

CT Open Data

Data Publication Guidelines

Introduction

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

Steps to Publishing Open Data

This document covers the steps for publishing data on the CT Open Data Portal and the CT Geo 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 should adopt a presumption in favor of openness to the extent permitted by law and subject to privacy, confidentiality, security, or other valid restrictions.

Agencies should use their high value data inventories as the source for identifying eligible datasets to publish as open data. The annual high value data inventories are listed in the [CT Data Catalog](#) published on the Open Data Portal.

The high-value data inventory can inform which datasets to consider for publication, but it is not enough on its own. The inventory catalogs the data you have, but 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 of this document.

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 or the Geo 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 agency³;

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."

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

Open Data Publication Guidelines

- 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.

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.

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 have any questions or concerns about a dataset's privacy risks or need guidance about how to best modify the data, contact the Data and Policy Analytics team 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.
- **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.

Open Data Publication Guidelines

- **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.
- **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.

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 CT Open Data Portal or the CT Geo 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 Portal administrators for each platform, 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 and the CT Geo Data Portal, as well as datasets that are already published elsewhere as open data.

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://aisp.upenn.edu/wp-content/uploads/2022/07/AISP-Toolkit_5.27.20.pdf