Washington State’s open data portal (data.wa.gov) was launched soon after 2009 and is now one of the largest and most well-established state portals in the country. It currently contains over 800 datasets, covering 14 categories from over 30 state agencies. Any of the state’s [197 agencies](https://data.wa.gov/dataset/Washington-State-Agencies-Listing/hsx3-pn9g) can publish to the portal giving it a broad scope and making it a valuable resource for the state. However, while allowing agencies to publish data with very little oversight is likely a driver of the portal’s success, the consequence is a proliferation of low-quality data and metadata. This has made finding the high quality data more difficult and reduces the reusability of data in general.

Consider this scenario. If you lived in Washington State and wanted to see some data on water quality tests you might decide to bypass a Google search and head straight to data.wa.gov. However, a search for “water tests” will return [over 500 results](https://data.wa.gov/browse?q=water+test&sortBy=relevance), many of which are pdf documents concerning lead tests in school water systems. Those documents, while very important, barely meet the first star of the [five stars of open data](http://opendatahandbook.org/glossary/en/terms/five-stars-of-open-data/). As savvy as you are, you filter the results to [just datasets](https://data.wa.gov/browse?limitTo=datasets&q=water+test&sortBy=relevance) and get four results, one of which shows [lead test data](https://data.wa.gov/Health/WA-School-Water-Lead-Test-Data/i3jn-y8vx) for all schools in the state. Not exactly what you started looking for but still interesting. Closer inspection shows that the metadata for that dataset is partially complete with a short description, the name of the owner, and the date updated. Those metadata are enough to convey what the data are but there are still key pieces of information missing, such as how frequently the data are updated, what license the data are published under, and what each column in the dataset means (the data dictionary). These gaps might mean you still need to contact the dataset owner to get more information before using the data. As it turns out, even missing those metadata elements, this dataset is of high quality compared to many of the other 800 datasets available.

The WA Office of the Chief Information Officer (OCIO) is the managing agency for data.wa.gov and in recent years has become aware of the proliferation of partly filled in metadata or poor-quality metadata. Data.wa.gov is not along; quality issued are a global problem in open government data portals (Kubler). In 2016, the portal software company Socrata [surveyed developers](https://benchmarkstudy.socrata.com/Developer-Survey/Developers-Rate-the-Current-State-of-Gov-Data-Acce/4j6z-qhyp) on their perceptions of open government data (OGD) and found that over half thought that metadata is inconsistent between datasets, data is not up to date, and data is not clean or accurate.

To address metadata quality issues on data.wa.gov, the OCIO approached the [Washington State Library](https://www.sos.wa.gov/library/) (WSL) and the Open Data Literacy program for help. My internship this summer was the first step in a potential partnership between the OCIO and the WSL to curate data.wa.gov to make it more usable.

This project consisted of two components: interview state agencies to understand data publishing behavior on the portal and assessing the current state of metadata quality on the portal.

Interviews

I interviewed eight agencies and one organization that uses data from the portal. Most agencies held positive views of the data.wa.gov and it became clear that every agency uses the portal to meet their unique data needs. In fact, publishing behavior is only generalizable to the extent that it is unique to every agency. Publishing behavior included publishing for specific users, publishing for transparency, and publishing for internal or interagency use. Some agencies know exactly who their main users are while others can only guess or have no idea. The primary users are not individual citizens. Rather they are other agencies, local governments, federal agencies, and 3rd parties such as businesses, nonprofits, and the media. Seven out of the eight interviewed publishers plan on continuing or increasing publishing on the portal. This emphasizes the need for curation as the number of datasets on the portal will likely increase.

The varying needs and publishing behaviors of the agencies suggest a curator will need to work closely with publishers to encourage better metadata practices. Any broad attempt to increase metadata quality on the portal will affect agencies in different ways and may produce unintended consequences.

Metadata Assessment

There are many ways to assess metadata quality (e.g. Kubler et al. 2018). I assessed metadata quality through completeness and understandability using a combination of variables from multiple other studies. Data.wa.gov is built on the [Socrata software platform](https://www.tylertech.com/products/socrata) and Socrata offers several APIs to collect information from portals. I used Python to access Socrata’s Discovery API for basic metadata and looped through the JSON files for every published asset to collect detailed metadata information.

While the OCIO provides excellent guidelines on open data best practices for data publishers, agencies are only required to fill out the title of a dataset on data.wa.gov. Perhaps for that reason, over half of the published datasets are missing 50% or more of the 11 available metadata elements. About 19% of datasets only have the title. The least filled out elements include license, posting frequency (how often the dataset is updated), period of time, originator, and the metadata language (typically English).

It is clear that agencies only fill out a fraction of the metadata, so it would be useful if they were filling out metadata that provided the most useful information. I selected five elements that if filled out properly would allow a dataset to pass the [CRAAP test](https://en.wikipedia.org/wiki/CRAAP_test). These are attribution (publishing agency), description, category, posting frequency, and license. Of these elements, the former three are filled out in some combination in about 60% of datasets, but 69% of datasets are missing posting frequency and 67% are missing license. A full 21% do not include any of these five elements.

I looked at a sample of 112 datasets to assess the understandability of what metadata agencies filled in. If information was confusing or incorrect it was scored as difficult to understand. Forty percent of dataset titles are enigmatic. Half of the sampled datasets had confusing temporal or spatial information, often both. Only 25% of datasets include a data dictionary (column descriptions).

The search experience and usability of data.wa.gov would be much improved if every dataset had at the least all five core elements completed with understandable information. Only seven percent meet all these criteria, 48% are either missing or have enigmatic information in one to two core elements. A curator could efficiently improve a significant portion of the metadata needs on the portal by focusing on this 48%.

There are other curation needs in addition to metadata quality. About 13% of sampled datasets contained the same data as one or more other datasets but from a different time period. These datasets should be combined to one dataset that is regularly updated. About eight percent of datasets are test or dummy datasets that should be unpublished or removed. Three percent of datasets do not meet the [OCIO’s definition of data](https://ocio.wa.gov/programs/open-data/guidance-open-data-definitions).

The takeaway from this work is that data.wa.gov is an important resource for state agencies but more than half the datasets need metadata improvements or other curation work. The data catalog needs curation to maintain its usefulness. However, data.wa.gov should not be compared to a centrally managed library catalog of books that can be developed and weeded as needed. A curator of the portal will have no control over what gets published and will have to focus on making incremental changes to maximize improvements. A curator should focus on finding ways to make it easy for publishers to include high quality metadata, should find ways to make the portal visible for both publishers and potential users, and develop a removal policy and procedure.

Here are specific recommendations to help guide a future curator:

1. Focus on the five core metadata elements
   1. Include controlled vocabularies when possible
   2. Provide metadata explanations to help publishers fill out metadata elements
2. If the partnership with the WSL becomes official, the WSL should advertise its curatorial role on the portal to encourage user feedback- putting a face to a feedback system may increase interaction
3. Remove datasets when necessary
   1. Implement a policy to identify datasets for removal
   2. Create a procedure that removes datasets in a transparent way
4. Run the assessment script available on GitHub several times a year to evaluate how curation efforts are affecting metadata quality

**What’s Next**

There is much more behind this analysis and it is all available on the ODL GitHub. This project was a massive learning experience for me. The end result is a collection of information that will help guide future curation efforts; however, the really useful information will come from evaluating future curation efforts and documenting successes and failures.