Recommendations and Guidelines for Data Dictionary Development

This document provides a set of generic guidelines recommended for use in the development of a data dictionary for any dataset managed by the California State Water Resources Control Board (Water Board). These recommendations were developed by the Water Board's Office of Information Management and Analysis (OIMA), and are intended to help the State Board's data managers understand the purpose and structure of a data dictionary while also encouraging a level of consistency of metadata structure and contents across datasets.

A data dictionary is a part of a dataset's metadata which provides a description of each of the fields or variables contained within that dataset. Its purpose is to thoroughly document the dataset and provide consumers, suppliers, managers, and/or administrators of the data with specific information about each variable that they may need in their given role (note that a data dictionary describes the repository of the data, but not necessarily the interface used to collect it, so other types of documentation may be required to describe and manage the interface).

The format and level of detail contained in the data dictionary is dependent on the context and intended audience for which it is being developed. For example, for open data applications, the intent is to provide, to the greatest extent possible, relatively simple and concise information which helps data consumers understand how to interpret and analyze the data appropriately. In addition, open data applications often have standardized templates and protocols for metadata (including data dictionaries) to provide a user-friendly interface for data consumers. In contrast, data dictionaries developed for an internal or non-public use (such as documentation for data submitters, dataset managers, and/or data system administrators) may contain additional levels or types of detail, and may be better served by a more customized format. As a result, this guide first describes the fields that are required of all data dictionaries, which provide a sufficient level of detail for a generic open data application (including the California Open Data Portal, at data.ca.gov), then describes an additional set of example optional fields which could be included in data dictionaries intended for other applications. It is up to the developer of the data dictionary to define the audience and context in which the data dictionary will be applied, and then use this information to select the appropriate fields and format for their intended application.

The State Water Board's Office of Information Management and Analysis (OIMA) can assist other Divisions, Offices, or Regions in the process of determining the appropriate format for a data dictionary, as well as assisting in developing and maintaining any workflows related to publication of data dictionaries for open data applications. However, the contents of a data dictionary should be populated primarily by the staff who are most familiar with that dataset, which may include staff who manage the dataset as well as those who regularly use the data. These staff will be best positioned to understand the factors that help provide context for the dataset and explain its intended uses, such as the drivers behind the collection of that particular data, details about the way the data in each field is generated, and potential issues with or limitations of the data in any field.

Required Fields

These are the fields that are required at a minimum for open data applications, and are compliant with most open data protocols (although the field names and order may differ, depending on the platform where the data is published). For applications not intended for open data publication, additional fields may be added as needed, and the information contained in the 'Field Description' can be split into multiple fields as described in the Optional Fields section below.

For data dictionaries describing datasets published to the California Open Data Portal, note that only the first four fields (Column, Label, Type, and Description) are accepted by the portal. Therefore, if you are developing a data dictionary solely for the purpose of publishing data to the California Open Data Portal, then you should only use the fields contained in this section for your data dictionary (note that while the '*Data Categorization'* field is not used by the California Open Data Portal, it is required for the Waterboards internal documentation and approval processes). Alternatively, if you would like to develop a data dictionary for multiple uses and audiences, and one of those uses includes publishing to the California Open Data Portal, you will need to be prepared to combine information from the relevant fields such that all of the information needed for an open data application can be contained within these four fields. Similarly, if you have a pre-existing data dictionary in a different format that you would like to use for publishing data to the California Open Data Portal, you will need to combine information from the relevant fields in your existing data dictionary such that the relevant information is contained in those four fields.

The required fields are:

- **Column**: the field name used in the data table, which may be abbreviated or truncated (note that when populating a data dictionary on the California Open Data Portal, the column name is automatically generated based on the input data file, and is not editable by the user)
- **Label**: the common name of the field, in plain English, without abbreviations, acronyms, or other truncated names
- **Type**: the type of data contained in the field, which describes the way the field's records are stored, and defines the meaning of the data and the types of operations that can be done on it; at present, the California Open Data Portal only accepts the following three data types (however, for data dictionaries developed for other applications, other types may be included):
    - *Text*: fields that include any type of data which can't (or shouldn't) be specified as any other type - for example, this can include plain text, fields with a mix of

    different data types (varchar), numeric values which aren’t intended for numeric operations (such as IDs), etc.
    - *Numeric*: fields in which all records can be recognized as a number, and for which some type of numeric operation (e.g., statistical summarization, numeric filtering, etc.) could logically be performed
    - *Timestamp*: fields containing a combination of a calendar date and time, in a standard format (e.g., YYYY-MM-DD HH:MM:SS); note that for publication to the California Open Data Portal, fields which only contain a date without any related time component can still be treated as timestamp type
- **Description**: a description of the information contained in the field, as well as any notes that provide context to help suppliers, maintainers, and users of the data understand how it can be interpreted, managed, and used appropriately - for example, this could include:
    - notes about the data source (e.g., if the records are calculated, a description of the calculation methods and source data)
    - known issues and limitations that users should take into account when analyzing the data, such as issues that could limit the types of operations that can be carried out on the data, limit the confidence of any conclusions drawn from the data, or ways that the data should be processed prior to use
    - valid values (e.g., a list of valid values for a list-defined entry, a list of all possible categories or levels for categorical data, or the possible range of values for continuous data) or examples of the data format
    - any additional information that provides useful context about the information contained in the field
- **Data Categorization**: a description of the sensitivity of the information contained in the field and limitations on its disclosure (see the ‘*Data Publication Sign-Off’* document for more information) - options include:
    - *Public Information*: any information not exempt from disclosure, per the California Public Records Act
    - *Confidential Information*: any information that is exempt from disclosure under the California Public Records Act. Examples include business trade secrets, enforcement actions, etc.
    - *Sensitive Information*: requires special precautions to protect from unauthorized use, access, disclosure, modification, loss, or deletion
    - *Personally Identifiable Information (PII)*: Information that could potentially be used to identify an individual, such as Name, Former Name, or Alias, Date of Birth, full or truncated Social Security Number (SSN), etc. (see the see the ‘*Data Publication Sign-Off’* document for more examples)

Optional Fields

These are examples of fields which could be included in applications for internal use (e.g., use by data submitters, dataset managers, and/or data system administrators). Many of these fields are split from the ‘Field Description’ field described in the Required Fields section above, to allow for a greater level of detail and a more user-friendly format. This is not an exhaustive list, and some of these fields may not be applicable to all datasets. Some optional fields include:

- **Valid Values / Examples**: a list of all possible categories or levels, or the possible range of continuous values, and/or examples of the data format
- **Known Issues**: any caveats that should be taken into account by users or managers of the data
- **Condition / Source:** the way the data is generated, for example provided by a data submitter through a form, auto-generated by a data system (e.g. dates/times, IDs, etc), auto-generated sensor data, calculated (including a description of the calculation methods and underlying data), etc.
- **Required / Optional**: whether or not the field is required to be populated be the data collector/submitter
- **Collection Period**: whether the variable is expected to be collected indefinitely, or only intended to be collected for a limited period of time
- **Other Fields**: any additional fields that may be useful to describe and manage the dataset

Example Data Dictionary Format

Column	Label	Type	Description	Data Categorization	Valid Values / Examples	Known Issues	Condition / Source	Required / Optional	Collection Period	Other Fields