Open Data Publication Guidelines

This document provides guidance on the publication of data maintained by Connecticut state agencies as open data. Open data is data that can be freely used, reused, and redistributed without legal or financial restrictions.[1] These guidelines focus on the publication of data on the Connecticut Open Data Portal (data.ct.gov). This guidance is intended for Agency Data Officers and other publishers of open data within Connecticut state government.

Overview

The purpose of the Open Data Portal, established under C.G.S. Sec. 4-67p, is to publish timely, well-documented, and easily accessible data from government agencies. The Open Data Portal is maintained by the Data and Policy Analytics unit in the Office of Policy and Management, in collaboration with data publishers in other Connecticut state agencies, to:

- Increase agency accountability, responsiveness, and (inter)agency efficiency;
- Improve public knowledge of the government and its operations, and encourage public participation with government agencies, policies and issues; and
- Empower citizens or third parties to create social, political, economic, or other value from open data.

This document covers the steps for publishing data on the Open Data Portal, including:

1. Identifying eligible datasets,
2. Evaluating datasets,
3. Preparing data for publication, and
4. Publishing open data.

Step 1. Find Eligible Datasets

All high-value data collected or possessed by a government agency can be considered for publication as open data.[2] Agencies must adopt a presumption in favor of openness to the extent permitted by law and subject to privacy, confidentiality, security, or other valid restrictions.

The data sources in your agency's high-value data inventory (the CT Data Catalog) should be the sources for your open data.

---

[1] Section 4-67o of the Connecticut General Statutes defines open data as: "Any data that (A) is freely available in convenient and modifiable format and can be retrieved, downloaded, indexed and searched; (B) is formatted in a manner that allows for automated machine processing; (C) does not have restrictions governing use; (D) is published with the finest possible level of detail that is practicable and permitted by law; and (E) is described in enough detail so users of the data have sufficient information to understand (i) the strengths, weaknesses, analytical limitations and security requirements of the data, and (ii) how to process such data."

[2] Section 4-67o of the Connecticut General Statutes defines high-value data as: "Any data that the department head determines (A) is critical to the operation of an executive branch agency; (B) can increase executive branch agency accountability and responsiveness; (C) can improve public knowledge of the executive branch agency and its operations; (D) can further the core mission of the executive branch agency; (E) can create economic opportunity; (F) is frequently requested by the public; (G) responds to a need and demand as identified by the agency through public consultation; or (H) is used to satisfy any legislative or other reporting requirements."

The high-value data inventory can inform which datasets to consider for publication, but it is not enough on its own. The inventory just catalogs the data you have, it does not help prioritize or evaluate risks from publication. While there are many benefits to publishing open data, it is important to weigh those advantages against privacy risks. If the data contains information about residents, agencies must take appropriate action to protect individual privacy before releasing the data as open data. For more information, see the "Privacy Considerations" section.

Remember that the owner agency retains ownership and record-keeping responsibility for its data regardless of whether it is published on the Open Data Portal.

Step 2. Evaluate Datasets

After your agency has identified data that may be published as open data, evaluate the data further to make sure they should be published. Make sure that the data:

- Belong to your agency[3];
- Have no fee for access, use, adaptation, or use of the data;
- Are free from legal, contract, or policy restrictions;
- Have a plan for predictable or regular updates; and
- Do not violate the privacy of individuals represented in the data. (More guidance on privacy considerations can be found in the next section of this document.)

Datasets should be considered for publication as open data if they are:

- Subject to reactive disclosure through right-to-know, freedom of information (state or federal) and/or public records laws;
- Shared with other agencies for operational purposes;
- Used for reports on federal, state, or nonprofit grants;
- Used by your agency for trend, statistical, or performance analysis;
- Frequently requested by the public or other government agencies; and/or
- Considered high impact and high value by your agency, partner agencies, the public, and other stakeholders, especially if that data is not already publicly available in a machine-readable format.

Also consider whether this should be published as a new dataset or consolidated with an existing dataset. If your agency has published the same (or similar) data in previous years, try to consolidate this new data with existing data rather than creating separate entries every year. This approach minimizes clutter on the portal and allows users to more easily track changes in the data over time. If you would like the more recent data to be more prominently featured, you can create a "filtered view" based on the larger dataset. For instructions on how to create a filtered view, refer to this article from Socrata or contact the Open Data Portal Administrators at OPM via dapa@ct.gov.

[3] Cases where multiple departments contribute to a single dataset are handled on a case-by-case basis.

Privacy Considerations

If the dataset has one or more of the following identifiers, your agency should consider modifying it to ensure that the privacy of the individuals represented in the data is protected:

- Unique identifiers (e.g., name, SSN) which can identify individuals with relative ease;
- Quasi-identifiers (e.g., birth date, ZIP code, gender/sex, race, ethnicity, age) which can identify people when taken in combination with other available data, or when the sample size is small enough; and
- Sensitive attributes (e.g., protected health or financial information) which could prove harmful and/or stigmatizing if the individuals they refer to were re-identified.

Possible forms of modification include aggregating the data or, if the identifiers add little to no value to the dataset, removing the identifiers entirely before publication.

Most of the time, datasets that contain unique identifiers cannot be made public in their raw form because of privacy laws and protections. There are some exceptions to this rule, including the State Licenses and Credentials dataset provided by the Department of Consumer Protection, which includes individuals with state-issued licenses or credentials. This dataset would not be useful in a modified form, and it serves a clear public good, so it may be published on the portal. If you think your dataset qualifies for such an exception contact the Open Data Portal Administrators at OPM via dapa@ct.gov.

If you have any questions or concerns about a dataset's privacy risks or need guidance about how to best modify the data, contact the Open Data Portal Administrators at OPM via dapa@ct.gov.

Step 3. Prepare Data for Publication

Once your agency has identified a dataset and evaluated it to be appropriate to publish as open data, the next step is to prepare the data for publication. Before you publish the data, data publishers should determine the following:

- **How should the data be formatted?** It is important to determine the structure of the dataset before publishing. The data format should be easy for users to interpret and should be set up in a way that facilitates appending future data if appropriate. Often data that is set up in a long format (rather than wide), sometimes referred to "tidy" data, is the best choice. (This format also facilitates disaggregation and filtering, as described in Step 2.)
- **What should the asset be titled?** Make sure to choose a name that is clear and easily understood by somebody who is not familiar with the data. Avoid acronyms that are not widely understood.
- **Should a date field be included in the dataset?** It is often helpful to include a column for the date that the data was last updated, especially if the data will be updated with any frequency. It is generally recommended to include a date field in your data.
- **What other metadata can be included?** You should complete as much metadata as possible. Not only does this help users understand your dataset, but it helps the portal stay organized and improves search functionality, making it more likely that your dataset is seen. For a description of every metadata field and its meaning, view the CT Open

Data Metadata Guidelines here. Use plain language--avoid confusing acronyms, titles, or other language, and provide context in the description areas. Use this checklist for plain language and this thesaurus for plain language. Also use person-centered language (e.g., "person who is incarcerated" rather than "inmate").

- **How often will the data be updated and what is the process for updating?** It is crucial that agencies update their data on a regular basis and at a predictable frequency. Remember that you should always do your best to consolidate new data with existing data, rather than creating new entries for more recent data. This approach lessens clutter on the portal and makes it much easier to track changes in the data over time. When possible, data updates should be automated. More information about automation options is available here, and the Open Data Administrators at OPM are also available to provide automation support. If a dataset is updated infrequently (for instance, annually), it may make more sense to update it manually.

Step 4. Publish Open Data and Update Data Access Plan

Once you have answered the questions above, you are ready to publish your data as an asset on the Open Data Portal. The person who will be responsible for maintaining the data on the Portal will need access to an account with publishing rights, and the ability to support regular or automated updates. Account creation should be coordinated with an Open Data Portal Administrator at OPM, who can be contacted via email at dapa@ct.gov.

Using the Open Data Access Plan template or your agency's existing Data Access Plan, update the plan for publication of open data for your agency. The plan should include datasets that you plan on publishing to the CT Data Portal, as well as datasets that are already published as open data.

Detailed instructions about how to publish an asset on the Open Data Portal can be found in the appendix at the end of this document. Additional support resources are available from Socrata here.

Acknowledgements

These guidelines were developed based on similar documents from other open data programs, including the examples listed below.

1. NYC OpenData, "New York City Open Data Playbook", https://opendata.cityofnewyork.us/wp-content/uploads/2020/05/2020-ODC-New-York-City-Open-Data-Compliance-Playbook-Final.pdf
2. Sunlight Foundation, "Open Data Policy Guidelines, https://sunlightfoundation.com/opendataguidelines/
3. DataSF, "Publishing Guidelines", https://datasf.org/publishing/guidelines/
4. DataSF, "Open Data Release Toolkit: Privacy Edition", https://datasf.org/resources/open-data-release-toolkit/
5. DataSF, "Open Data Release Form: Privacy Edition", https://www.plainlanguage.gov/resources/checklists/checklist/
6. The World Bank, "Benefits of Open Data", http://opendatatoolkit.worldbank.org/en/starting.html

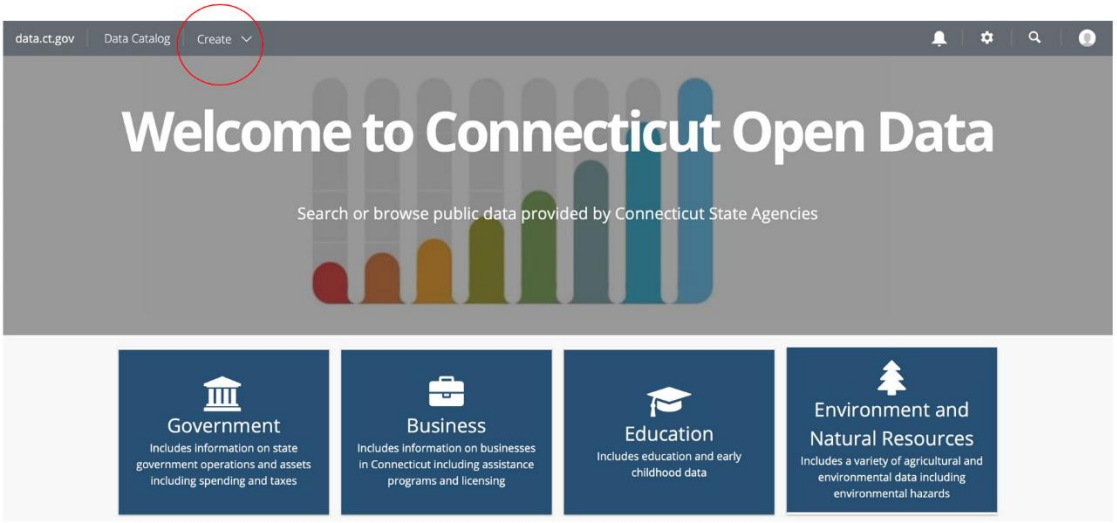
7. Centers for Disease Control and Prevention, “Plain language thesaurus for health communicators,” https://stacks.cdc.gov/view/cdc/11500/
8. Actionable Intelligence for Social Policy, “Centering Racial Equity Throughout Data Integration”, https://www.aisp.upenn.edu/wp-content/uploads/2020/08/AISP-Toolkit_5.27.20.pdf

Appendix

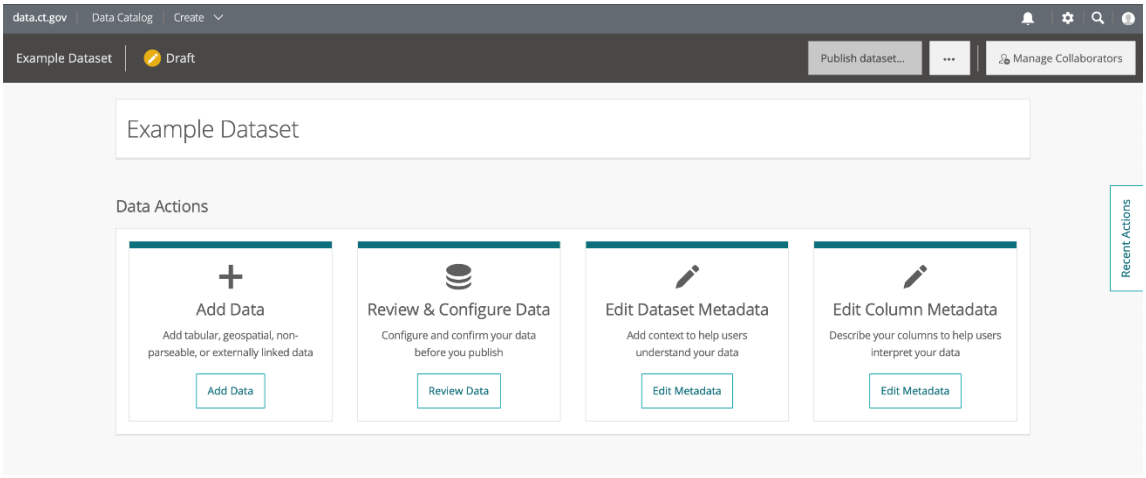
The steps below walk through the steps of creating a new asset on the Open Data Portal.

Create a New Asset

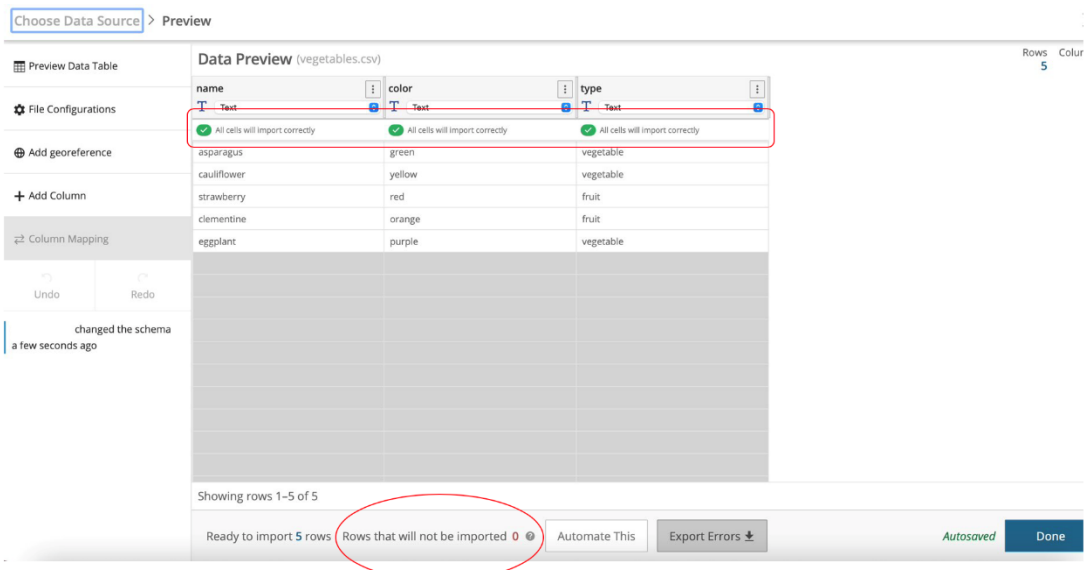
1. Sign into the Open Data Portal. If you don’t have an account on the portal, contact the Open Data Portal Administrators at OPM via dapa@ct.gov to get an account with publishing permissions.
2. At the top left of the screen, click “Create.”



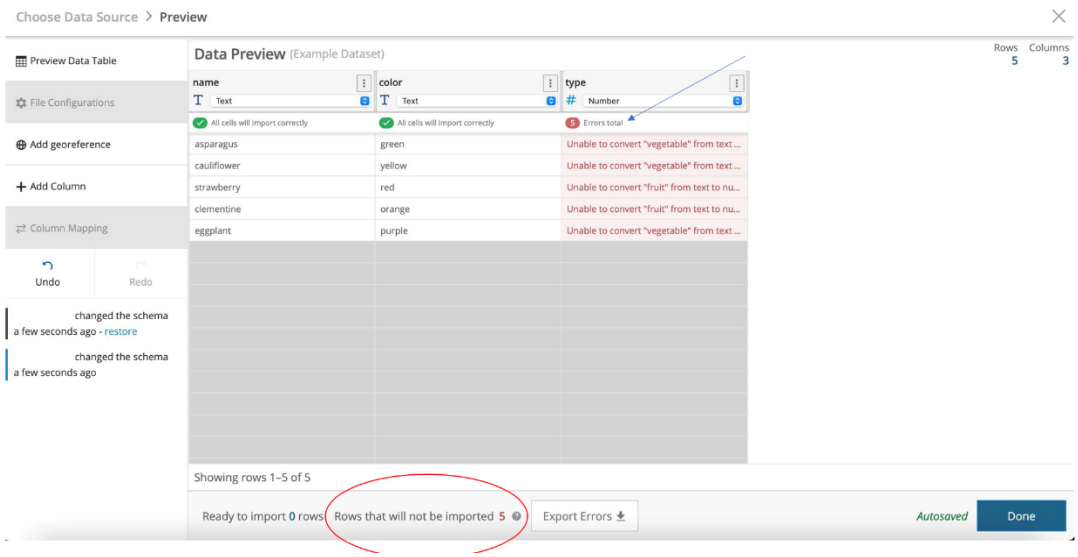
3. Select the type of asset (e.g. dataset, story, ESRI map layer) and name the asset. You’ll then be taken to the screen pictured below, referred to here as the landing page:



4. Click on "Add Data" to upload the data. Before importing the file, consider the following:
    - **Publish data in the rawest form that is appropriate**. Keep in mind that data may be useful to a variety of users and may have applications beyond what the originating agency imagines. Remember that raw data should not be published if it poses a risk of harm, as discussed in the section above.
    - Remove metadata, footnotes, disclaimers, or any other clarifying information from the spreadsheet before importing—all of this should be on the dataset's landing page, outside of the dataset itself.
    - The portal will only import the first tab in an Excel workbook, so make sure that the full set of data is compiled into one sheet, and that sheet is the first tab.
    - Avoid publishing shapefiles by themselves—when possible include the tabular data. This applies even if the dataset is organized by geography.
5. Import the file. If there are no errors importing the data, the screen will look like this:



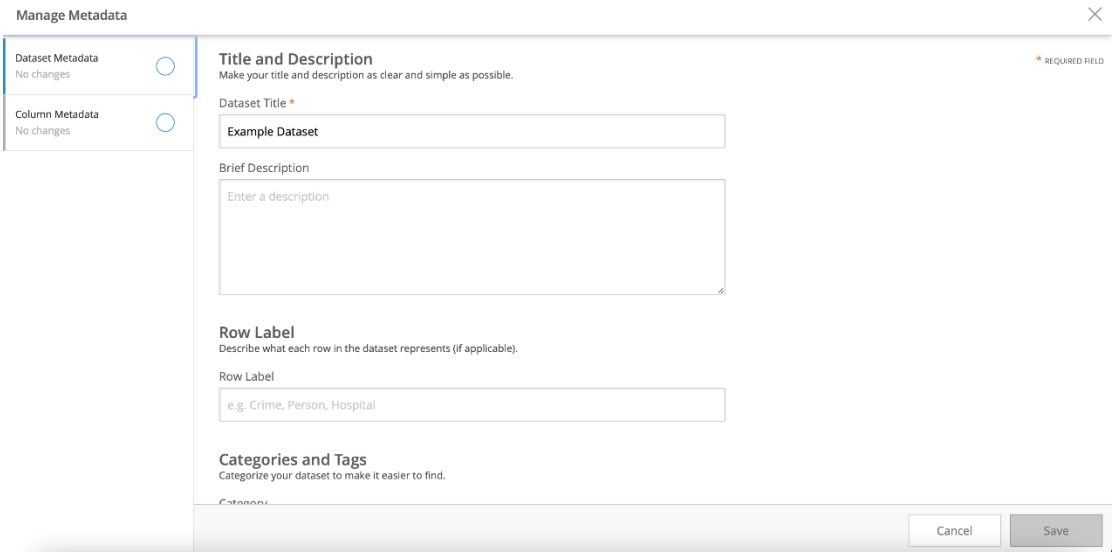
If there are errors, the screen will look like this:



Socrata will infer the data type. To change that data type, use the drop-down menu below the column name. If you would like to use location data, click “Add georeference” and follow the instructions there.

Fix any errors and press “Done.” You’ll be taken back to the landing page.

6. From the landing page, click on “Edit Dataset Metadata.” That will take you to this page:



Fill out every field of metadata that you possibly can. Not only does this help users understand your dataset, but it helps the portal stay organized and improves search functionality, making it more likely that your dataset is seen.

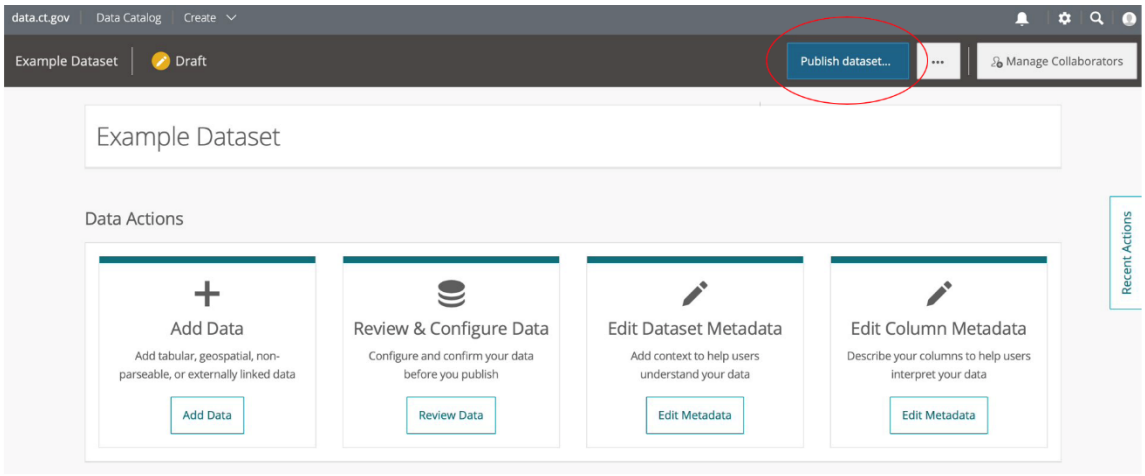
- For a description of every metadata field and its meaning, view the CT Open Data Metadata Guidelines here.

- Metadata should be as simple as possible—the goal of the Open Data Portal is to make data accessible and understandable to all, regardless of technical knowledge. With this goal in mind, use plain language--avoid confusing acronyms, titles, or other language, and provide context in the description areas. Use this checklist for plain language and this thesaurus for plain language. Also use person-centered language (e.g. "person who is incarcerated" rather than "inmate").
- Explicitly identify the license of your data. Generally, data should be identified as "Public Domain." You can also use a "Creative Commons" license, which requires that users attribute the source if they use the data in publications or projects. If applicable law prevents you from dedicating your dataset as public domain, you may have to apply for an Open Data license, but first contact the Chief Data Officer.
- Consider publishing model citations for your datasets in the dataset description.

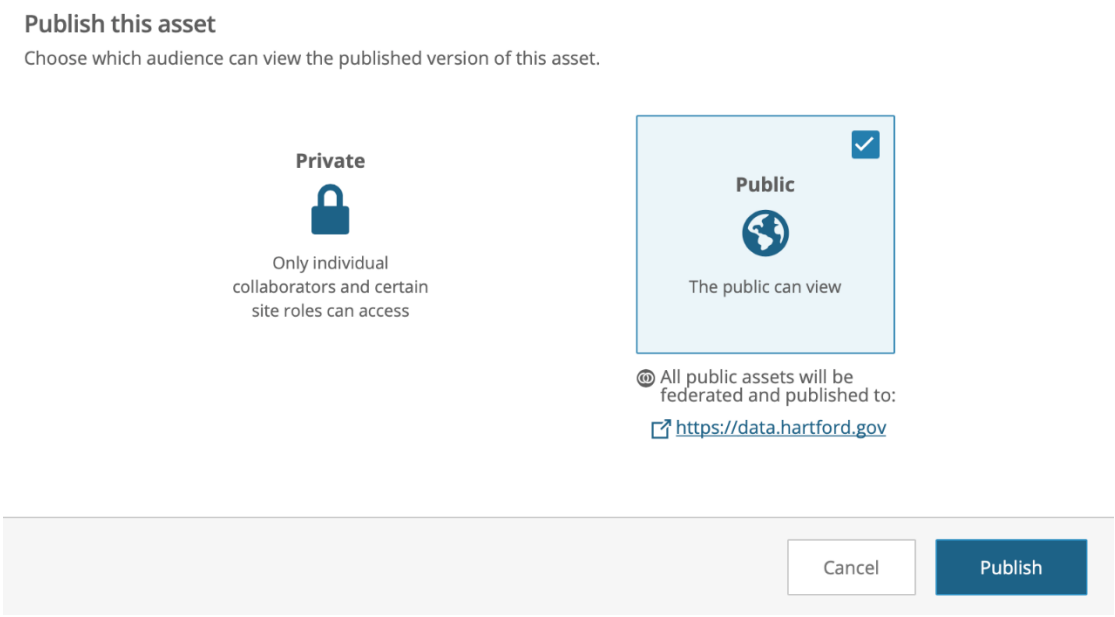
After you are finished with the Dataset Metadata, click "Save," then click "Column Metadata" on the left side of the screen. You can also reach this screen by clicking "Save," then "Done," then "Edit Column Metadata" on the landing page.

For column metadata, make sure to describe what each column measures. Do not just copy and paste your column names into the description field. Your data isn't useful if users don't understand what information you're tracking.

7. After filling out the Dataset **and** Column Metadata, you may publish the data. Click "Publish dataset." Note that you cannot click this button until you have uploaded data and filled out the required metadata:



That will take you to a screen where you can review the changes you've made. Pressing "Continue" will take you to this screen:



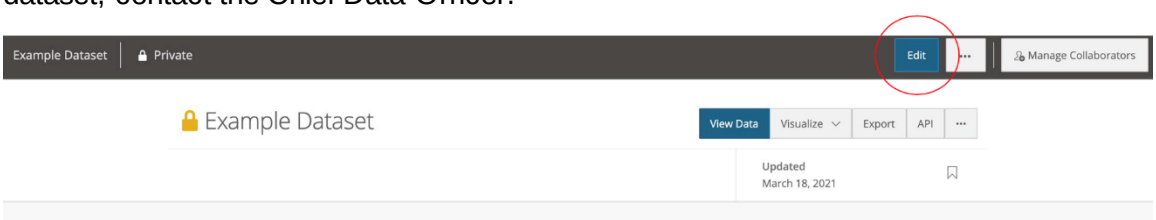
Choose “Public,” then click Publish.

8. If you would like to create a visualization (graph, map, etc.) using your data, scroll to the bottom of the landing page and click on “Create a Visualization.”

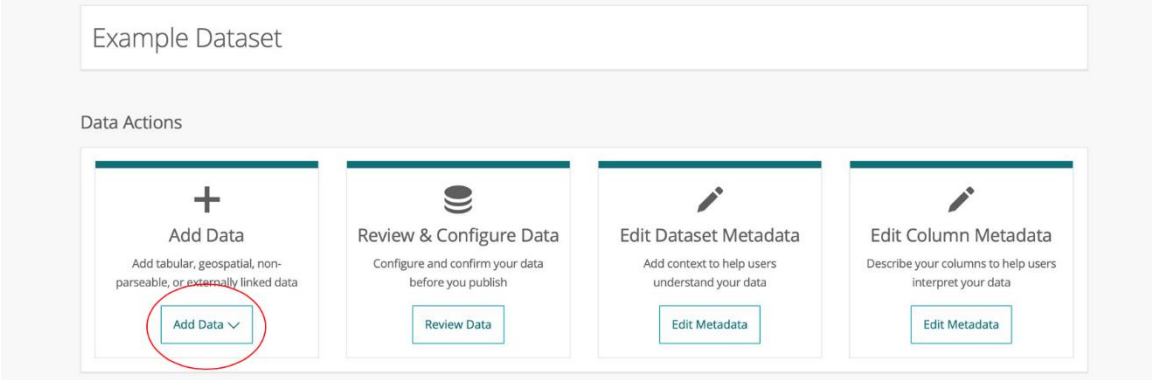
Update Existing Assets (Manually)

It is crucial that agencies update their data on a regular basis and at a predictable frequency. Remember that you should always do your best to consolidate new data with existing data, rather than creating new entries for more recent data. This approach lessens clutter on the portal and makes it much easier to track changes in the data over time.

1. If the dataset contains unique, quasi, or sensitive identifiers and the structure of the data has changed since the last update (e.g. the method of aggregation has changed), refer to the “Privacy Considerations” section of this document to make sure that it is still appropriate to publish as open data.
2. Sign in to the Open Data Portal. If you don’t have an account on the portal, reach out to the Chief Data Officer to get an account with publishing permissions.
3. Go to the landing page. Click on “Edit.” If you do not have editing permissions for the dataset, contact the Chief Data Officer.

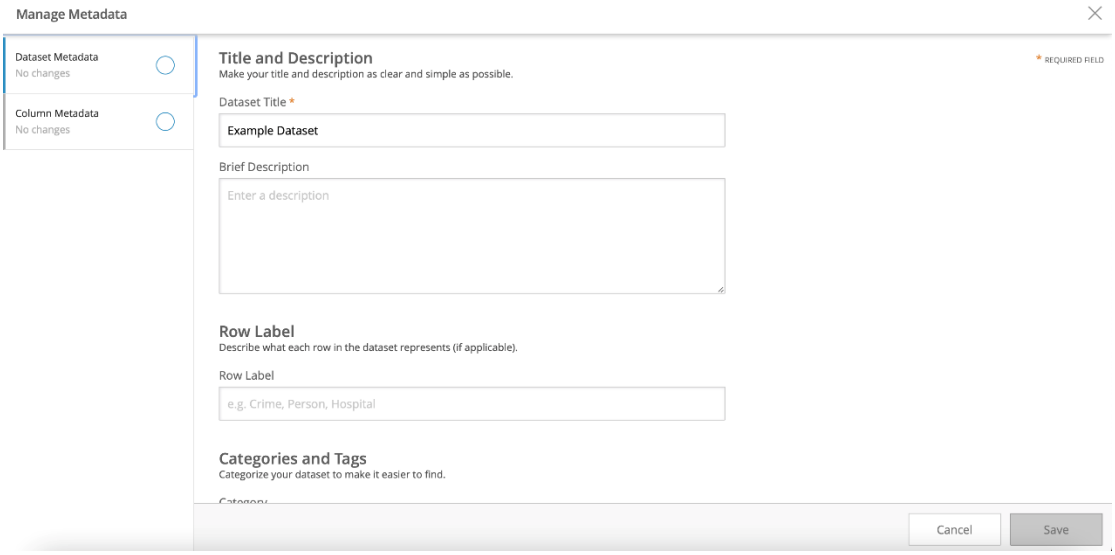


4. Click on “Add Data.”



Select "Append." This adds the new data to the existing dataset. Do not choose "Replace"—if you would like to remove data from the portal, consult the Data Retirement Policy. Follow the instructions to upload the file with the new data.

5. Click on "Edit Dataset Metadata." That will take you to this page:



Fill out every field of metadata that you possibly can. Not only does this help users understand your dataset, but it helps the portal stay organized and improves search functionality, making it more likely that your dataset is seen.

- For a description of every metadata field and its meaning, view the CT Open Data Metadata Guidelines here.
- Metadata should be as simple as possible—the goal of the Open Data Portal is to make data accessible and understandable to all, regardless of technical knowledge. With this goal in mind, use plain language--avoid confusing acronyms, titles, or other language, and provide context in the description areas. Use this checklist for plain language and this thesaurus for plain language. Also use person-centered language (e.g. "person who is incarcerated" rather than "inmate").
- Explicitly identify the license of your data. Generally, data should be identified as "Public Domain." You can also use a "Creative Commons" license, which

requires that users attribute the source if they use the data in publications or projects. If applicable law prevents you from dedicating your dataset as public domain, you may have to apply for an Open Data license, but first contact the Chief Data Officer.

- Consider publishing model citations for your datasets in the dataset description.

After you are finished with the Dataset Metadata, click "Save," then click "Column Metadata" on the left side of the screen. You can also reach this screen by clicking "Save," then "Done," then "Edit Column Metadata" on the landing page.

For column metadata, make sure to describe what each column measures. Do not just copy and paste your column names into the description field. Your data isn't useful if users don't understand what information you're tracking.

6. After filling out the Dataset **and** Column Metadata, you may publish the data. Click "Update" in the upper right section of the landing page. Note that you cannot click this button until you have uploaded data and filled out the required metadata:

That will take you to a screen where you can review the changes you've made. Pressing "Continue" will update the dataset.