Frequency with which changes and additions are made to the resource after the initial resource is completed – e.g. Annually, quarterly, monthly, weekly, daily, hourly, none planned.  \*when this entry is created so is a contact automatically loaded, please make sure it is correct | Frequency | Resource > Maintenance > Update Frequency  Resource > Maintenance > Load a contact (organization preferred) | Update Frequency – Annually  Load a contact – Caltrans |
| **DATA\_LIFE\_SPAN** | If Applicable | Estimated useful life span of this dataset – end of life year (YYYY) or leave blank if data has no end of life. |  | Resource > Maintenance > Maintenance Note | 2016 |
| **USE\_CONSTRAINTS** | Yes | There are 2 parts to this element:   1. Place the boilerplate disclaimer language as a Use Limitation. 2. Add another Use Limitation with 2 parts to it:    1. Enter License    2. Enter 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. | Additional Information Limitations, License | There are 2 parts to this element:   1. Resource > Constraints > + New Legal Constraints > + New Use Limitation - enter the “Boilerplate Disclaimer” in the text box 2. Resource > Constraints > + New Legal Constraints >    1. Use Constraints – click dropdown and select “Licence”    2. + New Use Limitation – enter the description of restrictions or legal prerequisites, e.g., “Creative Commons 4.0 Attribution” | 1. The data are made available to the public solely for informational purposes. Information provided is accurate to the best of our knowledge and is subject to change on a regular basis, without notice. While Caltrans makes every effort to provide useful and accurate information, we do not warranty the information to be authoritative, complete, factual, or timely. Information is provided on an "as is" and an "as available" basis. Caltrans is not liable to any party for any cost or damages, including any direct, indirect, special, incidental, or consequential damages, arising out of or in connection with the access or use of, or the inability to access or use, the Site or any of the Materials or Services described herein. 2. License - Creative Commons 4.0 Attribution |
| **METHODOLOGY** | Yes | Description of how the data was collected, produced, and processed. Note any assumptions made or data issues to be aware of. Note if data is collected from multiple sources, list who those sources are and what specific fields come from each source. | Additional Information – Data Methodology, Related Content | Resource > Lineage > Process Step > Process Description | This GIS layer was created using the 2009 Census TIGER roads file. From the 2009 TIGER data, highways were extracted and modified using BING imagery and street maps from the ESRI base map extension. The source of data is the Caltrans TSN (Transportation System Network) database. TSN is maintained by the Caltrans, DRISI, Office of Data Services and Technology |
| **DATA\_DICTIONARY** | If Applicable | Link to the data dictionary for this dataset. | Data Dictionary | Resource > Fields > Entity and Attribute Information > Details > Entity Type > Definition Source | <https://gisdata.dot.ca.gov/arcgis/rest/services/Highway/SHN_Postmiles/MapServer/0> |
| **PLACE** | If Applicable | Geographic locations characterized by the dataset, e.g. California, “Sacramento County”. | Spatial/ Geographic Coverage | Resource > Extents > Extent > Description | California |
| **TEMPORAL\_COVERAGE\_BEGIN** | If Applicable | Start date for which the dataset is applicable (YYYY-MM-DD) | Temporal Coverage, Start Date | Resource > Extents > Temporal Period Extent > Begin Date/Time |  |
| **TEMPORAL\_COVERAGE\_END** | If Applicable | End date for which the dataset is applicable (YYYY-MM-DD) | Temporal Coverage, End Date | Resource > Extents > Temporal Period Extent > End Date/Time | 2014-12-31 |
| **NEXT\_UPDATE** | If applicable | Date when the next dataset update is planned to be published (YYYY-MM-DD) |  | Resources > Maintenance > Next Update | 2016-11-18 |
| **SOURCE\_LINK** | If Applicable | Link (URL) for access to data and/or useful information about the data. | Homepage URL | Resource > Distribution > Digital Transfer Options > Online Resource > Linkage  Resource > Distribution > Digital Transfer Options > Online Resource > Function = Information | <https://postmile.dot.ca.gov/PMQT/PostmileQueryTool.html> |
| **DATA\_DICTIONARY\_TYPE** | If Applicable | 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 Dictionary Type | Resource > Fields > Entity and Attribute Information > Details > Entity Type > Definition | XML |
| **COORDINATE\_SYSTEM\_EPSG** | If Applicable | EPSG code from:  <https://spatialreference.org/ref/epsg/> |  | Resource > Spatial Reference > Reference Systems > Code  Resource > Spatial Reference > Reference Systems > Code Space = EPSG | 4326 |
| **VERT\_DATUM\_EPSG** | If Applicable | EPSG code from:  <https://spatialreference.org/ref/epsg/> |  | Resource > Spatial Reference > Reference Systems > Code  Resource > Spatial Reference > Reference Systems > Code Space = EPSG |  |
| **HORIZ\_ACCURACY** | If Applicable | Horizontal accuracy of collected data |  | Resource > Quality > Report > Report Type = Relative Internal Positional Accuracy  Resource > Quality > Report > Dimension = horizontal  Resource > Quality > Report > Measure > Description | 10 meters |
| **VERT\_ACCURACY** | If Applicable | Vertical accuracy of collected data |  | Resource > Quality > Report > Report Type = Relative Internal Positional Accuracy  Resource > Quality > Report > Dimension = vertical  Resource > Quality > Report > Measure > Description |  |
| **DATA\_STANDARD** | No | URI or URL documenting a standardized specification to which the dataset conforms. | Data Standard | Metadata > Details > Dataset URI |  |
| **NOTES** | No | Other notes about the dataset not covered by other metadata items. This item may be used to describe dataset limitations. | Additional Information | Resource > Details > Supplemental Information |  |

# Appendix F: Guidance for Defining Business Glossary Terms

A business glossary provides a central location providing shared definitions of common business terms. While a data dictionary includes descriptions of the meaning of each data element, a business glossary defines the core concepts that data may represent. For example, a business glossary might include terms like “State Highway System”, “State Highway Operations Program”, “STIP Project”, “culvert” and “gore point”.

Because there are many independent glossaries of terms maintained by Caltrans business units, building a unified Caltrans business glossary will be an incremental process. Initially, the glossary will contain those terms that are important for understanding concepts underlying designated corporate data assets.

## Corporate Business Glossary

| Data Element | Type | Required? | Value Domain and Description |
| --- | --- | --- | --- |
| ENTRY\_ID | Numeric | Yes | Auto-assigned unique identifier for entry (likely automated by SharePoint) |
| TERM\_NAME | Text | Yes | The full name of the glossary term |
| TERM\_TYPE | Single Select | Yes | Business Term (default)  Data Governance Term  IT Governance Term |
| ACRONYM | Text | No | A commonly used acronym for the term |
| TERM\_DEFINITION | Text | Yes | The meaning of the glossary term (avoid using the name of the term itself in the definition) |
| AUTHORITIVE\_REFERENCE\_NAME | Text – Lookup or enter | No | Document that provides the authoritative definition for this term.  Pick list from existing names + option to enter a new one |
| AUTHORITY\_REFERENCE\_LINK | Text | No | A link to the authoritative reference |
| AUTHORITY\_OWNER | Text | No | Organization that owns the authoritative source of this definition  Default to “Caltrans”  Pick list from existing names + option to enter a new one |
| TERM\_OWNER\_UNIT | Single Select | Yes | Organizational unit responsible for providing or updating the definition of this term  Use Business Unit Column of Caltrans Active Directory Employee List |
| ENTERPRISE\_DATA\_STEWARD | Single Select | Yes | Picklist of the EDS units – populates the EDS name |
| STATUS | Single Select | Yes | Proposed (Default)  Adopted  Rejected  Archived |
| AUTHOR | Single Select | Yes | Name of the person that proposed the term and provided the definition  Select from Caltrans Active Directory Employee List  Default to current user name |
| AUTHOR\_UNIT | Auto-Populate | Yes | Business unit of the term author  Lookup from Caltrans Active Directory Employee List |
| SYNONYMS | Text | No | A comma-separated list of synonyms for the term (including any additional acronyms) – to facilitate search in the future technology solution (use “choice option” in SharePoint) |
| COMMENTS | Text | No | To be used during glossary development to note any issues, questions or actions to be taken for particular glossary terms |
| ADD\_DATE | Date Auto-Populate | Yes | Date that this catalog entry was created |
| EDIT\_DATE | Date Auto-Populate | Yes | Date that this catalog entry was last updated |
| EDIT\_USER | Auto-Populate with AD Name | Yes | Name of person who added or last modified the entry |

## Style Guidelines for Writing Definitions

1. Don’t include the term you are defining within the definition.
2. Name the general class of objects to which the concept underlying the term belongs and identify distinguishing characteristics of the concept. (e.g. “A state highway is a roadway owned and maintained by Caltrans.”
3. Be succinct – don’t include extraneous information not essential to understanding the term.
4. Avoid unnecessary jargon – make the definition understandable to a general audience.
5. Use active voice and first person singular unless the term being defined is plural or a collective noun.

Do rely on authoritative sources for definitions (e.g. legislation or existing glossaries) but make sure the source is in the public domain before copying the definition verbatim. Paraphrase definitions that may be copyrighted.

# Appendix G: Guidance for Creating a Data Dictionary

## Purpose

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

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

If a physical database already exists, some the 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. Note that while this guidance can be used to produce a data dictionary for an entire database, its primary intent was for use in documenting the data elements within a dataset – which is defined as a collection of data elements prepared specifically for posting on a data portal or sharing with others.

## Data Dictionary Items

The following basic set of data dictionary items are to be included for documenting corporate data. Mappings to Caltrans ISO 19139 GML 3.2 geospatial metadata elements and the California Open Data elements are provided in the table below. Some elements are a crucial part of a GIS datasets Metadata and this should always be reviewed. For data in a GIS format, items marked with an asterisk are managed from within the GIS software and should not be edited using a metadata editor.

The worksheet “List of Values” must be completed if the data has defined values (whole words or codes) or uses a reference dataset as a source (can be a link to a URL) in your table that needs documenting. Some List of Values examples are city names, county abbreviations, culvert sizes, pavement types, employee classifications, etc.

| **Caltrans Data Dictionary Element** | **Required?** | **Description** | **CA Open Data** | **ISO 19139 GML 3.2** | **Example Entry** |
| --- | --- | --- | --- | --- | --- |
| **TABLE\_NAME\*** | Yes | The name of the table or dataset containing this data element. This should match with the DATASET\_NAME used for the dataset metadata. |  | Resource>Fields>Entity and Attribute Information>Details>Label | ROUTES |
| **FIELD\_NAME\*** | Yes | The name of the data element exactly as it appears in the table (or dataset column name). | Field Name | Resource>Fields>Entity and Attribute Information>Details>Attribute>Label | CARRIAGE\_FLAG |
| **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 Title | Resource>Fields>Entity and Attribute Information>Details>Attribute>Alias |  |
| **FIELD\_DESCRIPTION** | Yes | A description of the data element’s meaning and contents. Avoid obvious descriptions that don’t add new information beyond the Field Title. | Description | Resource>Fields>Entity and Attribute Information>Details>Attribute>Definition | Single character designator assigned to roadway geometry that defines whether it represents the primary (P) or secondary (S) side of the roadway as defined by the Federal Highway Administration’s Highway Performance Monitoring System guidelines. |
| **FIELD\_TYPE\*** | Yes | Choose:   * Text * Text with Formatting * Short * Long * Double * Float * Boolean (Yes/No) * Date\* * DateTime * Geometry\*\* * Binary or BLOB * GUID\*\* * Object ID\*\*   \*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> | Data Type | Resource>Fields>Entity and Attribute Information>Details>Attribute>Type | TEXT |
| **FIELD\_LENGTH\*** | Yes | Number of characters or digits needed for this data element |  | Resource>Fields>Entity and Attribute Information>Details>Attribute>Width | 1 |
| **DOMAIN\_TYPE** | If Applicable | 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   If DOMAIN\_TYPE of Range is used, UNITS, ALLOWABLE\_MIN\_VALUE and ALLOWABLE\_MAX\_VALUE must be completed.  List of Values sheet needs to be used to define values of Enumerated and Coded domains.  Items marked with an asterisk are the only domain types supported by Esri for use 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> |  | 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 | CODED |
| **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.  Enter NA otherwise. |  | Resource>Constraints>Security Constraints>Classification = Confidential | 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.  Enter NA otherwise. |  | 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) | 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.  Enter NA otherwise. |  | 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) | 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.  Enter NA otherwise. |  | 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) | NA |
| **SYSTEM\_NAME** | If Applicable | The name of the system or database containing this data element. |  | Resource>Lineage>Data Source>Source Citation>Titles>Title | Linear Referencing System |
| **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”. |  | Resource>Fields>Entity and Attribute Information>Details>Attribute>Definition Source | Federal Highway Administration |
| **FIELD\_PRECISION\*** | If Applicable | If FIELD\_TYPE is Float or Double, enter number of digits to the right of the decimal point |  | Resource>Fields>Entity and Attribute Information>Details>Attribute>Precision |  |
| **UNITS** | If Applicable | Choose:   * Units of measurement (e.g., inches, miles, feet, tons, miles per hour, inches per mile) * Currency (e.g., USD) * Percent   If UNITS is used, DOMAIN\_TYPE must be set to Range and ALLOWABLE\_MIN\_VALUE and ALLOWABLE\_MAX\_VALUE must be completed. |  | Resource>Fields>Entity and Attribute Information>Details>Attribute>Range Domain>Units |  |
| **ALLOWABLE\_MIN\_VALUE** | If Applicable | Minimum \*allowable\* value that should be accepted.  If ALLOWABLE\_MIN\_VALUE is used, DOMAIN\_TYPE of Range must be used and UNITS and ALLOWABLE\_MAX\_VALUE must be completed. |  | Resource>Fields>Entity and Attribute Information>Details>Attribute>Range Domain>Minimum |  |
| **ALLOWABLE\_MAX\_VALUE** | If Applicable | Maximum \*allowable\* value that should be accepted.  If ALLOWABLE\_MAX\_VALUE is used, DOMAIN\_TYPE of Range must be used and UNITS and ALLOWABLE\_MIN\_VALUE must be completed. |  | Resource>Fields>Entity and Attribute Information>Details>Attribute>Range Domain>Maximum |  |
| **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. |  | Resource>Fields>Overview Description>Summary |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **List of Values Worksheet** |  |  |  |  |  |
| **Caltrans Data Dictionary Element** | **Required?** | **Description** | **CA Open Data** | **ISO 19139 GML 3.2** | **Example Entry** |
| **SYSTEM\_NAME** | If Applicable | The name of the system or database containing this data element. |  | Resource>Lineage>Data Source>Source Citation>Titles>Title | Linear Referencing System |
| **TABLE\_NAME\*** | Yes | The name of the table or dataset containing this data element. This should match with the DATASET\_NAME used for the dataset metadata. This is auto populated in ArcGIS Pro. |  | Resource>Fields>Entity and Attribute Information>Details>Label | ROUTES |
| **FIELD\_NAME\*** | Yes | The name of the data element exactly as it appears in the table (or dataset column name). This is auto populated in ArcGIS Pro. |  | Resource>Fields>Entity and Attribute Information>Details>Attribute>Label | CARRIAGE\_FLAG |
| **CODE** | If Applicable | Word, letters, or numbers used to shorten a value |  | Resource>Fields>Entity and Attribute Information>Details>Attribute>Enumerated/Codeset Domain | P |
| **CODE\_DEFINITION** | If Applicable | Full value that has been shortened into a code. |  | Resource>Fields>Entity and Attribute Information>Details>Attribute>Enumerated>Definition | Primary |
| **SOURCE\_NAME** | If Applicable | The name of the source of the list of values. |  | Resource>Fields>Entity and Attribute Information>Details>Attribute>Codeset Domain>Name | USPS Publication 28 - Postal Addressing Standards |
| **SOURCE\_LINK** | If Applicable | Link (URL) to documentation or other pertinent information related to the list of values. |  | Resource>Fields>Entity and Attribute Information>Details>Attribute>Codeset Domain>Source | https://pe.usps.com/text/pub28/welcome.htm |

# Appendix H: Guidance for Specifying Business Rules

## Purpose

Business rules provide the foundation for data quality management. They can be used to build in data validation logic within data entry applications, and to develop procedures for data validation and cleansing once data are already in hand. They also are used to provide guidance for manual data review.

Data entry and validation logic for individual data elements should be based on information provided in the data dictionary. This information includes allowable and expected ranges (min and max) for numeric and date data elements, and lists of values for enumerated and coded data elements. Usage notes for data elements should also be included in the data dictionary to supplement the more general-purpose field descriptions, and provide extended technical documentation of complex logic for calculated, composite or overloaded fields.

Business rules that go beyond validation of individual data element values that should be covered within the business rule catalog are described below.

## Types of Business Rules

The following types of business rules are to be considered and included in the business rules catalog:

|  |  |
| --- | --- |
| **Rule Type** | **Rule Description** |
| **Conditional Rules** | Conditional rules applied to one data element based on values of other data elements. |
| **Geo-Processing and Linear-Referencing Rules** | Rules applied to a data element or the record itself that utilize the location of a record to validate data entry. These rules may validate data entry or record creation based on a GIS feature or information embedded within other common location referencing information. |
| **Cardinality Rules** | Rules applied to and describing the relationships between the data entities within a given dataset (one to one, one to many, etc.). These rules are used to inform or validate the relational model. |
| **Dataset Level Rules** | Rules applied to the dataset at large, typically applied to aggregation of values across records for comparison against other attributes of the dataset or an expected value. |
| **Usage Notes/Descriptions** | Details of calculations or other business logic that is too detailed to include in the data dictionary but important for understanding and validating the data element. |

## Business Rules Catalog Contents

The following items are documented for each rule in the catalog:

| Section | Item | Contents |
| --- | --- | --- |
| ID | BR\_ID | A unique identifier for the business rule that can be used for cross-referencing. |
| Rule Scope | SYSTEM | The name of the application, system or database being documented (if applicable) |
|  | DATASET\_NAME | The name of the table or dataset being documented. Use the actual physical dataset name (if available) or a descriptive name for a future dataset being planned. This should match the DATASET\_NAME in the data catalog and dataset metadata file. |
|  | DATA\_ELEMENT | The name of the data element being documented. Use the actual physical field name if you are documenting a data element within a single existing table. If you are documenting a future data element or a data element that appears across multiple sources, use a descriptive name instead. |
|  | ASSOCIATED\_ELEMENT | The name(s) of the associated data elements referenced in the rule (for conditional rules or cardinality rules). Use the actual physical field names if you are referencing data element(s) within a single existing table. If you are referencing future data element(s) or data element(s) that appear across multiple sources, use descriptive name(s) instead. |
| Rule Description | RULE\_TYPE | Type of business rule (one of the predefined types) |
|  | RULE\_DESCRIPTION | Brief (non-technical) description of what the rule says |
|  | POLICY | Yes – if the rule is based on a documented agency policy or legislation  No - Otherwise |
|  | CALCULATION | Yes – if the business rule provides documentation of a formula used to compute the value of the evaluated data element  No- Otherwise |
| Rule Specification – Geoprocessing/ LRS | GEO\_BUFFER | For Geoprocessing rules - buffer distance to be used for applying geoprocessing for data validation  For other rules – enter NA |
|  | GEO\_LRM | For Linear Referencing rules – Linear Referencing Method to be used for validation  For other rules – enter NA |
| Rule Specification – Detailed Business Logic | LOGIC\_DESCRIPTION | A detailed specification of the logic of the rule – suitable for translation into code by a programmer. If too lengthy for documentation within a spreadsheet cell, indicate “See Data Documentation Package” and provide the specification within this document. |
| References | REFERENCES | Links to document(s) with backup information for the rule – e.g. policy documents, user manuals, etc. |

## Examples

### Conditional Business Rule Examples

* **Pavement Roughness (IRI)** – building from the Simple Business Rule examples, consider a dataset which includes Pavement Roughness information collected across the entire Caltrans network. If this dataset included a Functional Classification assigned based on the individual sections collected, Pavement Roughness limits could be assigned based on the Functional Classification.

For example, Interstate pavement may be expected to be paved under ideal conditions, with the highest quality materials and equipment, and achieve the highest performance, while Local roads may have many constraints that impact the ability to achieve and maintain a smooth ride. Other classifications may be realistically expected to perform at some intermediate performance levels.

Given this expectation for varied performance by Functional Classification, different minimum and maximum thresholds for the expected Roughness values could be identified. The rules template allows this to be identified within the “Other Rules” item and then the detailed business logic provided in the “Detailed Rule Business Logic” item.

Assuming the functional classification is trusted, this data validation applies to the Pavement Roughness measurement, and is based on the Functional Classification value assigned to that record. In this example, Expected Ranges for IRI are set by Functional Classification:

|  |  |
| --- | --- |
| **Rule Attribute** | **Rule Attribute Value** |
| Evaluated Data Element(s): | Pavement Roughness |
| Associated Data Element(s): | Functional Classification |
| Allowable Value Rule: | NA |
| Expected Value Rule: | [Pavement Roughness] on Interstate pavement should be flagged if less than 40 or greater than 200  [Pavement Roughness] on Local pavement should be flagged if less than 40 or greater than 500  [Pavement Roughness] on pavement of other functional classifications should be flagged if less than 40 or greater than 350 |

* **Task Status and Task Percent Complete** – building from the Simple Business rule examples, consider a project tracking dataset which includes Task Status and Task Percent Complete.

Based on the Status the Percent Complete can be validated and vice versa. Given that both data elements come from the same source, it may be that either or both may be in error if they are not consistent, so the business rule should apply to both data elements:

|  |  |
| --- | --- |
| **Rule Attribute** | **Rule Attribute Value** |
| Evaluated Data Element(s): | Task Status; Task Percent Complete |
| Associated Data Element(s): | NA |
| Allowable Value Rule: | Flag if [Task Status] is “Not Started”, and [Task Percent Complete] is greater than 0%  Flag if [Task Status] is “Underway”, and [Task Percent Complete] is 0% or 100%  Flag if [Task Status] is “Complete”, and [Task Percent Complete] is less than 100% |
| Expected Value Rule: | NA |

### Geo-Processing and Location-Based Rule Examples

* **Pavement Roughness Functional Classification** – by example of the previously discussed Pavement Roughness dataset, the Functional Classification assigned to a Pavement Roughness record could be validated against official Functional Classification information through geo-processing or through comparison through another shared location referencing approach.

*Geo-Processing Example*

| **Rule Attribute** | **Rule Attribute Value** |
| --- | --- |
| Evaluated Data Element(s): | Functional Classification |
| Associated Data Elements: | Functional Classification feature class in the Production GIS System |
| Analysis Approach | Geo-Processing |
| Rule General Description | Compare the Pavement Roughness line feature against the Functional Classification feature class to identify all Functional Class features within the buffer distance which are on the same route. Select the highest functional class (Interstate is highest, Local is lowest) from the identified Functional Class features. Compare the selected Functional Class value against the Pavement Roughness functional classification. |
| Buffer Distance | 50 feet |
| Allowable Value Rule: | NA |
| Expected Value Rule: | Flag the Pavement Roughness record’s [Functional Classification] if it does not match the [Functional Classification] selected through the geo-processing |

*Linear Referencing Example*

|  |  |
| --- | --- |
| **Rule Attribute** | **Rule Attribute Value** |
| Evaluated Data Element(s): | Functional Classification |
| Associated Dataset(s) and System(s): | Functional Classification - enterprise dataset |
| Analysis Approach | Linear Referencing |
| Rule General Description | Compare the Pavement Roughness against the Functional Classification dataset using the Route, Direction, and Caltrans Postmile to identify all Functional Class data that overlaps the Pavement Roughness record location. Select the highest functional class (Interstate is highest, Local is lowest) from the identified Functional Class records. Compare the selected Functional Class value against the Pavement Roughness functional classification. |
| Buffer Distance | NA |
| Allowable Value Rule: | NA |
| Expected Value Rule: | Flag the Pavement Roughness record’s [Functional Classification] if it does not match the [Functional Classification] selected through the analysis. |

### Cardinality Rule Examples

* **Pavement Roughness** – it is common for a DOT to identify a set of management section records against which Pavement Roughness and other condition data are collected on a cycle. It may be that each Pavement Roughness record should be assigned to Management Section, and that a Management Section should have zero or more Pavement Roughness measurements over time.

Based on these expectations, cardinality rules can be documented

|  |  |
| --- | --- |
| **Rule Attribute** | **Rule Attribute Value** |
| Evaluated Entity(s): | Pavement Roughness |
| Associated Entity(s): | Pavement Management Section |
| Relationship General Description: | Pavement Roughness records must be assigned to a specific Pavement Management section. A Pavement Management section may be assigned to multiple Pavement Roughness records. |
| Allowable Value Rule | Flag all roughness records which do not have a related pavement management section |
| Expected Value Rule | NA |

* **Projects and Tasks** – building from the previous example, a project should have 1 or more tasks, while a task must always be assigned to a specific project.

Based on these expectations, cardinality rules can be documented

**Project Rule**

|  |  |
| --- | --- |
| **Rule Attribute** | **Rule Attribute Value** |
| Evaluated Entity(s): | Project Information |
| Associated Entity(s): | Task Information |
| Relationship General Description: | A project should have 1 or more tasks (unless in early stages of data entry and/or project scoping) |
| Allowable value Rule | NA |
| Expected Value Rule | Flag all project records which do not have a related task |

**Task Rule**

| **Rule Attribute** | **Rule Attribute Value** |
| --- | --- |
| Evaluated Entity(s): | Task Information |
| Associated Entity(s): | Project Information |
| Relationship General Description: | A task must be assigned to a specific project |
| Allowable Value Rule | Flag all task records which do not have a related project |
| Expected Value Rule | NA |

### Dataset Level Business Rule Examples

* **Pavement Roughness** – many DOT’s evaluate their pavement network based on a planned collection schedule. Assuming the DOT has a monthly collection schedule, and that collected data should not take more than three months to uploaded into the system, the following dataset level quality rule could be identified.

|  |  |
| --- | --- |
| **Rule Attribute** | **Rule Attribute Value** |
| Evaluated Data Element(s): | Pavement Roughness Length |
| Associated Data Elements: | Collection Date, Collection Schedule |
| Rule General Description: | For the current collection year, sum mileage  for all months outside of the most recent three months (as these records may not be fully processed into the systems) and compare the total collected mileage against the established collection schedule for the summarized months. |
| Allowable Value Rule: | NA |
| Expected Value Rule: | Flag the Pavement Roughness dataset if the collected mileage is less than 80% of the scheduled collection mileage. |

### Usage Notes or Description Examples

* **Pavement Roughness** – as part of many data collection efforts, “Intelligent IDs” may be automatically generated which contain valuable information that can be decoded from the data element. For example, Pavement Roughness is typically collected by a road profiler, which may have a unique name designating the particular vehicle. This unique designation may be combined with the date, time, collection project name, speed of collection, or other information that might be used to generate a unique identifier of the collection. This data element may be embedded within the dataset and can be explained for those who may be evaluating the dataset.

| **Usage Attribute** | **Usage Attribute Value** |
| --- | --- |
| Evaluated Data Element(s): | Pavement Roughness ID |
| Rule General Description | The Pavement Roughness ID is an intelligent identifier which has embedded business meaning in the ID assigned |
| Usage Details | This 20 character number is comprised of the following information:   * Characters 1-3: Road Profiler Unique ID * Characters 4-6: Operator Unique ID * Characters 7-11: Collected Route Number * Characters 12-13: Collection Start MP * Characters 14: Collected Lane * Characters 15: Collected Lane Direction * Characters 16-17: Collection Average Speed * Characters 18-20: Special Route Identifiers |

1. Many Source Databases that are used to maintain data are also used for reporting. Only databases that are created specifically to integrate and transform data from other source systems should be categorized as Reporting Databases. [↑](#footnote-ref-2)
2. Note that the California Open Data Handbook uses a different definition of dataset: “*any organized collection of data. The most basic dataset is composed of data elements in a table. Each column represents a particular variable. Each row corresponds to a given value of that column’s variable. A dataset may also present information in a variety of non-tabular formats, such as an extended mark-up language (XML) file, a geospatial data file, or an image file. Dataset is a flexible term and may refer to an entire database, a spreadsheet or other data file, or a related collection of data resources.” The Caltrans definition is not inconsistent with this, but is intended to make a clear distinction between datasets designed for sharing or distribution and databases (and their constituent tables) are part of transactional systems and not suitable for sharing in their current form.* [↑](#footnote-ref-3)