# California State Water Resources Control Board's Open Data Publishing Guide

## Table of contents

**1 Introduction** **1**

**2 Background** **2**

**3 Governance** **2**

**4 Preparing Data** **3**

4.1 Permissions and Privacy . . . . . . . . . . . . 3

4.2 Dataset Development . . . . . . . . . . . . 4

4.3 Data Formatting . . . . . . . . . . . . 5

4.4 Data Quality . . . . . . . . . . . . 5

**5 Preparing Metadata** **6**

**6 Dataset Publishing Mechanics** **6**

6.1 Dataset Naming Conventions . . . . . . . . . . . . 7

6.2 Publishing Data on the California Open Data Portal . . . . . . . . . . . . 8

> ℹ Note
>
> This is a copy of the original guide that was published on Sharepoint here.

## 1 Introduction

This document provides guidance to staff of the California State Water Resources Control Board (State Water Board) on the process of publishing data in an open data format, with the aim of creating a consistent and repeatable set of guidelines and procedures for staff to follow.

It is a living document that will be updated as new issues are identified or new workflows are created. The guide is being developed by the State Water Board's Office of Information Management and Analysis (OIMA), which is leading the State Water Board's Open Data Initiative, and will serve as a resource to facilitate open data publishing efforts and coordinate open data efforts across the State Water Board's departments and programs.

This publishing guide is not meant to be an exhaustive guide to open data principles and practices, but is instead meant to be a practical and concise guide for State Water Board staff. To find more information about the principles and importance of open data, see guides such as the California Health and Human Services Agency's Open Data Handbook or White House's Project Open Data, among many other open data resources.

## 2 Background

On July 10, 2018, the State Water Board adopted resolution number 2018-0032, titled Adopting Principles of Open Data as a Core Value and Directing Programs and Activities to Implement Strategic Actions to Improve Data Accessibility and Associated Innovation. This resolution commits the State Water Board to following several core principles for open data, including striving to make all critical public data available in machine readable datasets with metadata and data dictionaries. The resolution also directs OIMA to develop a Data Management Strategy that will guide the State Water Board's efforts to implement practices and procedures consistent with open data principles, and this Open Data Publishing Guide is one part of that Data Management Strategy.

Open data principles generally specify that datasets should be public, accessible, described, reusable, complete, timely, and managed post-release. Managing data as an asset in this manner promotes transparency, accountability, and innovation, by making it as easy as possible for anyone to analyze and utilize publicly available government data, and providing derivative value that often cannot be predicted. Ultimately, making public data more accessible and usable helps facilitate the process of turning the data that we already collect into actionable information, and democratizes access to that data.

## 3 Governance

In the State Water Board's efforts to implement open data principles, OIMA will generally act as a Data Coordinator, serving as an intermediary between data providers and any published datasets. While these roles and responsibilities are flexible and will vary depending on the particular program and datasets being published, they will generally be defined as follows:

- **Data Coordinator**: OIMA, who will work collaboratively with the data providers to obtain the required permissions to publish their data in an open data format, ensure that the published datasets are compliant with open data standards and consistent with other State Water Board datasets, and help to publish the datasets in publicly accessible location and refresh them on a regular basis.

- **Data Providers**: Program staff who collect and maintain the source datasets (sometimes referred to as Data Stewards), who will take primary responsibility for approving the datasets to be published and identifying the specific fields to include in those datasets, writing metadata describing the data, responding to and approving any required revisions to the dataset after publication, and answering questions about the data or the program the data derives from.

The governance of data must be addressed at the organizational level to ensure data is secure, non-public data does not get published as open, and, most importantly, that data needed for core business functions and services is made accessible, has known and useful quality characteristics and is sufficiently documented and archived. As part of the 2019 Data Management Strategy effort the Water Boards will develop and adapt governance systems to address these core principles.

# 4 Preparing Data

Prior to publishing datasets in an open data format, an approval process must be completed and several steps must be taken to develop and transform data and metadata to ensure that they are consistent with open data standards for usability and documentation. This section describes the framework for dataset preparation. As noted above, each step in this process will be a collaborative effort between OIMA and the program staff providing the data.

## 4.1 Permissions and Privacy

In general, the State Water Board is committed to proactively releasing publishable state data. The initial step in the publishing process is to verify that the dataset is publishable, meaning that the data are (1) not exempt from disclosure, per the California Public Records Act, (2) are not prohibited from being released by any laws, regulations, policies, rules, rights, court order, or any other restriction, and (3) do not contain sensitive or personally identifiable information. If any unpublishable, sensitive, or personal data is present in the dataset, that data must be removed, masked, or otherwise censored in such a way that it meets standards for publication.

In addition, the data provider must be identified and must approve a plan for publishing the data in an open data format (including approving the data fields to be published, the metadata to be included, the forum for publishing the data, and plans for maintaining and updating the published dataset), even if the dataset is already made publicly available by other means. OIMA will ensure this process is completed and documented by requiring an approval document to be signed by managers of the program providing the data. This is meant to ensure that the data provider is aware of where their data is being published, so that they are prepared to answer questions that may arise about the data or the information the dataset reveals, and so they can help maintain the dataset and address any changes that may occur over time. The dataset approval document is available as an appendix to this handbook.

## 4.2 Dataset Development

To identify and prioritize data for publishing in an open data format, departments should assess their existing datasets for value, quality, completeness, and appropriateness in accordance with the definition of publishable state data. High value data are those that can be used to increase the State entity's accountability and responsiveness, improve public knowledge of the State entity and its operations, further its mission, create economic opportunity, or respond to a need or demand identified after public consultation (for a more detailed discussion of the factors involved in identifying and prioritizing datasets for publishing in an open data framework, see the CHHS Open Data Handbook section on Publication Guidelines).

Once a dataset has been identified for publishing in an open data format and the required permissions have been obtained, the data provider must determine the level of detail to include in the dataset. The aim is to reach an appropriate balance between usability and comprehensibility of the dataset versus completeness of the data provided. Data providers should consider that datasets which contain information that is irrelevant, ambiguous, excessively redundant, or otherwise incomprehensible to the public may become overwhelming to potential users, and may also be unreasonably large to process. On the other hand, data providers should consider that users of any given dataset may come from a variety of fields and specialties, some of who may envision a use for data not anticipated by the data provider, so care should be taken when omitting data of questionable utility as it can be difficult to anticipate all potential uses of a dataset.

There are no strict rules for determining what level of detail is appropriate when selecting the particular data fields to include in a published dataset in an open data format, but a general best practice is to ensure that the metadata associated with each dataset is complete and comprehensive. The data provider should be able to produce comprehensive data dictionaries describing all data fields provided, as well as comprehensive overview documents describing the dataset where appropriate (such as background information about the programs collecting the data, relationships to related datasets from other programs or agencies, data collection methods, known limitations of the data, etc.). The overall aim of the dataset development process is to apply an intentional decision-making process when selecting the specific information to provide in any given dataset, which considers the tradeoffs inherent in those decisions and ensures that datasets are designed to maximize the public's ability to understand and interpret the data, while also facilitating interoperability with other datasets.

## 4.3 Data Formatting

All data published in an open data framework should be presented in a machine-readable format that enables efficient processing with modern software tools and applications. This means that data should be provided in a common, non-hierarchical, plain text file format (such as comma separated values, or CSV). In addition, for any data that is exported from a database system or converted from another file format, checks should be performed to locate and clean any elements within the dataset which could cause formatting problems when converted to a plain text format, such as special characters embedded within the data.

Data providers should also consider the structure of the datasets which they provide. This means that providers should seek to structure their data in a manner that, to the greatest extent possible, facilitates analysis of the data. In general, the data should be provided in such a way that every observation is a row, and every variable is a column – this is sometimes referred to as tall or tidy data. Note that this may contrast with the way that data is presented for publication, and may also not be consistent with the way that data is often stored to maximize efficiency (e.g., principles of database design). However, formatting data in this manner helps to facilitate data analysis by providing data in a consistent format that is easy to work with using a wide range of data analysis tools.

For more information about the principles of data formatting for efficient analysis, see resources such as the California Open Data Portal guide to data preparation or other resources listed below.

## 4.4 Data Quality

In general, data quality assurance should be a part of the data collection and management efforts that occur prior to publishing datasets in an open data format. Therefore, extensive data cleaning should not generally be a significant part of the open data publishing process. However, to the extent possible, the data provider should work with OIMA to develop a plan to perform basic checks for usability before publishing the data, such as:

- checking to make sure that special characters that could cause formatting issues are removed;
- checking to make sure that all records within a field are of a consistent type (for example, if a field should be numeric, ensuring that all records can be treated as a number);
- checking for and addressing obvious outliers; and
- verifying the reasonableness of any geospatial data (for example, making sure that any points with associated location data are within an expected area when plotted on a map).

Despite the data provider's best efforts, publishing data in an open data format may sometimes expose data quality issues and create a need to address data quality concerns that are raised as a result. Therefore, it is also important to be aware of and/or plan for potential needs to:

- revise the published dataset (e.g., a dataset published on an open data portal);
- apply the revisions to the source dataset (e.g., an internal database used to store the data); and
- notify users of the published dataset of significant changes to previously published data

## 5 Preparing Metadata

Metadata is information that describes a data resource. It helps users understand the contents of a data resource and how it can be appropriately interpreted and applied, and it also provides the information needed to organize data repositories so that users can find the specific data resource they want.

There are a variety of different types of information contained in metadata. Structural metadata describes the format of the file containing the data and its location (e.g., file type, URL, etc.), and descriptive metadata to tell users what information the file contains (e.g., title, description, publisher). The more information you provide about your data resources, the easier it will be for users to find. For publishing to the California Open Data Portal (data.ca.gov), the user is prompted to enter all of the required information when creating a new dataset or data resource (more guidance can be found here).

Metadata also includes a data dictionary, which tells the user exactly what is in the data resource and how it can be used. It lists all the variables, both by the name they go by in the resource as well as a plain English name, the format (e.g., text, numeric, date) for each, and a description of what information the user can find in that variable. A Data Dictionary helps the user decide if the data resource will provide the data that is needed to answer their question, and also if it is in a form they can use. Guidance for development of a data dictionary can be found in the Recommendations and Guidelines for Data Dictionary Development document.

## 6 Dataset Publishing Mechanics

Once a dataset and associated metadata consistent with the open data principles described above have been prepared, they should be published on a platform that makes them accessible to the general public. Accessibility includes a number of considerations, including how easy it is to discover and locate the dataset (especially for those who do not already know that the dataset exists), and what options are available to connect to the dataset (e.g., APIs, direct downloads, or other web-services). The state of California maintains an open data platform at data.ca.gov, which the State Water Board currently plans to use as its primary platform for hosting its open data; however, there are many other platforms available to publish data in a manner consistent with open data principles. The following considerations should be taken into account when selecting an open data platform and publishing open data:

- A persistent link to the dataset should be maintained whenever a dataset is published.
- Consistent nomenclature for titles of datasets and data resources should be used; guidelines are available in the section below.
- Significant changes to the structure of the dataset (e.g., changing which fields are included in the dataset) should be avoided to the extent possible.
- Corrections and revisions to the data should be anticipated, and a plan for how to deal with these issues should be developed prior to publishing (e.g., the authority and/or responsibility of the data provider and the data coordinator to make changes to the published dataset in response to any identified data errors should be clearly defined).

Another important aspect of the data publishing process is the development and implementation of a plan to ensure that any published dataset is updated on a consistent basis to include new data and to reflect any revisions to previously published data. The appropriate frequency of updates will vary for each dataset, and will depend on the frequency of data collection as well as the degree to which the data is needed to make time-sensitive decisions. In general the following factors should be taken into consideration when developing a plan to maintain the dataset:

- The process of updating datasets should be automated to the greatest extent possible. This can help to ensure that datasets are updated consistently, and can help to avoid errors. In addition, a staff member should be assigned to periodically check the dataset to ensure the updates are working as expected. OIMA can lead the technical aspects of this process, but will need to coordinate with the data provider to determine the proper frequency of updates, and to perform periodic checks of the dataset to ensure that updates are occurring as expected.
- If automation isn't possible, OIMA and the data provider should coordinate to develop a schedule and plan to regularly update the dataset manually.

## 6.1 Dataset Naming Conventions

The title of datasets uploaded to the open data portal should follow the naming convention described below. This naming convention will assist data users in identifying the datasets of interest to them.

In general (see exception discussed below), the title should consist of two parts separated by a hyphen. The first part of the title should describe the Water Board business area associated with the dataset. The second part of the title is a brief description of the data contained in the dataset. An example title following this convention for a dataset involving drinking water system service area boundaries is:

- Drinking Water – Water System Service Area Boundaries

To facilitate consistency in the first part of the title, the following list of Water Board business areas should be used whenever possible:

| Business Area Name | Description |
|---|---|
| Drinking Water | Data involving the protection of drinking water. Typically, data related to the regulation of drinking water systems. |
| Financial Assistance | Data related to financial assistance programs for infrastructure improvement and environmental protection projects. Typically, data involved in the administration of loans and grants. |
| Water Quality | Data involving the protection of the environment, public health, and all beneficial uses. Typically, data related to the regulation of discharges of waste to the environment. |
| Water Rights | Data related to ensuring proper water resource allocation and efficient use. Typically, data involved in the regulation of water rights. |

The second part of the title is a brief description of the dataset. Use simple, clear terms that are readily understood by individuals outside of the Water Boards and avoid the use of acronyms.

The exception to the two-part title convention involves datasets related to the Water Quality business area. For these datasets, an additional descriptor for the type of waste discharge associated with the dataset is used between the business area name and dataset description. The suggested list of descriptors is groundwater, storm water, and surface water. An example title for a dataset involving enforcement action information related to surface water discharges is:

- Water Quality – Surface Water – Active and Historical Enforcement Action Information

## 6.2 Publishing Data on the California Open Data Portal

The California Open Data Portal at data.ca.gov is currently the primary platform that OIMA is using to publish State Water Board data in an open data format. The California Open Data Portal is managed by the California Government Operations Agency, who maintain detailed instructions for publishing data on the portal. OIMA can lead or assist data providers in this task, provided that the dataset and metadata have been developed consistent with the guidelines described above.