Curation Assessment of the Washington State Open Data Portal

Andrew Mckenna-Foster

2019

An Open Data Literacy Project

Sponsor:

Kathleen Sullivan – Washington State Library

Contents

Executive Summary

The Washington State 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, more than double the number in 2018, in 14 categories from over 30 publishers. Any of the state's 197 agencies can publish to the portal with few restrictions, resulting in a broad range of data and varying data and metadata quality. This model of unmediated deposits is likely a driver of the portal's success, but the consequence is the presence of poor-quality datasets. The solution may be to actively curate the portal.

The Office of the Chief Information Officer (OCIO) manages the portal and has partnered with the Washington State Library (WSL) and the Open Data Literacy (ODL) program at the University of Washington Information School to assess the quality of metadata and data on data.wa.gov. With the ultimate goal of providing information that can inform future curation work on data.wa.gov by the WSL, this report is the first step in the collaboration.

This project consists of two data collection components: gathering information on publishing habits by interviewing a sample of the agencies that publish to the portal and assessing the quality of data and metadata on the portal. I interviewed eight agencies and one portal user. I assessed all the datasets on the portal using a framework of five dimensions: Format, Discovery, Contact, Temporal Information, and License. Within those dimensions, I evaluated all datasets on metadata existence and a sample of datasets on metadata understandability. I also identified a set of five core metadata elements that are particularly helpful for finding, understanding, and using a dataset: Description, Category, Update Frequency, Data provided by (or Attribution), and License.

Agencies publishing data on data.wa.gov overall find it a useful and important resource that helps them achieve their goals. Publishing behavior is only generalizable to the extent that it is unique to every agency. The open data portal is essential to the operations of some agencies, a convenient tool for other agencies, and not considered useful at all by a minority. There was a general interest from agencies on how to use the portal in better ways and each could list pros and cons from their experiences publishing data. Interviewees suggested state agencies, local governments, media, and various nonprofits are likely the heaviest users of data on the portal.

Results show that metadata quality on data.wa.gov is far from perfect, similar to many open government data portals around the world. Of the published datasets, 75% are missing half or more of the available metadata elements. Nineteen percent of datasets do not include any optional metadata and only have a dataset name. Publishers fill in 2.6 ± 1.7 (SD) of the five core metadata elements and 21% of datasets have none of the core elements. Sixty-two percent of publishers provided some indication of what department published the data, but the entries are not standardized, and it is difficult to easily summarize the information. Posting frequency and License are the least filled in of the core elements. Licenses are only included with 33% of datasets making reuse difficult. As for understandability, about 40% of datasets have enigmatic titles and temporal and spatial information is often confusing. Fewer than 25% of datasets have a data dictionary.

There are many ways a curator can improve metadata and data quality on data.wa.gov. I propose curation activities that could efficiently improve quality without undue burden on publishers or the curator, and that increase the visibility and use of the portal.

Metadata and Data Quality Recommendations:

- Establish a policy and procedure for removing datasets
- Remove test and dummy datasets
- Add a controlled vocabulary for 'Data Provided By' and 'Posting Frequency'
- Remove the 'Period of time' and 'Originator' elements
- Determine if adding custom tool tips/help call-outs to the metadata entry form is possible
- Assist agencies aggregate datasets that are part of a time series
- Run the metadata analysis several times a year

Visibility and Use Recommendations:

- Create a direct user feedback option on the homepage for data.wa.gov. This should highlight that the query is going to a librarian
- Identify datasets or groups of datasets that would make good stories
- Advertise the portal to potential users outside the government- nonprofits, teachers, political groups

All the resources, code, and datasets used for this project are available at https://github.com/OpenDataLiteracy/WSL-AMF. Interview recordings and notes are stored in a private GitHub repository accessible to the partners on this project.

Introduction

Open government data (OGD) as an official concept in the USA is just over 10 years old. In 2009, President Obama signed a presidential memorandum making open data part of the US government's regular operations (White House Memorandum 2009). Soon after 2009, open data portals proliferated at city, county, tribal, and state levels throughout the USA. Agencies at all levels were encouraged to inventory their datasets and publish what they had available, sometimes regardless of whether it was complete or perfect (e.g. see guidelines in the Open data handbook). In the last five years, a growing concern for data and metadata quality has emerged across the globe and there are many efforts to assess the quality of data portals. Socrata, a software platform commonly used by governments and the current platform for data.wa.gov, conducted a 2016 survey of developers and their use of open government data. They found that half of respondents thought that there were problems with metadata consistency between datasets, outdated data, and data cleanliness and accuracy. Their responses are supported by multiple studies that suggest many OGD portals contain significant numbers of datasets that have poor quality metadata and data (Umbrich et al 2015, Vetrò et al. 2016, Kubler et al. 2018).

Washington State Open Data Portal

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, more than double the number in 2018, in 14 categories from over 30 publishers. Any of the state's 197 agencies can publish to the portal with few restrictions of data and metadata quality, resulting in a broad range of data and varying data and metadata quality. This model of unmediated deposits is likely a driver of its success, but the consequence is the presence of poor-quality datasets.

The Office of the Chief Information Officer (OCIO) manages the portal and was aware of quality issues. There are old datasets that have not been updated, datasets with very little metadata, and datasets in formats that are not machine readable. The OCIO was left in a quandary: how can resources be directed to efficiently increase metadata and data quality while maintaining low barriers to publishing?

OCIO initiated discussions with the Washington State Library (WSL) and the Open Data Literacy (ODL) program at the University of Washington Information School on how the portal could be curated by WSL. WSL has a long history curating government documents and state related information and is a natural candidate for a curatorial role. In recent years it has taken a leading role in developing open data knowledge and practice in libraries around the state.

These discussions led to the formation of this project with a goal of gathering information on portal use and metadata/data quality to inform the possible partnership between WSL and OCIO.

The following questions guided the project scope:

1. How do agencies view the portal and what are their experiences publishing data?
2. How well are WA state agencies filling in metadata fields?
3. Do datasets meet the basic standards to be interoperable?
4. Are titles, descriptions, and keywords understandable and useful?
5. What might curation and collection development look like for the state portal?

The Metadata Quality Landscape

The Washington state open data portal is not alone in this metadata quality problem. There are many recent papers that reveal this to be a global issue with no consensus on how to address it. These papers typically attempt to define the scale of metadata quality issues and propose ways to benchmark metadata quality for future comparisons by defining various frameworks.

Kučera et al. (2013) is an early attempt to address metadata quality and proposes using the overarching dimensions of Accuracy, Completeness, Consistency, and Timeliness to assess both portals and portal datasets. Velijkovic et al. (2014) assessed the national data.gov portal and gave it a 67.5% rating for openness and completeness. They only looked at descriptions to assess completeness. Umbrich et al. (2015) designed a system to continuously monitor 82 CKAN portals from 35 countries using core metadata elements (title, description, tags, license and organization) and other available metadata elements categorized into six dimensions (Retrievability, Usage, Completeness, Accuracy, Openness, and Contactability). They found that a majority of portals had a completeness level of 60-80%. Vetrò et al. (2016) compared datasets within specific dataset categories from local level portals of mixed quality and a national level portal of recognized high quality. The municipal portals had poorer metadata quality under seven dimensions compared to the centralized portal, but overall compliance with the e-Government Metadata Standards was high. However, they identified the lack of data dictionaries, lack of dataset update information, and large numbers of empty cells or rows as the main quality issues. Máchová and Lnénicka (2017) used human assessors to rate national portals under approximately four dimensions that encompassed Technical Aspects, Availability, User Communication, and Dataset Metadata. They conclude that national portals show a lack of harmonization and there is a need for metadata standardization. Kubler et al. (2018) presents one of the most comprehensive assessments using an automated system to monitor 250 portals from three software platforms. They use five dimensions (Existence, Conformance, Retrievability, Accuracy, and Open Data), each with multiple subdimensions and a total of 21 variables. They report that many datasets are missing license, temporal, and spatial information and conclude that portal manager do not pay enough attention to metadata quality. Finally, the OpenDataMonitor, a project partly funded by the European Commission, provides continuous monitoring of national level OGD portals in Europe. As of August 2019, it rates metadata completeness across portals at 58%. This completeness metric accounts for license, author, organization and either date released, or date updated.

All of the recent attempts at metadata quality assessment agree that low quality metadata is a widespread problem. The fact that there are many ways to define quality metadata and that there is no agreement on the best combination of dimensions or the granularity of variables within those dimensions, makes the scale of the problem and any potential solutions difficult to understand. It may be that more comprehensive assessments will be useful for local level portals while simpler assessments will better meet the needs of national or global aggregating portals. This project aims to work at a level that is balance between detailed assessment and actionable information.

Methodology and Results

This project consists of two data collection components: gathering information on publishing habits by interviewing a sample of the agencies that publish to the portal and assessing the quality of data and metadata on data.wa.gov.

Agency Interviews

I selected agencies for interviews to represent a wide range of publishing habits including frequency of publishing, frequency of updating datasets, number of datasets published, and the number of dataset downloads. I conducted interviews under a University of Washington Institutional Review Board approval. From each interviewee I obtained verbal consent to both conduct and record the interview. The interview questions are in Appendix A. Details from these interviews will be used by the OCIO and WSL to inform future curation work.

Agency Publishing

Interviews with agency representatives show that data.wa.gov serves many purposes. Publishing agencies have disparate levels of sophistication and they have varying expectations. Several have incorporated data.wa.gov into their core operations while others only use it sporadically. Publishing behavior is only generalizable to the extent that it is unique to every agency.

Why Publish

All interviewed agencies have an open data plan as required by the State. The open data portal is essential to the operations of some agencies, a convenient tool for other agencies, and not considered useful at all by others. All agencies were interested in how to use the portal in better ways and could list pros and cons from their experiences publishing data. Several agencies use the portal to meet transparency and data storage needs and do not know who most of their users are. In some cases, an agency will regularly maintain their datasets for the unknown users while in other cases datasets get published as part of an open data push and then are no longer updated. Two agencies know who their main users are and publish with them in mind. At least one interviewed agency uses the portal as both an intra and inter-agency data sharing tool.

Publishing Behavior Examples

The wide variety of publishing habits is visible in the collection of datasets on the portal. Examples of agency behavior include:

- Publish one dataset that is updated regularly
- Published a group of datasets that have never been updated
- Regularly publish data for inter or intra agency use (data is public but has poor metadata)
- Manage multiple datasets that are updated multiple times each day

Publishing Plans

Three agencies said their publishing on data.wa.gov will increase, two will remain the same, two are uncertain, and one expects a decrease in publishing. A majority of the agencies were interested in learning about using the portal more effectively or hearing about future open data developments.

Portal Feedback

Almost all the agencies reported that they are constantly trying to find better ways to share visualizations, insights, and information both with agency staff and external users. One of the most common complaints is that the visualization tools on the Socrata platform are more limiting compared to other software. PowerBI and Tableau were mentioned in a majority of interviews as software agencies either use or are trying to use.

Two agencies listed poor search experience and two listed the lack of a data suppression feature as downsides to data.wa.gov. The extra curation steps of masking or removing values before uploading to data.wa.gov to protect individuals may prohibit publication of data.

On the other hand, most publishers find the portal interface easy to use, very cost effective (the platform is paid for by the OCIO), and a good way to make data available. Agencies mentioned the easy to learn web interface, the ease of filling in metadata, useful assistance from Socrata representatives, and useful API features. The user I interviewed also highlighted the useful filtering options on the Socrata platform.

A few agencies know what features would help them use the portal more. One agency wondered if searching by agency would be possible. Two agencies mentioned the portal as a source for authoritative datasets and one of the agencies wondered if it were possible to certify datasets as up to date. The need to compare data from different levels, such as county data and school district data was mentioned in two interviews.

Users

Users of data.wa.gov are varied and still obscure. Interviewees suggested state agencies, local governments, media, and various nonprofits are likely the heaviest users of data on the portal. Some of the most accessed datasets feed search functions, visualizations and summaries on agency websites. According to the portal metrics, most of the data use on the portal is through an API. A majority of data transferred from the portal to users may flow through agency websites first. Agencies use the portal to provide data to the media, identify what information other agencies may have available, or gather information from other agencies for their own use. However, only one agency specifically mentioned using data directly from the portal.

Individual citizens are the most obscure users and are likely the least common users. Only one agency was aware of a private citizen directly using their data from the portal.

Out-of-state users appear to be federal agencies (EPA, fisheries) or private data aggregators such as followthemoney.org. or USAFacts. I talked with one data aggregator who stressed the importance of having access to complete datasets. An agency might choose to make a subset of data fields available as open data not realizing the omitted fields are essential for some users.

Impact for a Curator

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 sweeping attempts to increase metadata quality on the portal will affect agencies in different ways and may produce unintended consequences.

Metadata and Data Quality Assessment

Data.wa.gov is built on the Socrata platform which offers fourteen metadata elements for dataset publishers along with the option to include a description for every column in the dataset (the data dictionary). Other important elements, like date created, date updated, number of downloads and number of page views are automatically filled. Uploads can be tabular data or documents (e.g. pdf or MS Word). Socrata offers multiple APIs to access data from portals and also includes a dataset containing most of the metadata for every asset (i.e. dataset, file, chart, map, etc) on the portal.

I examined data and metadata quality at two levels of assessment:

1. Metadata existence
2. Metadata understandability and data quality

Metadata existence assessment is based on 473 datasets that were available on data.wa.gov on August 5, 2019 (over 300 recently added datasets tagged with "reportcard" were not used in this analysis). Understandability and data quality assessment is based on datasets sampled from those available June 18, 2019.

Metadata Existence

I built an assessment framework informed by other assessment studies. These studies use a variety of strategies and rationales to evaluate portals. Each assess the completeness of slightly different combinations of metadata elements. I created a framework based on five dimensions that are used most commonly: Format, Findability, Contact, Date, and License. Data.wa.gov automates aspects of the first four and provides the publisher additional options within each (Table 1).

Table 1. The metadata assessment framework is based on a synthesis of other recent assessments. On data.wa.gov, some metadata elements within the assessment dimensions are required or automatic, while others are optional.

Dimension	Typical Elements	Kučera et al. 2013	Umbrich et al. 2015	Vetro et al. 2016	Kubler et al. 2018	ODM* 2019	Data.wa.gov Required or Automatic Elements	Data.wa.gov Optional User Elements
Format	Machine Readable Open File Format	✓	✓	✓	✓	✓	Meets open criteria if uploaded as tabular	User Choice
Discovery	Title Description Keywords	✓	✓	✓	✓		Name	Description Keywords Category
Contact	Email Author Organization URL		✓	✓	✓	✓	Upload Contact (Owner)	Data Provided By Originator
Temporal	Date Created Date Updated Temporal Information	✓		✓	✓	✓	Date Created Date Updated	Posting Frequency Period of Time
License	License	✓	✓		✓	✓	None	License

*ODM = Open Data Monitor

For dataset level metadata existence, I created a Python script in a Jupyter Notebook that downloads basic metadata for published assets from the Socrata Discovery API as well as more detailed metadata elements from the API endpoint for each public dataset. The script creates CSV files of the metadata and appends summarized data to another CSV file for longitudinal analysis (there is an additional script that summarizes this file). Additionally, this Python script can be used on any Socrata portal with slight modifications to properly ingest custom metadata elements added by portal managers.

Format

The Format dimension is assessed through the file format and whether it is nonproprietary and machine readable. It is also important that the data is structured with consistent columns and rows. Approximately 25% of the assets on the portal are not in a machine-readable format and are mostly pdfs or MS Word documents. Socrata makes any uploaded tabular data available in multiple open formats. However, of the tabular data, 12% (13/108) are not structured appropriately for machine readability. They have summary rows among data rows or may contain rows that are not part of the dataset.

Metadata Existence Methods and Results

Of the fourteen available metadata elements, the name of the dataset is the only element currently required by data.wa.gov. Additionally, the Socrata platform provides a way to contact a dataset owner that keeps the owner identity anonymous unless the owner specifies their contact information in the Contact Email element. For this reason, I ignored the Contact Email metadata element in the assessment. A publisher can also add up to six notes. I did not include these in the existence assessment because they are completely open ended and difficult to standardize. That leaves eleven optional available metadata elements for analysis (Table 2).

Of these eleven options, five are particularly helpful for finding, understanding, and using a dataset: Description, Category, Update Frequency, Data Provided By (or Attribution), and License. These will be referred to as core metadata elements. Originator is not included since most publishers put this information in the 'Data Provided By' element. The data dictionary completeness is included in the data quality assessment described in the next section.

Table 2. Metadata elements available to data publishers on data.wa.gov and what elements were included in the assessments. Each was evaluated for presence. The core elements, if filled out properly, would allow a dataset to be understandable and reusable.

Element	Included in Available Assessment	Included in Core Assessment	Dimension
Categories	Categories	Categories	Discovery
Description	Description	Description	Discovery
Data Provided By	Data Provided By	Data Provided By	Owner
Posting Frequency	Posting Frequency	Posting Frequency	Temporal
License	License	License	License
Period of Time	Period of Time		
Row Label	Row Label		
Keywords	Keywords		
Source Link	Source Link		
Originator	Originator		
Metadata Language	Metadata Language		
Name			
Notes			
Alternate Contact Email			

As the OCIO expected, metadata on data.wa.gov needs improvement. Of the published datasets, 75% (357/473) include six or fewer metadata elements (Fig. 1). Nineteen percent (88/473) of datasets do not include any optional metadata and only have a dataset name. There is a clear spike at six metadata elements and these datasets most likely are including Category, Tags, Description, Attribution, Attribution Link, and License.

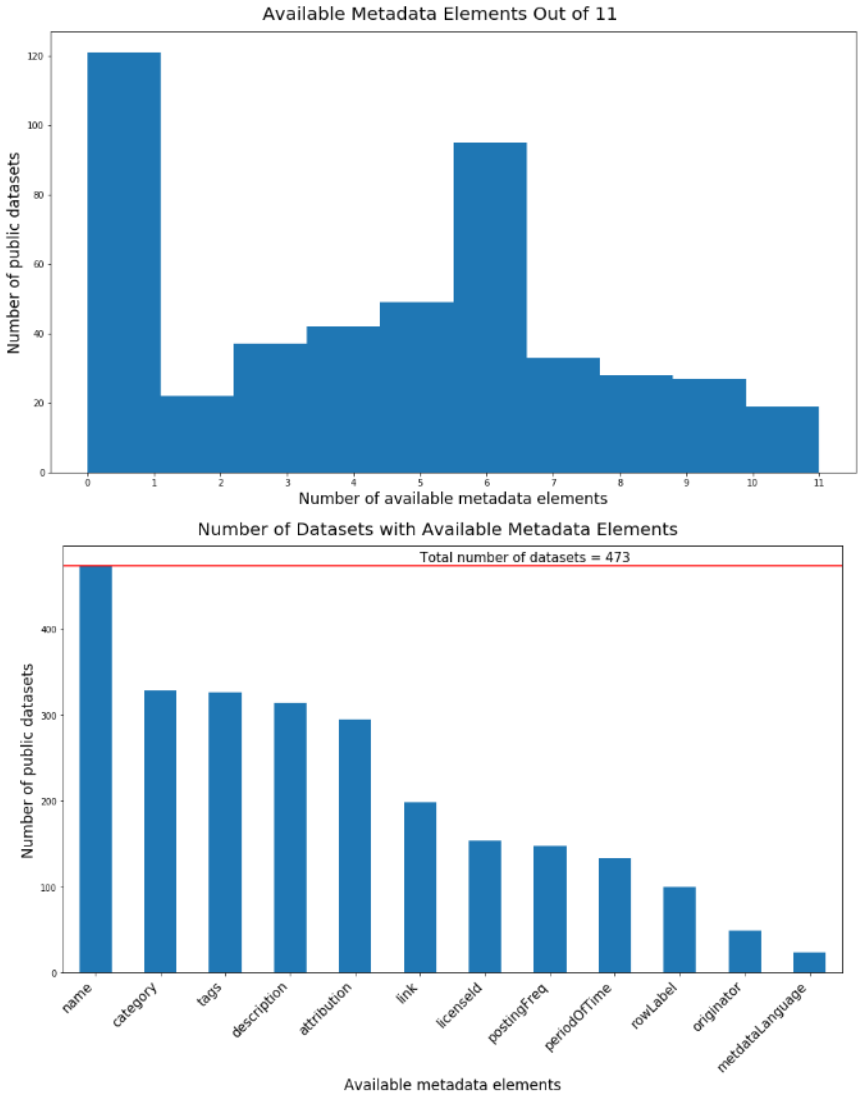

Figure 1. Available metadata elements filled out by data publishers on data.wa.gov.

Publishers fill in 2.6 ± 1.7 (SD) of the five core metadata elements and 21% of datasets have none of the core elements. Sixty-two percent of publishers provided some indication of what department published the data, but the entries are not standardized and it is difficult to easily summarize the information. Posting Frequency and License are the least filled out of the core elements. Licenses are only included with 33% (154/473) of datasets. State agencies publishing on data.wa.gov may purposively not include license information because they assert that they do not own the data in a way that allows them to license it.

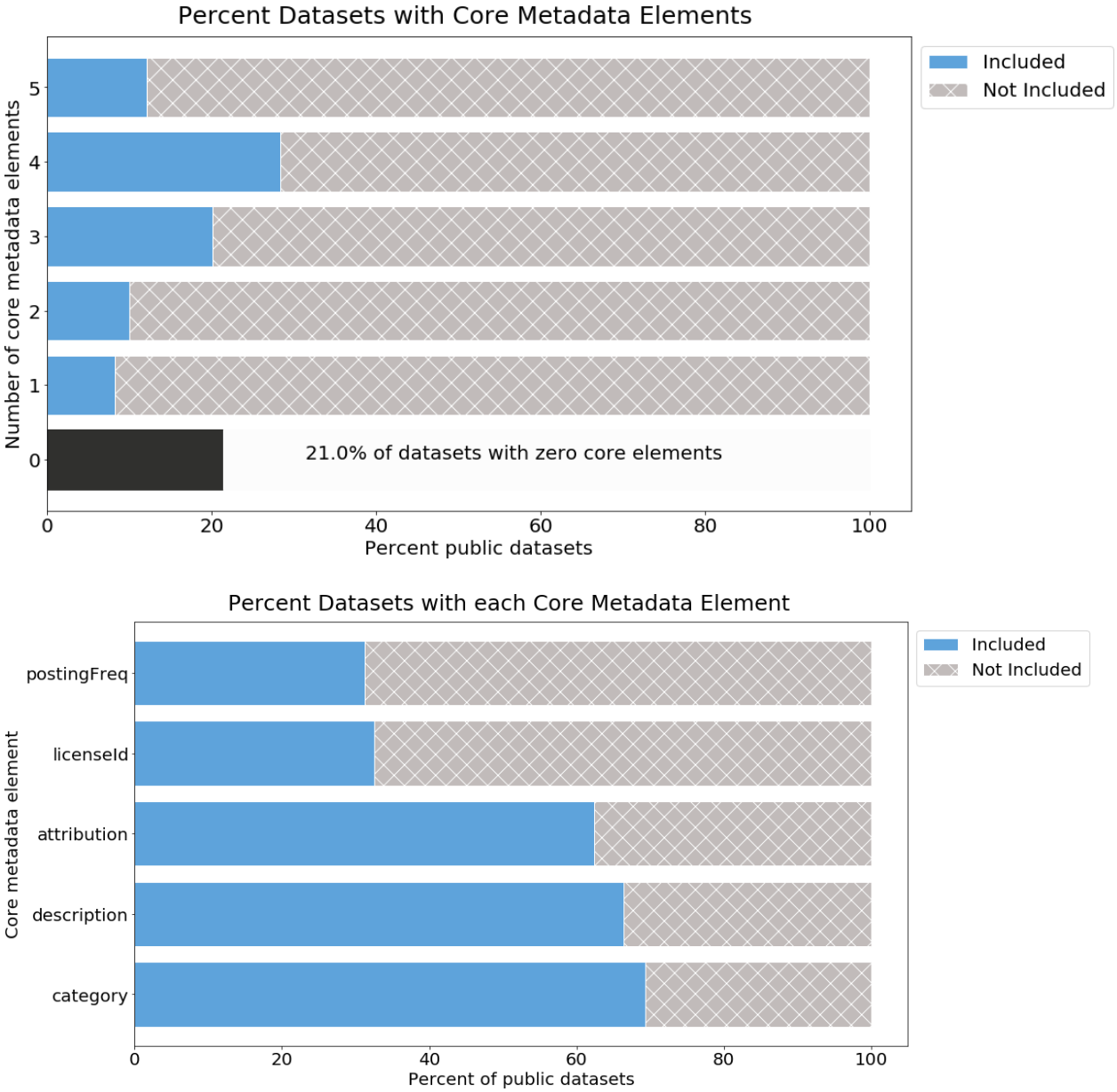

Figure 2. The number of core metadata elements filled in (top) and what percentage of datasets fill in each of the core elements (bottom).

Metadata Understandability and Data Quality

I used a stratified random design to select a sample of datasets from published datasets available on June 18, 2019. I divided the list of datasets into groups by the number of downloads (<100, >101 but <1500, and >1500) and then by year created (2012/2013, 2014/2015, 2016/2017, and 2018/2019) for a total of 12 groups. After randomly ordering the datasets within each group, I selected the first 10 for each group. Some groups did not contain 10 datasets and one selected dataset was the asset list, which I removed, leaving 116 datasets. Four of these datasets had been removed from the portal by the time I tried to retrieve them leaving a final sample of 112 datasets, representing 25% of published datasets. I visited the URL for each dataset, downloaded the dataset, assessed the metadata understandability,

data quality, and the data dictionary. Since assessing understandability is subjective, I used a very basic ranking system to assign scores (Table 3).

Even when a metadata element is filled out for a resource, the content may not always be understandable. Half of the sampled datasets have enigmatic information for at least one of the studied areas (Fig. 3). This could affect the trustworthiness and ultimately the reusability of data. The OCIO provides some guidance on producing understandable metadata. More detailed guidelines may be useful.

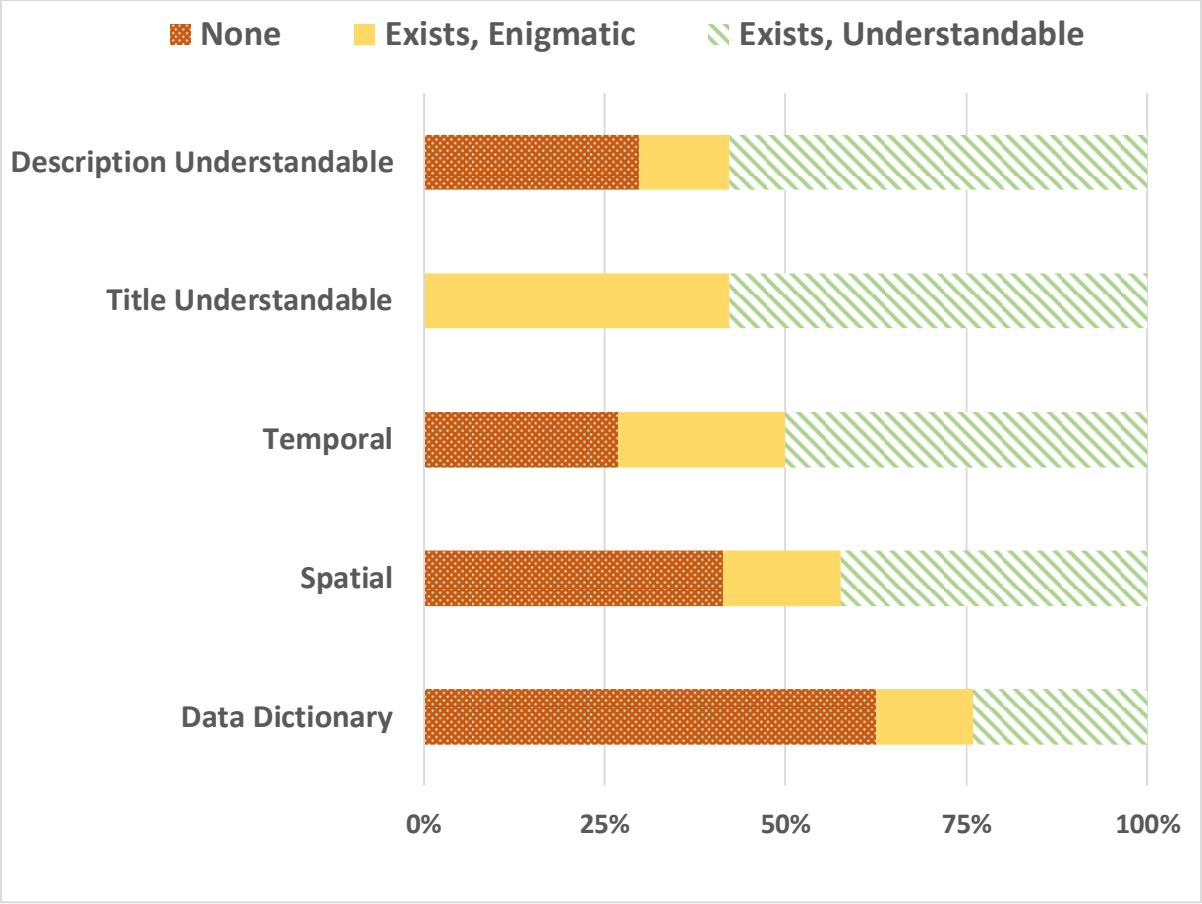

Figure 3. Understandability assessment for information contained in metadata elements.

Table 3. Metadata elements included in the understandability assessment and the criteria used for each.

Metadata Element	Assessment	Criteria
Title	Understandability	(1-No, 2-Yes) - contains subject, context and important qualifiers (e.g. "Snake Barriers Removed" = 1 contains subject and qualifier, but missing context; "Washington Fuel Tax Active Licensees" = 2, an informative title.
Description	Existence and Understandability	0=none 1=No, 2=Yes - explains purpose and history of dataset. If missing either of those, 1.
Each Row is…	Existence and Understandability	0=none 1=not understandable, or partial 2=yes, understandable
Data Dictionary	Existence and Understandability	0=none 1=not understandable, or partial 2=yes, understandable
Spatial	Existence and Understandability	0= none 1= Spatial extent exists but unclear 2= Spatial extent is clear (e.g. "Washington Fuel Tax Active Licensees")
Temporal	Existence and Understandability	0=none 1=Temporal information exists but is unclear or conflicts with itself 2= Temporal information in temporal elements, title or description.
Dataset Retrievable	Download success	0=no 1=yes
Resource Retrievable	Access URL functions	0=no 1=yes
Data Understandable	Are the values in the cell understandable without a dictionary? (money, counts, county names, etc.) (e.g. okay if you don't know what the money value is for) Ignore missing contextual metadata	0=missing 1=no 2=yes
Cell Accuracy	Number of columns with problems (same values spelled the same, consistency in format within column)	0= >10 columns have accuracy issues 1= 4-10 columns 2= 1-3 columns 3= all columns accurate
Curation Needs	Number of columns in need of curation (summed columns removal, summed rows removal, untidy data format, type formatting, special characters.)	0= >10 columns need curation 1= 4-10 columns 2= 1-3 columns 3= No curation needed
Currency/Usefulness	How current or useful is the dataset?	0= old/needs update 1 = old/still useful (like data about an event) 2= up to date/useful
Belongs on Portal?	Would the information be more useful in another delivery package?	1=yes, 0=no - is it data? more usable as a static webpage?

Title

About 42% of titles are difficult to understand (Fig 3). Ideally, an understandable title should include the subject, spatial information, and what was measured while avoiding the use of acronyms (Table 4). Opaque titles many times appear to be for datasets that were created for an internal report or to share with another agency. It appears the publisher created the title thinking of a specific group of users rather than an anonymous user.

Table 4. Examples of titles on data.wa.gov that are not understandable with corresponding suggested edits.

Current Title	Proposed Title
Master Content	Washington Water & Salmon Fund Finder Grant and Loan Information
DEL Office Locations	Department of Early Learning Office Locations
Coast Complete Metrics	Coastal Stream and Estuary Restoration Metrics
Imaged Documents and Reports	Index of Digitized Documents and Reports from the Public Disclosure Commission

Description

About 30% of datasets were missing descriptions and an additional 13% had enigmatic descriptions. A useful description should explain why data was collection, how data were collected, and include any caveats useful for interpreting the data. Explaining missing values is particularly useful.

Temporal and Spatial Information

About 28% of the sampled datasets were either missing or had enigmatic temporal or spatial information. Another 41% had issues with both. Temporal values are not always put into the two available temporal metadata elements (Posting Frequency and Period of Time) and often appear in the title or description. About 15% (16/108) of sampled datasets have cryptic or confusing temporal information. Often, the values in Posting Frequency and Period of Time conflict with information in the Title or Description or with the Date Updated element.

Data Dictionary

Over 60% of sampled datasets did not have a data dictionary. An additional 15% had a data dictionary that was either incomplete or difficult to understand.

Core Elements and Understandability

The search experience and usability of data.wa.gov would be much improved if every dataset had all five core elements completed with understandable information. Only seven percent meet all these criteria, 28% are either missing or have enigmatic information in one core element, and 71% have two or more elements missing or containing enigmatic information.

Record Level Data Quality

Data quality at the record level is overall pretty high on data.wa.gov. Most tabular datasets are in a tidy format with the most common curatorial need being filling empty cells.

Keyword Selection

Good keywords increase findability for a resource and should offer alternative search terms than what is in the title and description. About 69% of public datasets contain keywords. Most of these datasets

have keywords that overlap with words in the title or in the data provided by section. This could be beneficial for portal users searching for datasets through the API and restricting their search to keywords. Currently, the portal contains over 500 datasets that are related to school report cards or lead testing in school drinking water (Fig. 4). These datasets have good keywords and are very findable. Removing those keywords reveals that publishers often include the year as a keyword. For data.wa.gov, keywords should be the main concepts or subjects along with alternative spellings or words, acronyms, and larger concepts.

Figure 4. Keywords from all datasets available at the time of this report (left) and keywords from all datasets available at the time of this report but with the most common seven words removed. Wordcloud made with https://www.wordclouds.com/

Other Curation Needs

Even when a dataset has high quality metadata, it may require other curation work to be a useful dataset. Other curation needs include:

1. Individual datasets that contain the same data but from different dates
2. Datasets that are test or dummy datasets
3. Datasets that do not meet the OCIO's definition of data

About 13% of the sampled datasets appear to be part of a larger time series and are candidates for aggregation. The publisher may have reasons for splitting the time series, but this should be noted in the description for each dataset.

About eight percent of datasets (38 of 470 datasets that were public on June 24, 2019) are test or dummy datasets that should be unpublished. Often this information is included in the title or description. It is difficult to easily subset these datasets because data.wa.gov also includes a large number of assets on lead tests in water which appear in search results using the word "test."

Three percent of sampled datasets (3 of 112 datasets) do not meet the OCIO's definition of data. Two of these datasets are the contact information of organizations that commented on a public document along with links to a pdf of their comments. The third is one column containing a list of brake manufacturers reporting to the agency.

Recommendations

Washington State's open data portal is akin to an unmediated information sharing resource. Publishers can publish and remove anything they choose – there is no higher authority gatekeeping what is available. Curating a collection like this is different than curating a traditional, centrally managed library collection. A portal curator has no control over what becomes available. This is similar to what the steward of a community book exchange (e.g. a Little Free Library®) or the owner of a community bulletin board faces. The largest curation activities in these situations are removing stagnant or poor-quality items and making the exchange interface findable and navigable (e.g. Churchill et al. 2003, Kullenburg et al. 2018). Both of these are applicable to data.wa.gov.

However, data.wa.gov also has extra layers of complexity that make curation after publication difficult. Unlike a community bulletin board, the portal is a source of information that may be highly connected and have many dependencies. For example, there are connections to visualizations and applications that make it risky to remove datasets, even with publisher permission. These complexities make curation much more difficult.

The following recommendations account for the portal's connectivity and focus on helping publishers provide higher quality metadata and making the portal itself more visible and usable.

Metadata and Data Quality:

- Establish a policy and procedure for removing datasets (Appendix B)
- Remove test and dummy datasets
- Add a controlled vocabulary for 'Data Provided By' and 'Posting Frequency'
- Remove the 'Period of time' and 'Originator' elements
- Determine if adding custom tool tips/help call-outs to the metadata entry form is possible
- Focus efforts on datasets missing one or two core metadata elements
- Assist agencies aggregate datasets that are part of a time series (e.g. Summer Low Flow Trend Indicator datasets and OSPI report card datasests)
- Run the metadata analysis several times a year to update the portallog.csv file

Visibility and Use:

- Create a direct user feedback option on the homepage for data.wa.gov. This should highlight that the query is going to a librarian
- Identify datasets or groups of datasets that would make good stories
- Advertise the portal to potential users outside the government- nonprofits, teachers, political groups

Acknowledgments

As the WSL representative and the main sponsor of this project, Kathleen Sullivan provided me with invaluable guidance, advice, mentorship, and encouragement. From the OCIO, Will Saunders provided essential technical and operational information and also patiently explained all the behind-the-scenes complexities of data.wa.gov. The WSL staff, especially Evelyn Lindberg, Cindy Aden, and Judy Pitchford,

generously gave their time to teach me how the WSL succeeds in its incredibly diverse work. Thanks to the ODL team: Nic Weber, Bree Norlander, Carole Palmer, and Kaitlin Throgmorton for helping at every stage of this project. And thank you to all the agencies and the one data user who all made time for interviews and provided insightful answers to my questions.

References

Churchill, E. F., Nelson, L., & Denoue, L. (2003). Multimedia Fliers: Information Sharing With Digital Community Bulletin Boards. In M. Huysman, E. Wenger, & V. Wulf (Eds.), *Communities and Technologies* (pp. 97–117). Springer Netherlands.

Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018). Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly*, *35*(1), 13–29. https://doi.org/10.1016/j.giq.2017.11.003

Kučera, J., Chlapek, D., & Nečaský, M. (2013). Open Government Data Catalogs: Current Approaches and Quality Perspective. In A. Kő, C. Leitner, H. Leitold, & A. Prosser (Eds.), *Technology-Enabled Innovation for Democracy, Government and Governance* (pp. 152–166). Springer Berlin Heidelberg.

Kullenberg, C., Rohden, F., Björkvall, A., Brounéus, F., Avellan-Hultman, A., Järlehed, J., … Westberg, G. (2018). What are analog bulletin boards used for today? Analysing media uses, intermediality and technology affordances in Swedish bulletin board messages using a citizen science approach. *PLoS ONE*, *13*(8). https://doi.org/10.1371/journal.pone.0202077

Máchová, R., & Lnénicka, M. (2017). Evaluating the Quality of Open Data Portals on the National Level. *Journal of Theoretical and Applied Electronic Commerce Research*, *12*(1), 21–41. https://doi.org/10.4067/S0718-18762017000100003

Umbrich, J., Neumaier, S., & Polleres, A. (2015). Quality Assessment and Evolution of Open Data Portals. *2015 3rd International Conference on Future Internet of Things and Cloud*, 404–411. https://doi.org/10.1109/FiCloud.2015.82

Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: An open data perspective. *Government Information Quarterly*, *31*(2), 278–290. https://doi.org/10.1016/j.giq.2013.10.011

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, *33*(2), 325–337. https://doi.org/10.1016/j.giq.2016.02.001

White House Memorandum. 2009. Transparency and Open Government. Accessed August 29, 2019 at https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government

Appendix A: Interview Questions

Code: Date: Time:

Describe ODL, Interested in assessing the state data portal, define metadata etc. Looking for background information.
Ask for consent, ask to record, ask for consent again.

Why did your agency start publishing data to data.wa.gov?

How do you choose what data to publish?

One dataset you have published is ( ). What is your experience with the open data portal?

Has your experience matched expectations when publishing data on data.wa.gov?

Do you know who uses your data? What format do you expect them to require data?

How does your agency obtain information from other agencies?

Does your agency use data.wa.gov to find data?

What are your data publishing plans for the next few years? Status quo, more, less?

Is there anything else you think I should know?

Appendix B: Possible Removal Policy and Procedure

(New York City has an excellent dataset removal policy for reference.)

Possible Removal Policy

Ideally, data should meet the OCIO's definition of data:

> *Data means final versions of statistical or factual information that:*
>
> - *Are in alphanumeric form reflected in a list, table, graph, chart, or other nonnarrative form, that can be digitally transmitted or processed;*
> - *Are regularly created or maintained by or on behalf of an agency and controlled by such agency; and*
> - *Record a measurement, transaction, or determination related to the mission of the agency.*

Data that should be considered for removal:

- Assets that are not data
- Assets that lack meaningful metadata rendering the data useless
- Assets that are part of a regularly updated series spread across multiple datasets should be aggregated
- Old assets that should be deleted or updated:
  - Data that cover a finite time period
  - Are part of a regular data collection regime (new data is available every year)
  - Have not been updated in two or more years.
  - Datasets with posting frequencies and updated dates that do not match
- Assets that are test or dummy datasets and are more than one year old.

Possible Removal Procedure

1. Create a public dataset that holds metadata for removed datasets, hereafter referred to as the Removed Asset Dataset. This dataset should have four additional columns: Date Removed, Reason for Removal, Future Action, Date of Future Action.
2. Contact the data owner and ask if the data is in use, and if not request permission to unpublish the dataset.
   a. If data is in use, suggest steps to make metadata clear or data more usable.
   b. Data is not in use, request permission to remove and go to step 3.
3. Copy metadata from asset list for dataset and all related assets and add it to the Removed Asset Dataset. Fill in date removed and reason for removal.
4. Unpublish the dataset and any associated assets.
5. Categorize dataset as either archive or deletion in the Future Action column. Records retention rules for WA may determine when data can be fully removed. If known, that date should be added to the Date of Future Action column.