

Data Quality Management Plan for:

<Dataset>

Version XXX

<Business Unit>

10/28/2019

Contents

1. [Introduction](#_bookmark0) [3](#_bookmark0)
   1. [Background](#_bookmark1) [3](#_bookmark1)
   2. [What is Data Quality?](#_bookmark2) [3](#_bookmark2)
   3. [How Good is Good Enough?](#_bookmark3) [4](#_bookmark3)
   4. [Data Quality Management Strategies](#_bookmark4) [4](#_bookmark4)
   5. [Document Purpose](#_bookmark5) [10](#_bookmark5)
2. [Data Quality Management Plan](#_bookmark6) [12](#_bookmark6)
   1. [Version Control and References](#_bookmark7) [12](#_bookmark7)

[DQMP Version](#_bookmark8) [12](#_bookmark8)

[DQMP Contributors](#_bookmark9) [12](#_bookmark9)

[Supporting Documentation](#_bookmark10) [12](#_bookmark10)

* 1. [Data Quality Objectives](#_bookmark11) [13](#_bookmark11)
  2. [Current State Assessment](#_bookmark12) [13](#_bookmark12)

[Current Data Quality](#_bookmark13) [14](#_bookmark13)

[Current Data Quality Management Practices](#_bookmark14) [14](#_bookmark14)

* 1. [Quality Improvement Strategies](#_bookmark15) [15](#_bookmark15)

[Current Sources of Errors](#_bookmark16) [15](#_bookmark16)

[Data Quality Improvement Strategies](#_bookmark17) [15](#_bookmark17)

* 1. [Tracking and reporting](#_bookmark18) [16](#_bookmark18)

[Tracking](#_bookmark19) [16](#_bookmark19)

[Reporting](#_bookmark20) [17](#_bookmark20)

# Introduction

## Background

Caltrans has adopted a set of core data principles, based on those established by the American Association of State Highway and Transportation Officials (AASHTO). These principles recognize that data is an asset that needs to be maintained – much like our physical assets. Good stewardship of data involves ensuring that data are of sufficient quality to serve its intended purposes.

Data are used in all areas of the transportation decision-making process from planning to design to operations to performance management. Furthermore, data is increasingly being used externally by citizens and customers to inform their personal decisions, and by stakeholders to assess the aggregate performance of a transportation organization. Significant human and system resources are consumed in the collection, manipulation and dissemination of data whether of high quality or not, so it is essential that the most effective use of public funds is achieved through appropriately directed attention to data quality and the procedures to realize quality.

## What is Data Quality?

In general, “quality data” means data are fit for their intended purposes. There are several aspects of data quality, including accuracy, timeliness and completeness. Data may be highly accurate but produced too late to be of value to users. Conversely, data may be provided in real time, but lack reliability. Usability – the ease with which an end user can access and make use of the data to meet their need – is also an important consideration. Data may be accurate and timely, but if it is not accessible or requires specialized expertise to transform into a useable form, it is not meeting user needs. The following table lists dimensions of data quality to be considered in a data quality management plan (DQMP).

#### Table 1. Data Quality Dimensions

|  |  |
| --- | --- |
| **Data Quality Description**  **Dimension** | |
| **Accuracy and Precision** | Accuracy is the extent to which the data provide an unbiased representation of the true value. Precision is a measure of exactness – or closeness of multiple observations to each other. For spatial data, Positional Accuracy is an important characteristic and is typically expressed as a probability that observations are within some distance of their true location or their location as recorded in an independent source (e.g. 95% of data points are within 5 feet of their actual or  reference location). |
| **Validity** | The degree to which data conforms to established formats, data types, and value ranges that can be formally specified – as well as the degree to which data  matches what is expected based on business knowledge and context. |
| **Completeness** | The absence of gaps in data – including gaps in coverage of the target scope as well as gaps in the content of individual records (missing values for required data  elements). |
| **Consistency** | For a given dataset, consistency is the degree to which data records collected at different points in time or for different locations can be compared and aggregated because they have the same collection methodologies, formats and  meanings. |

|  |  |
| --- | --- |
| **Data Quality Description**  **Dimension** | |
| **Currency and Timeliness** | How old the data are, which is a function of how frequently they are collected or updated, and the amount of time required for processing and publication. |
| **Granularity** | The extent to which data are provided at the right level of detail to meet the needs. This is related to the level of measurement precision as well as the definition of what each record represents (e.g. an entire bridge versus each  element of a bridge). |
| **Uniqueness / Non-repudiation** | The extent to which there is a single authoritative source for each data fact of interest, and information derived from different data sources is not contradictory (e.g. two databases with information about active projects that have different numbers of projects). |
| **Accessibility** | The degree to which the data are made available in a useable form (e.g. machine-  readable, mappable, documented) |
| **Reputation** | The extent to which data is trusted. For example, a dataset from an external source that is perceived to have a bias may not be trusted even if it passes formal validation checks. |

## How Good is Good Enough?

While people always want the data they use to be as good as possible, no dataset is perfect – and higher levels of data quality are costly to achieve. Improving data quality may mean more detailed data collection, more frequent updating, independent validation processes, and sophisticated error checking procedures. Determining the level of data quality that is appropriate requires judgement about how good is good enough given how the data will be used. Risks of poorer data quality must be considered together with costs to improve data quality.

## Data Quality Management Strategies

The following strategies should be considered to improve the quality of existing data, and to ensure quality of new data collected or assembled. Not all of these strategies will be applicable in every situation and this list is not intended to be comprehensive.

#### Table 2. Example Data Quality Management Strategies

|  |  |
| --- | --- |
| **Category Data Quality Management Strategy** | |
| **Data Specification** | Document meaning of each data element |
|  | Document data collection procedures |
|  | Produce field manual for data collection personnel |
|  | Specify valid ranges and value domains for coded elements |
|  | Define business rules that can be used for data validation |
| **Data Quality Analysis** | Data profiling to understand current level of completeness and validity |
|  | Root cause analysis to understand sources of errors |
| **Data Collection/ Entry/**  **Import Quality Control** | Data collection personnel training and certification |
|  | Independent verification (e.g. separate team independently collects data for  5% of sites and compares values to main data collection team) |
|  | Validation checks/import checks |
|  | Built-in data entry restrictions (pick lists) and validations |

|  |  |
| --- | --- |
| **Category Data Quality Management Strategy** | |
|  | Equipment inspection and certification |
|  | Equipment calibration |
|  | Data review by manager or subject matter expert (SME) and feedback |
|  | Formal data acceptance criteria and feedback |
| **Data Quality Assurance** | Manual review of sample records (sample size, frequency, checks performed) |
|  | Manual review of mapped records (sample size, frequency, checks  performed) |
|  | Manual review of aggregate results (frequency, checks performed) |
|  | Manual review of current against prior observations (sample size, frequency, checks performed) |
|  | Automated validation scripts/cleaning |
|  | Data exporting and reporting validation |
|  | Formal sign-off by manager on data quality |

* 1. 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):

Associated Data Element(s):

Task Status; Task Percent Complete 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):

Associated Data Elements:

Functional Classification

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):

Associated Dataset(s) and System(s):

Functional Classification

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):

Associated Data Elements:

Pavement Roughness Length 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. Document Purpose

This document provides a framework for assessment, documentation and improvement of data quality management practices for Caltrans for any data set or system. The provided structure allows for describing existing conditions, a desired to-be state, and the implementation planning of gap resolution regarding quality checks performed during the collection, entry, maintenance and reporting of data as well as issue reporting and overall communication of data quality.

In this template:

* + - <instructions are shown in blue, enclosed in brackets>
    - *examples are shown in green italic*

When completing this template, delete the instructions and examples and fill in information for your selected dataset(s).

# Data Quality Management Plan

This data quality management plan describes the specific steps taken to ensure quality and fitness of use of <DATA SET/SYSTEM NAME> managed under the <OFFICE NAME> within the <DIVISION NAME>. The plan has the following sections:

* Version control and references
* Data quality objectives
* Current state assessment
* Quality improvement strategies
* Tracking and reporting

## Version Control and References

### DQMP Version

<Use this table to keep track of the initial (baseline) version of the plan and subsequent revisions>

|  |  |  |
| --- | --- | --- |
| **Version** | **Date** | **Summary of Work** |
| *1.0* | *01.01.2020* | *Baseline version* |
| *1.1* | *01.01.2021* | *Updated tracking and reporting* |
|  |  |  |
|  |  |  |

### DQMP Contributors

<Provide the name(s) and contact information for the authors and contributors to this document*.>*

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Business Unit Name and District/Division** | **Email** | **Phone** |
| *Miranda Pool* | *OMI/DRISI* | [*mpool@dot.ca.gov*](mailto:mpool@dot.ca.gov) | *999-888-7777* |
| *Jamie Klein* | *OMI/DRISI* | [*jklein@dot.ca.gov*](mailto:jklein@dot.ca.gov) | *888-111-2222* |
|  |  |  |  |
|  |  |  |  |

### Supporting Documentation

<Provide references to supporting documentation including:

* Data collection/field manual
* Data processing procedures
* Dataset metadata
* Data dictionary
* Business rules
* Edit checks/Validation processes
* Applicable standards or requirements>

|  |  |  |
| --- | --- | --- |
| **Document** | **Link/Location** | **Comments** |
| *Data Dictionary* | *https://dot.ca.gov/data/xxx* | [*created*](mailto:mpool@dot.ca.gov) *10 years ago, needs updating* |
| *Data Collection Manual* | *See Miranda Pool* | *manual used by data collectors to ensure consistency* |
|  |  |  |
|  |  |  |

## Data Quality Objectives

This section documents the intended purposes of the dataset(s) covered under this plan and establishes objectives to be achieved to serve these purposes.

*Intended Users and Uses*

<Describe who will be using the data and for what purpose. There is no need to be comprehensive or exhaustive, but attempt to capture the major categories of users and what they will be doing with the data. Include any external (federal or state) requirements for which the data are used.>

|  |  |  |
| --- | --- | --- |
| **Use** | **Caltrans User** | **External User(s)** |
| *HPMS Reporting* | *OHSIP* | *FHWA* |
| *Traffic Analysis* | *District Traffic Operations* | *Consultants* |
|  |  |  |
|  |  |  |

*Objectives*

<List data quality objectives related to the different data quality dimensions listed in Table 1. These objectives should be based on consideration of user needs as well as what can be realistically achieved given costs and resourcing for data quality improvement. Data quality objectives should ideally be measurable and reportable. Data quality objectives should be at an appropriate level of detail for target users of the dataset(s) to understand and validate.>

#### Table 3. Data Quality Objectives

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Quality Dimension** | **Objective** | **Measurement/ Tracking Method** | **Comments** |
| *Validity* | *All numeric values should fall within their specified min-*  *max values* | *Data validation script* |  |
| *Completeness* | *The data set should cover no less than 95% of the state highway system’s lane mileage, as reported in the most recent annual HPMS*  *report.* | *Data validation script* |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## Current State Assessment

This section documents the current level of data quality and the current procedures in place for data quality control and quality assurance.

### Current Data Quality

<Document your understanding of the current quality of the dataset(s) **using the objectives listed above**. This may require development of data validation scripts and/or use of available data profiling tools to complete.>

#### Table 4. Current Data Quality Assessment

|  |  |  |
| --- | --- | --- |
| **Data Quality**  **Dimension** | **Current State** | **Comments** |
| *Validity* | *Unknown* | *Business rules for minimum and maximum values*  *have not yet been specified.* |
| *Completeness* | *The dataset covers 90% of the state highway system lane*  *miles.* |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

### Current Data Quality Management Practices

This section documents current practices in place for data quality management to provide a baseline for identifying improvements.

<Refer to Table 2 for a list of example data quality management practices. For each practice that is currently in place, describe who performs it and what the deliverables or results are.>

#### Table 5. Current Data Quality Management Practices

|  |  |  |
| --- | --- | --- |
| **Practice** | **Who Does it** | **When is it Done** |
| *Maintain a data collection manual* | *Business Data Steward* | *Updated annually based on feedback from quality assurance staff* |
| *Identify and report data anomalies* | *Business Data Steward and All*  *Users* | *Verified on an ongoing basis* |
| *Check DMI calibration for Automated Pavement Condition Survey data collection equipment.* | *Contractor provides test results and certification paperwork to the State Pavement*  *Engineer* | *Testing and calibration is performed every 30 days* |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Quality Improvement Strategies

This section documents current sources of errors and strategies for improving current data quality and preventing future errors from occurring.

### Current Sources of Errors

<Document current sources of data quality issues and possible ways to address them. Consider the following categories:

* + - Insufficient training or guidelines for personnel performing data collection, processing or entry tasks
    - Equipment reliability or calibration issues
    - Manual data entry
    - Lack of data entry controls (pick lists and built in validation)
    - Lack of formal validation procedures
    - Lack of data standards
    - Changes to data specifications or collection methods over time
    - Lack of feedback mechanisms allowing data users to report issues.>

#### Table 6. Current Sources of Error

|  |  |  |
| --- | --- | --- |
| **Source of Error** | **How Errors Occur** | **Possible Improvements** |
| *Data maintained in flat files* | *Manual transfer processes not always performed correctly* | *Switch to managing data in a database and automate*  *transfer process* |
| *No validation on entry of project number* | *Project number may be left blank, or mis-keyed* | *Incorporate validation of project number into the data entry program* |
| *High turnover among data collection staff* | *Lack of experience leads to inconsistent interpretation of*  *instructions* | *Data collector training and certification* |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

### Data Quality Improvement Strategies

<Two types of data quality improvement strategies should be considered: (1) fixing existing errors in the data (e.g. filling in missing values or correcting mis-coded items) and (2) preventing future data quality issues from occurring by addressing the sources of error listed in Table 6. See Table 2 for example data quality management strategies. Based on priorities and resource availability, create a list of **new** data quality management practices that you intend to implement.>

#### Table 7. Planned Data Quality Management Practices

|  |  |  |
| --- | --- | --- |
| **Practice** | **Who Does it** | **When is it Done** |
| *Data collector training and certification program* | *Vendor to do training; OMI to ensure certification* | *Prior to every data collection cycle* |
| *Develop and run data validation scripts based on business rules (e.g. LRS geometry cannot have loop-backs, PROJ\_ID*  *cannot have NULL values)* | *Business Data Steward and Custodian* | *1Integrate checks to be performed monthly to confirm data is in compliance. All non- compliant data will be corrected within 10 working days.* |
| *Perform cross-checks of mileage totals against*  *multiple sources* | *Business Data Steward and*  *Custodian* | *Cross-checks are performed in May of each year* |
| *Manually review all observations that have changed over 25% from*  *the prior year* | *Business Data Steward* | *Review is performed as each batch of data is delivered.* |
|  |  |  |
|  |  |  |

## Tracking and reporting

This section documents processes to be followed to track implementation of new data quality management strategies, and to inform users about current data quality.

## Tracking

<Create a list of trackable implementation steps to put the planned data quality management practices in place.>

#### Table 8. Data Quality Management Plan Implementation Steps

|  |  |  |
| --- | --- | --- |
| **S**tep | **L**ead | **T**arget Completion |
| *Create business rules* | *John Smith* | *Jan 31, 2020* |
| *Validate business rules*  *with SMEs* | *John Smith* | *Feb 15, 2020* |
| *Implement and test 1Integrate scripts* | *Mary Jones* | *Feb 28, 2020* |
| *Create report for cross-*  *check of mileage* | *Andrew Cooper* | *Feb 28, 2020* |
| *Create report for comparison with prior year data* | *Andrew Cooper* | *March 15, 2020* |
| *Initial run of validation* | *Mary Jones* | *March 15, 2020* |
|  |  |  |
|  |  |  |

## Reporting

<Describe how you will report progress on implementation of planned data quality improvement strategies.>

Responsible Manager to Receive Reports: *Fred Flintstone, Chief, Rock Division*

Reporting Interval: *Annual* – by July 31

<Describe how and where you intend to document current data quality for data users.>

* Verbally – on request
* Data quality report posted with data
* Data quality description included in dataset metadata (Notes element)
* Other: