Appendix A. – Open Data Publishing Process

Background and purpose

Open Data is an important part of any organization, especially in the public sector. It's essential to have a robust process for vetting datasets for publishing as it improves the quality of data. High quality and well documented open data are a critical component of data governance, as well as making sure data is accurate, secure, and properly used. This document should be used as guidance for how to publish Caltrans data as open data.

For a workflow diagram of the entire publishing process, see *Figure 11 - Open Data Publishing Workflow* on page 17 of this document.

1. Complete the Data Classification Checklist

The first step in the publishing process is to classify your data as **corporate data**. Corporate data is data that falls into one or more of the following categories, or a combination thereof:

A. **Master Data** identifies and describes the people, places and things that are fundamental to agency activities and appear across multiple agency information systems, like Project Short Names.
B. **Reference Data** is permissible values for other data fields, like County codes and names.
C. **Control Agency Required Data** is data that is required by law or directly impacts agency funding or resource allocation decisions like SB1 reports or Federal Highway Performance Monitoring System data.
D. **Shared Data** is used across multiple internal business units or is valuable to either Caltrans' partners or the public. An example of shared data is the Caltrans Fact Book.
E. **Purchased/Licensed Data** are produced as a result of state/federal grants, contracts, or cooperative agreements.

We must also identify any **sensitive, confidential, personally identifiable information** that needs to be removed from the dataset, as sharing this data is prohibited. Additionally, if there is a potential risk to Caltrans, its employees, or the public if the data was shared then those risks need to be documented and it should be protected accordingly.

The **Data Classification Checklist** has been developed as an online form to help classify Caltrans data and identify any data that should be exempt from publishing. To complete this form click here.[1]

[1] PowerApps Data Classification Checklist Survey

2. CTDATA Review & the Data Classification and Security Committee

CTDATA staff will review the results of the data classification checklist. If there is a need for additional questions to be asked, or concerns about sharing the data, they may refer you to the **Data Classification and Security Committee** (DCSC) for additional review. The DCSC will then schedule a meeting to review your dataset and determine if the data:

A. Can be published as Open Data.
B. Needs some elements redacted prior to publishing as Open Data.
C. Should be marked as Restricted or Non-Public.

Once the DCSC has made their determination regarding the data, the results shall be recorded in Caltrans' Corporate Data Catalog. If the data can be published as Open Data, then it is time to complete the required data documentation.

You may request a review by the DCSC at the time of submitting your data classification checklist.

3. Choose a License

All data that is classified as "public (open data)" shall have a license applied to it. The default license for Caltrans Open Data shall be Creative Commons Attribution 4.0 International.[2]

Visit creativecommons.org[3] to learn more about Creative Commons licensing.

Other licenses may be used if approved by Caltrans Legal and the Data Classification and Security Committee.

[2] https://creativecommons.org/licenses/by/4.0/?ref=chooser-v1
[3] https://creativecommons.org/

4. Complete Data Documentation

CTDATA has developed standard data documentation that must accompany data published internally and externally, including Open Data. These templates promote consistency in how we document, provide context to, and share information about Caltrans data.

You can read the Data Documentation Guide and download the latest required data documentation templates on Onramp.[4] Once the data documentation is completed, submit them for review by CTDATA staff by emailing CTDATA@dot.ca.gov

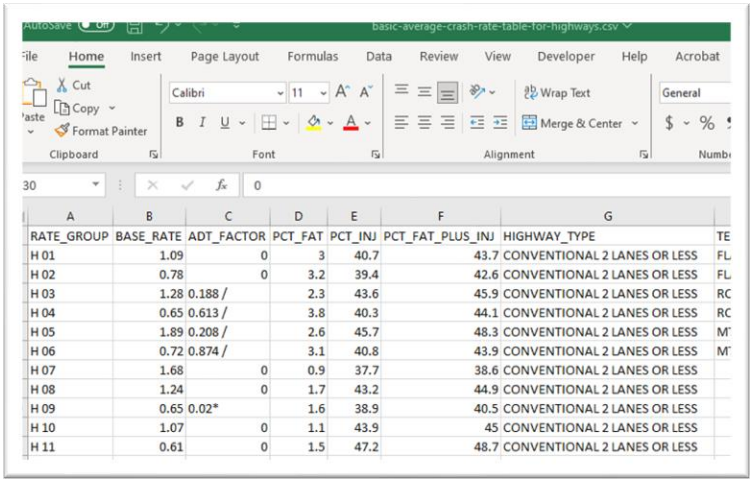
[4] https://datagovernance.onramp.dot.ca.gov/data-documentation-package-and-templates

5. Publishing Open Data

Once you have completed the previous steps, you are now ready to publish your data. There are two processes for publishing data as Open Data. To decide which process is correct, you will need to determine if your data is Tabular or Geospatial.

1. Tabular data

Tabular data is observations and attributes in the form of a table organized with rows and columns. This is often saved as a CSV or Excel. This kind of data can be shared publicly on the California Open Data Portal[5]. An example is the 2021 Crash Data on State Highway System[6] which has several tabular datasets within one page. The example below is one of those datasets, Basic Average Crash Rate Table for Highways[7] and downloads as a CSV.



RATE_GROUP	BASE_RATE	ADT_FACTOR	PCT_FAT	PCT_INJ	PCT_FAT_PLUS_INJ	HIGHWAY_TYPE	TE
H 01	1.09	0	3	40.7	43.7	CONVENTIONAL 2 LANES OR LESS	FL
H 02	0.78	0	3.2	39.4	42.6	CONVENTIONAL 2 LANES OR LESS	FL
H 03	1.28	0.188 /	2.3	43.6	45.9	CONVENTIONAL 2 LANES OR LESS	RC
H 04	0.65	0.613 /	3.8	40.3	44.1	CONVENTIONAL 2 LANES OR LESS	RC
H 05	1.89	0.208 /	2.6	45.7	48.3	CONVENTIONAL 2 LANES OR LESS	M
H 06	0.72	0.874 /	3.1	40.8	43.9	CONVENTIONAL 2 LANES OR LESS	M
H 07	1.68	0	0.9	37.7	38.6	CONVENTIONAL 2 LANES OR LESS	
H 08	1.24	0	1.7	43.2	44.9	CONVENTIONAL 2 LANES OR LESS	
H 09	0.65	0.02*	1.6	38.9	40.5	CONVENTIONAL 2 LANES OR LESS	
H 10	1.07	0	1.1	43.9	45	CONVENTIONAL 2 LANES OR LESS	
H 11	0.61	0	1.5	47.2	48.7	CONVENTIONAL 2 LANES OR LESS	

Figure 1 - Example of Tabular Data

Go to the Tabular Publishing section below for instructions on publishing this kind of data.

2. Geospatial data

Geospatial data is information that has a location component that is mappable on the surface of the planet. This location can be displayed as a point, line, or polygon on a map, and is typically in the form of a feature service, file geodatabase, and/or shapefile, but other tabular file types can contain geospatial data as well. This information is published publicly on the

[5] https://data.ca.gov/
[6] https://data.ca.gov/dataset/2021-crash-data-on-state-highway-system
[7] https://data.ca.gov/dataset/2021-crash-data-on-state-highway-system/resource/57411d59-fc84-4da2-87e5-68a5e7db7d9c

California State Geoportal,[8] like the Caltrans Districts[9] dataset example below. Datasets posted to the Geoportal are automatically replicated and posted to the California Open Data Portal for indexing.

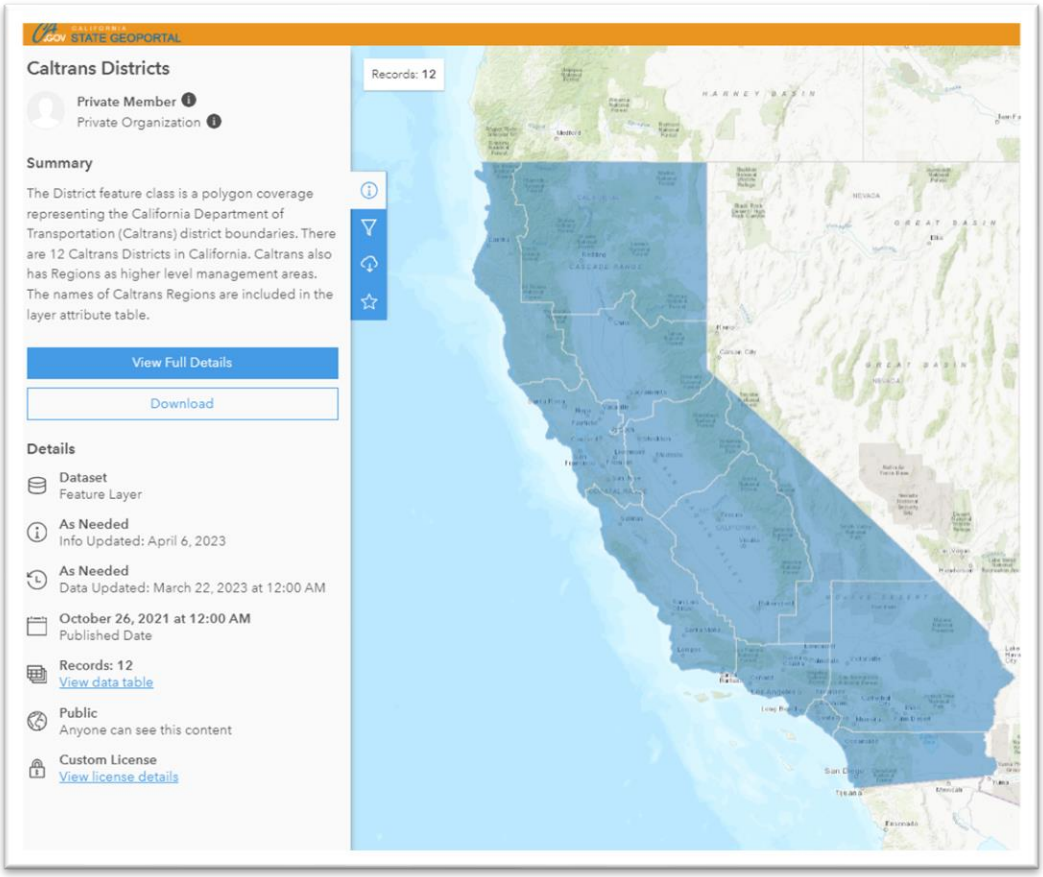


Figure 2 - Example of Geospatial Data

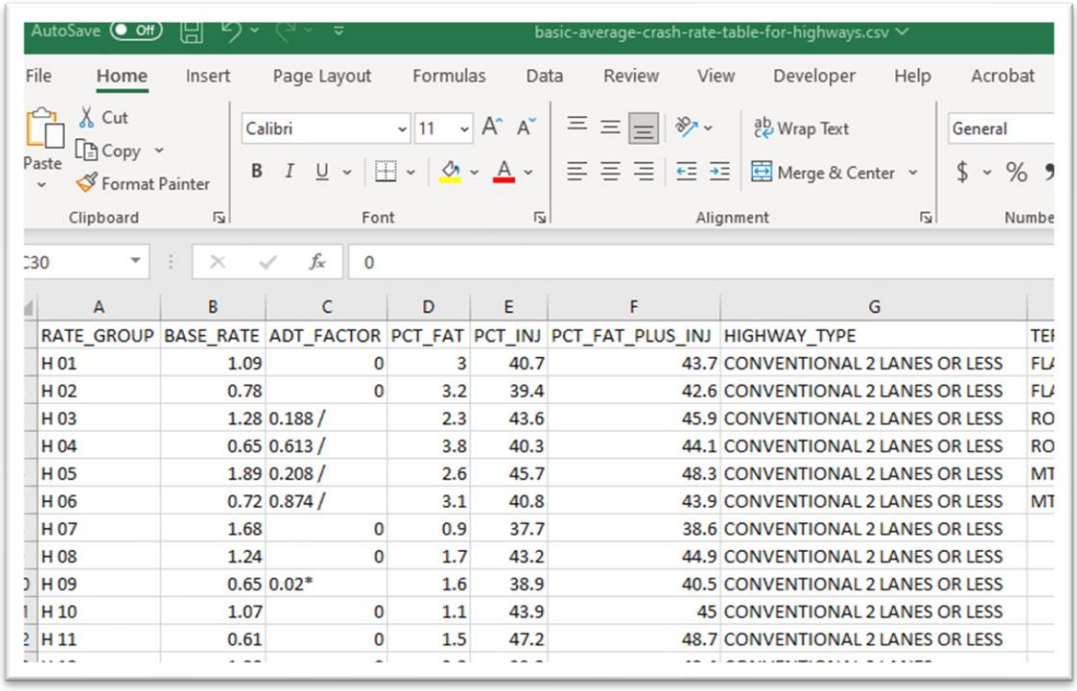
Go to the Geospatial Publishing section below for instructions on publishing this kind of data.

[8] https://gis.data.ca.gov/
[9] https://gis.data.ca.gov/maps/0144574f750f4ccc88749004aca6eb0c

6. Tabular Data Publishing

A. Create or Update the Dataset

The data custodian or data steward shall create or update the data, e.g., spreadsheet, table, csv, etc. The screenshot below is from the Basic Average Crash Rate Table for Highways[3] which is downloaded as a CSV.



	A	B	C	D	E	F	G	
	RATE_GROUP	BASE_RATE	ADT_FACTOR	PCT_FAT	PCT_INJ	PCT_FAT_PLUS_INJ	HIGHWAY_TYPE	TE
	H 01	1.09	0	3	40.7	43.7	CONVENTIONAL 2 LANES OR LESS	FL
	H 02	0.78	0	3.2	39.4	42.6	CONVENTIONAL 2 LANES OR LESS	FL
	H 03	1.28	0.188 /	2.3	43.6	45.9	CONVENTIONAL 2 LANES OR LESS	RO
	H 04	0.65	0.613 /	3.8	40.3	44.1	CONVENTIONAL 2 LANES OR LESS	RO
	H 05	1.89	0.208 /	2.6	45.7	48.3	CONVENTIONAL 2 LANES OR LESS	MT
	H 06	0.72	0.874 /	3.1	40.8	43.9	CONVENTIONAL 2 LANES OR LESS	MT
	H 07	1.68	0	0.9	37.7	38.6	CONVENTIONAL 2 LANES OR LESS	
	H 08	1.24	0	1.7	43.2	44.9	CONVENTIONAL 2 LANES OR LESS	
)	H 09	0.65	0.02*	1.6	38.9	40.5	CONVENTIONAL 2 LANES OR LESS	
1	H 10	1.07	0	1.1	43.9	45	CONVENTIONAL 2 LANES OR LESS	
2	H 11	0.61	0	1.5	47.2	48.7	CONVENTIONAL 2 LANES OR LESS	

Figure 3 - Example of Tabular Data

B. Check Data Quality

Ensure:

1. The data is current and includes any updates.
2. The data is accurate, understandable, and concise.
3. There is a Data Quality Management Plan (DQMP, aka, DDP-10), which is on the Data Documentation Package and Templates[10] site (more info below).
4. Per the metadata, provide the data updates on a set schedule (e.g., annually, monthly).

1. Complete the Data Documentation Templates

The Data Documentation Package and Templates[6] site has all required templates that meet open data requirements. DDP-6 (Dataset Metadata),

[10] https://datagovernance.onramp.dot.ca.gov/data-documentation-package-and-templates

DDP-8 (Data Dictionary), and DDP-10 (Data Quality Management Plan) are required for these datasets.

2. Submit your dataset and data documentation for review

This is submitted to the Caltrans Open Data Coordinator using the email ctdata@dot.ca.gov. This is monitored by the CTDATA team, and they will review it.

3. Publication

Once review is complete, and any corrections are made, the CTDATA team will publish it to the California Open Data Portal[1].

7. Geospatial Data Publishing

A. Create or update GIS dataset

The data custodian and/or business data steward typically does this using GIS software that makes it possible to create, analyze, manage, and share geospatial data and maps. The screenshot below is from an ArcGIS Pro Project used for the Caltrans Districts and displays the geospatial data on the map as well as the associated attribute table with information about what is on the map.

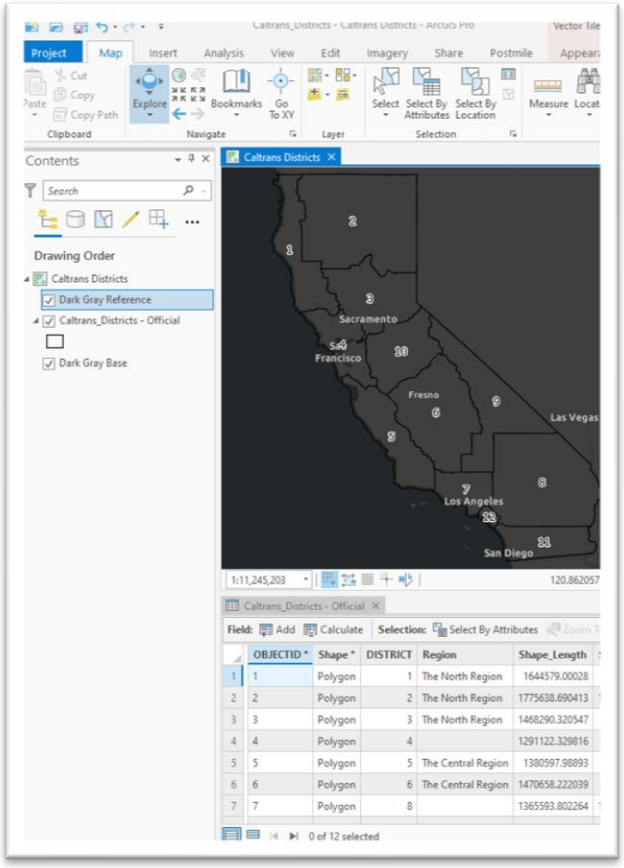


Figure 4 - Example of Geospatial Data in ArcGIS

B. Check Data Quality

Ensure:

1. The data is current and includes any updates.
2. The data is accurate, understandable, and concise.
3. There is a Data Quality Management Plan (DQMP, aka, DDP-10) implemented, which is on the Data Documentation Package and Templates[6] site (more info below).
5. Per the metadata, provide data updates on a set schedule (e.g., annually, monthly).

Caltrans_Districts

Field: Add Calculate | Selection: Select By Attributes Zoom To Switch

	OBJECTID *	Shape *	DISTRICT	Region	Shape_Length	Shape_Area
1	1	Polygon	1	North	1644579.00028	46599103942.556686
2	2	Polygon	2	North	1775638.690413	126694960969.822067
3	3	Polygon	3	North	1468290.320547	54587476873.405869
4	4	Polygon	4		1291122.329816	33833865147.604084
5	5	Polygon	5	Central	1380597.98893	48530897490.658821
6	6	Polygon	6	Central	1470658.222039	77711292318.891174
7	7	Polygon	8		1365593.802264	104917451858.524307
8	8	Polygon	10	Central	1205887.22955	45913241767.334824
9	9	Polygon	11		926255.51473	33286886469.761681
10	10	Polygon	7		714383.949603	23587047310.747429
11	11	Polygon	9		1972516.507051	66235517300.760201
12	13	Polygon	12		260845.307991	3555329112.183568

C. Create or update metadata in ArcGIS Pro

1. Ensure the correct Metadata Style is selected, "ISO 19139 Metadata Implementation Specification GML3.2." See screenshot below from the selection in ArcGIS Pro.

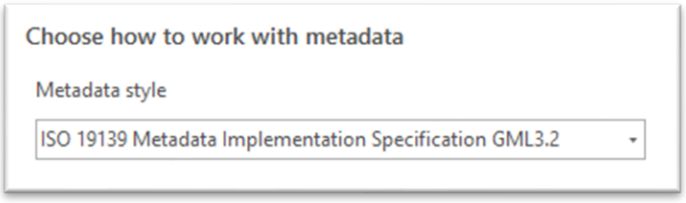


Figure 5 - Metadata Specification in ArcGIS Pro

   More information on how to create this in ArcGIS Pro can be found in the Metadata slides on the Trainings Resources[11] page.
2. Use the Metadata Editor in the Catalog Pane. One way to get there is to right click the feature class from within the file geodatabase.

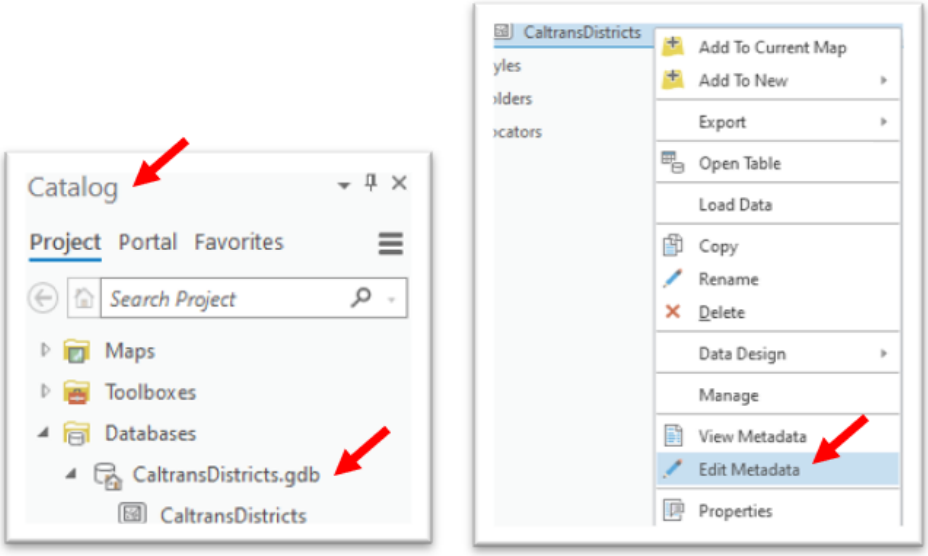


Figure 6 - Accessing the Metadata Editor in ArcGIS

3. Fill out the metadata following Caltrans standards on the Data Documentation Package and Templates page.
4. If you are updating the metadata, make sure to remove old metadata from any part of the editor that might hold a contact. These are:
   a. Overview > Citation Contacts
   b. Metadata > Contacts
   c. Resource > Point of Contact

[11] https://datagovernance.onramp.dot.ca.gov/training-resources

d. Resource > Maintenance > Contact
e. Resource > Lineage > Process Step > Processor Contact
f. Resource > Distribution > Distributor Contact

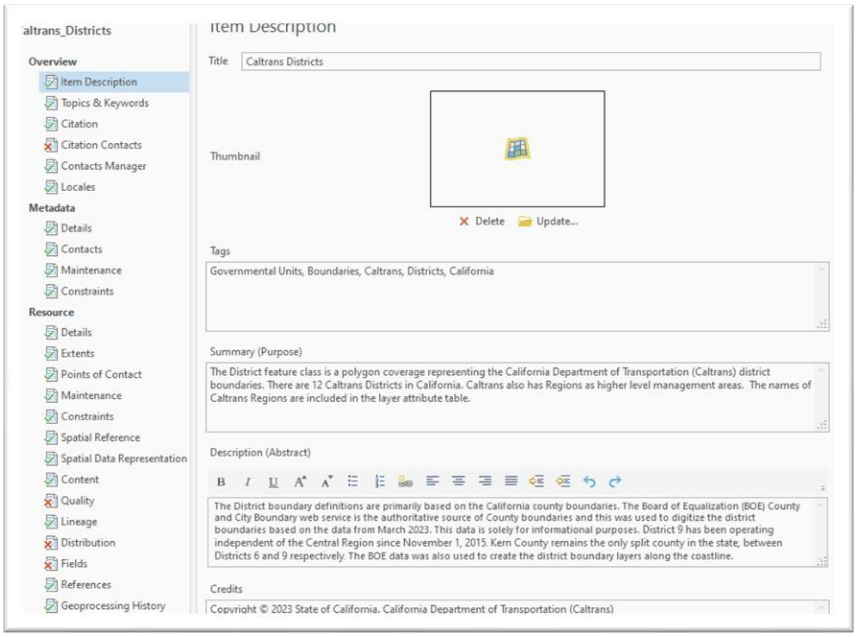


Figure 7 - Example of "Overview > Item Description" in the metadata editor of ArcGIS Pro

D. Use the Data Documentation Package and Templates

The Data Documentation Package and Templates[6] page provides information that ensures all open data requirements are completed. DDP-6 (Dataset Metadata) and DDP-8 (Data Dictionary) are especially important for these datasets. ESRI has differing standards so you MUST follow these Caltrans standards.

E. Create a Thumbnail

Follow the Thumbnail Guidelines found on the Resources for GIS Projects page under Thumbnail Guidance. The thumbnail image shall be sent along with your data. An example thumbnail image is below.

Figure 8 - Example Thumbnail

F. Submit an Open Data Request

The GIS Open Data Request[12] form can be found on the GEEP site and is reviewed by the Enterprise GIS Program team. There are several pages starting with the screenshot below. Fill out the form completely to submit your request.

Figure 9 - GIS Open Data Request Form

[12] https://sv03tmcpo.ct.dot.ca.gov/portal/apps/sites/#/geep/pages/open-data-request

G. Review dataset in the CA State Geoportal

Enterprise GIS staff notifies CTDATA to add the dataset to the CA State Geoportal[4]. CTDATA will reach out to you after this final step is completed and you will need to review what is posted. Make sure to click on the View Full Details button (see Caltrans Districts dataset example from the CA State Geoportal below) to see more information about the dataset including the tabular data and metadata.

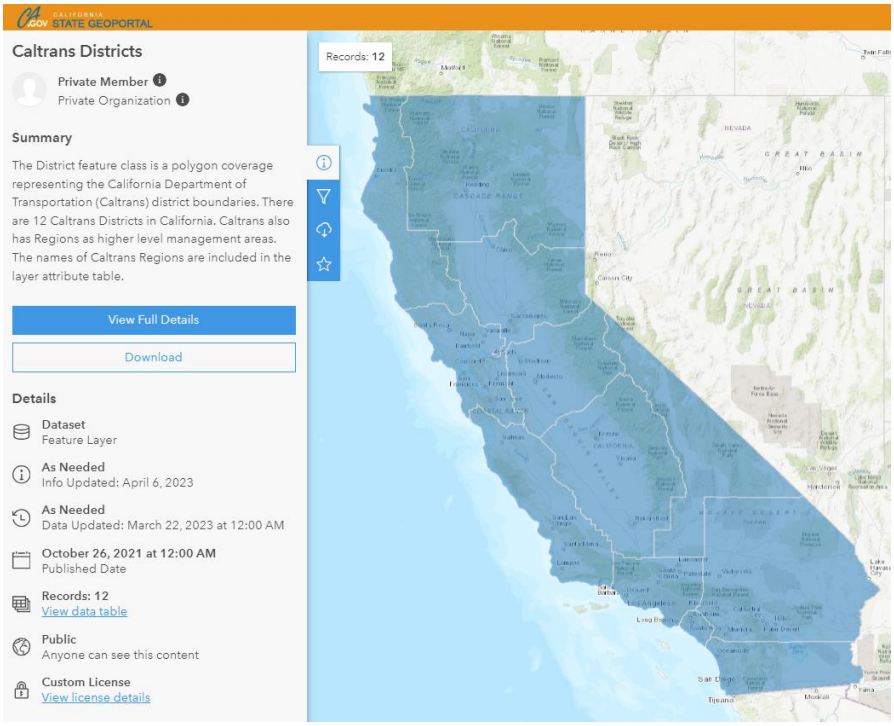


Figure 10 - Example Dataset; Caltrans Districts

8. Schedule Data Refresh and Data Documentation Review

Once you have published your data it is vital that you create a plan to refresh the dataset and go through the steps outlined in this document again. The recommendation is that datasets be refreshed based on their value to the public if no legal requirements mandate a refresh rate.

If the data elements in your dataset have not changed, then repeating the process outlined in this document should be very quick. If you have added fields or changed data element attributes, then corresponding changes to your data documentation are required to publish.

-END PROCESS-

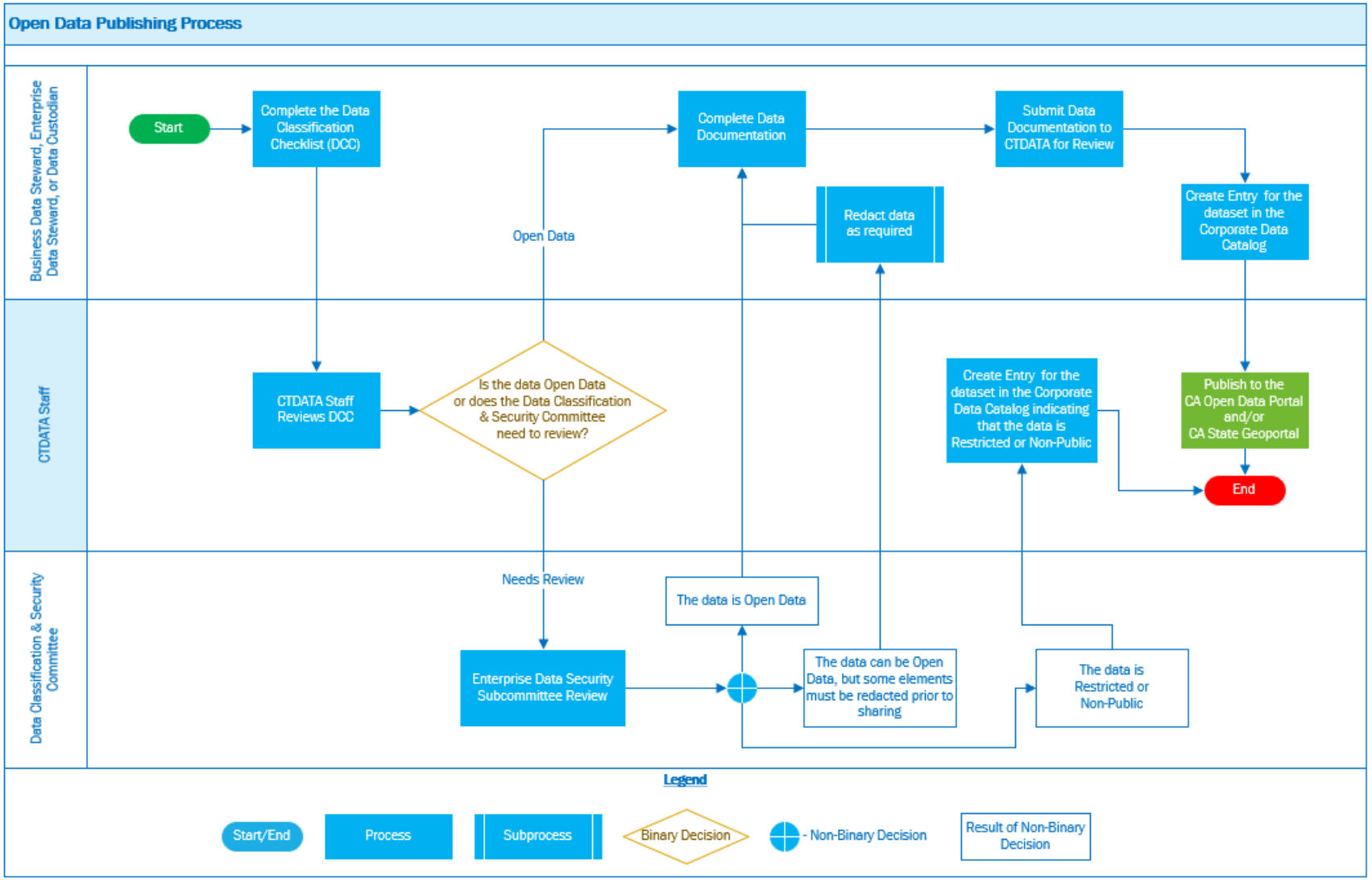


Figure 11 - Open Data Publishing Workflow

-End of Document-