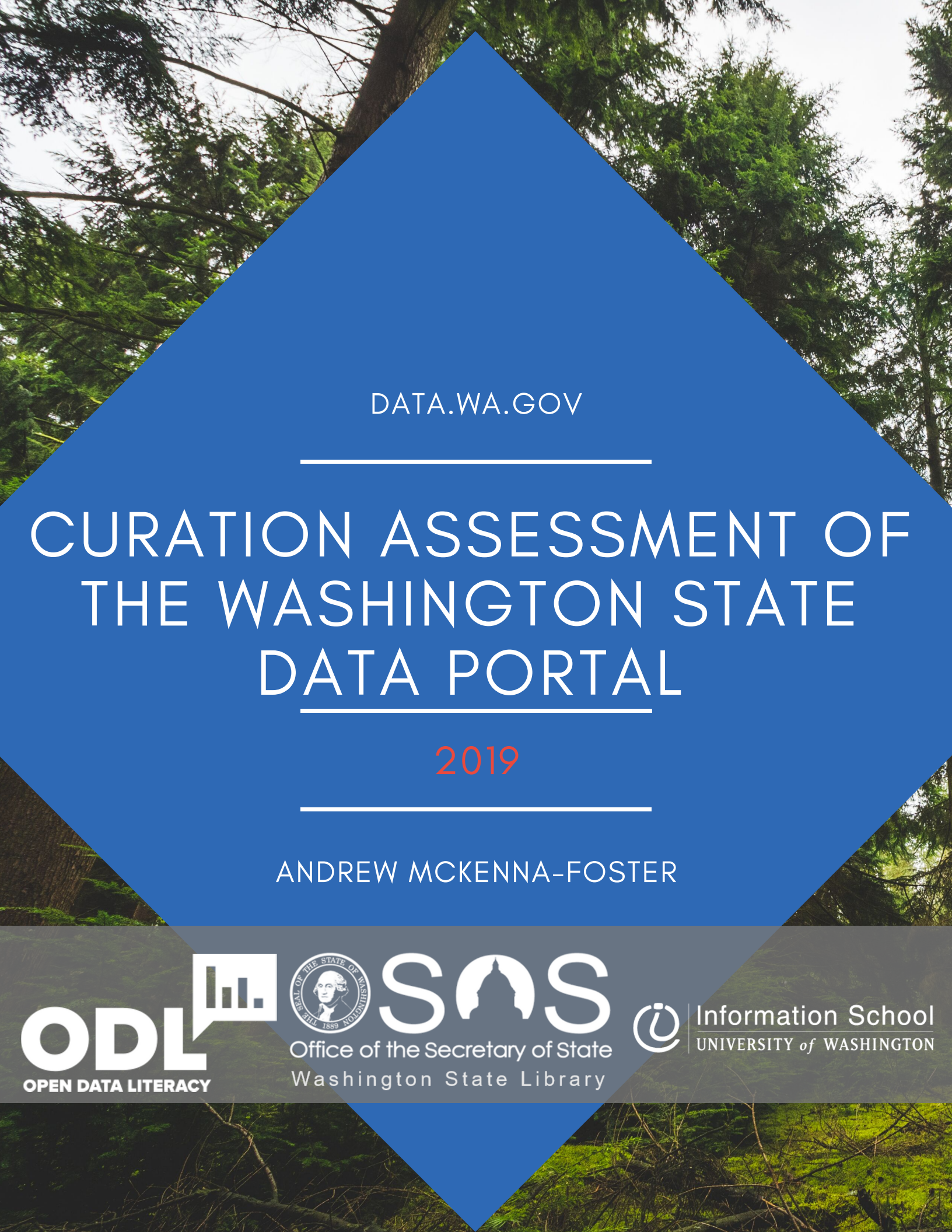


DATA.WA.GOV

CURATION ASSESSMENT OF THE WASHINGTON STATE DATA PORTAL

2019

ANDREW MCKENNA-FOSTER

ODL
OPEN DATA LITERACY

Office of the Secretary of State
Washington State Library

Information School
UNIVERSITY of WASHINGTON

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.

AGENCY INTERVIEWS

Agency data publishing behavior is only generalizable to the extent that it is unique to every agency

Agencies are overall very positive toward data.wa.gov
It is an integral part of their operations

8

State agencies Interviewed

1

Portal user interviewed

Publishing Behaviors

User focused - Know their users
Upload and forget - Publish but do not update
Internal use - Publish for inter/intra agency use
Transparency - Main goal is transparency

Users

Other state agencies
Local governments
Media
3rd parties- nonprofits, companies, etc.

A majority of publishers will continue or increase their use of data.wa.gov

Interviews with agency representatives show that data.wa.gov serves many purposes and is an important resource for Washington. 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. 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. Agencies do not like the limited visualization options, difficulty in searching, or lack of a data suppression feature. However, they express overall satisfaction citing the easy to learn web interface, the ease of filling in metadata, useful assistance from Socrata representatives, and useful API features. Publishers reported that their portal use will continue or increasein the future.

METADATA ASSESSMENT

Started around 2009, data.wa.gov is one of the oldest and most established state data portals

Any state agency can receive permission to publish data. To encourage publishing, there is no firm enforcement of metadata quality.

Metadata Existence

Incomplete metadata is a global problem
Complete metadata increases potential data re-usability

Metadata Understandability

Metadata values should be understandable
Data dictionaries should be included

Data Collection

Metadata is available through APIs
Python scripts available at:
https://github.com/OpenDataLiteracy/WSL-AMF

2,663

Datasets, charts, files, and maps

887

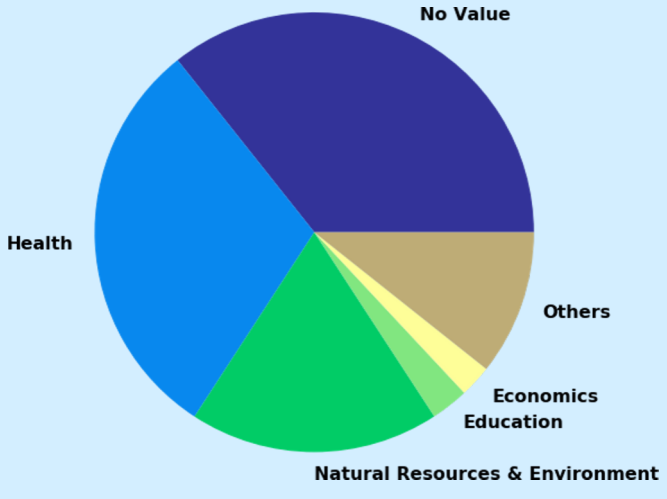
Tabular datasets

>30

Agencies publish data

Data.wa.gov stores datasets covering 14 categories. Within these categories, many datasets concern permitting, licensing, environmental testing, or public schools.

Most datasets are accessed through an API. The top downloaded dataset currently has over 950,000 views. The overall number of downloads ranges from zero to over 760,000.



AVAILABLE METADATA ELEMENTS - EXISTENCE

Data.wa.gov provides 14 metadata elements for publishers to populate. Only one of these, Name (or dataset title), is currently a required field. One element is an alternate email field and another is a notes field, and these were left out of the analysis. This leaves 11 optional metadata elements to assess for existence.
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).

Of 473 sampled datasets:

Most commonly completed elements:

- Category
- Tags
- Desctription
- Attribution
- Attribution Link
- License

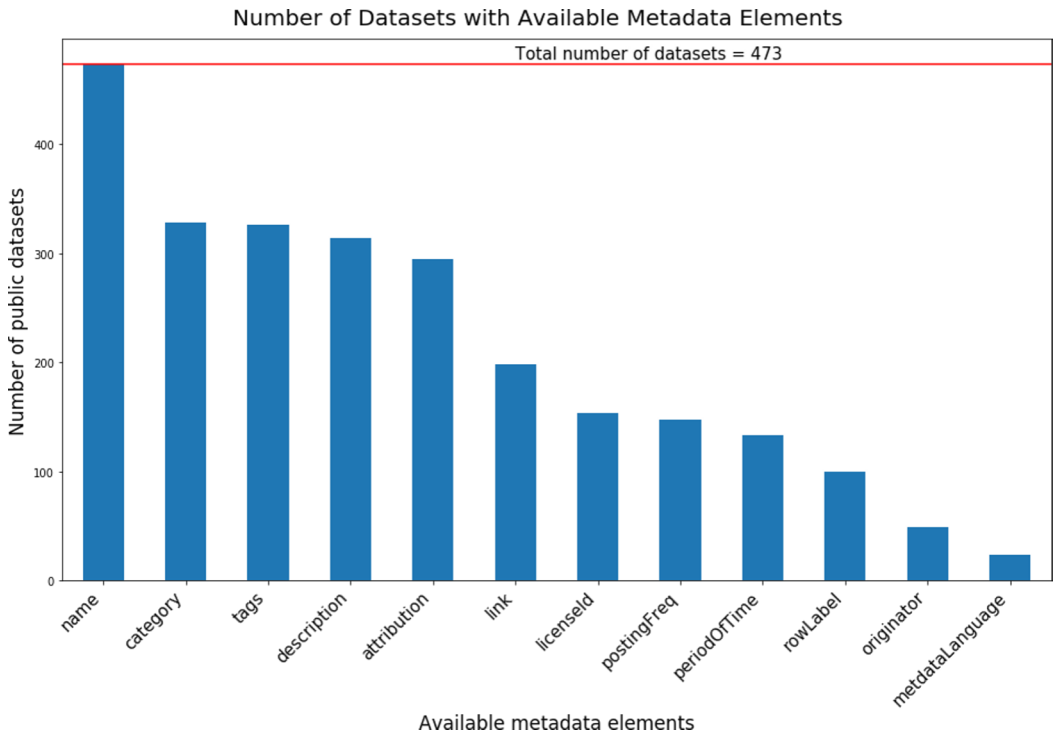
25%

Complete more than six metadata elements

19%

Only have a title

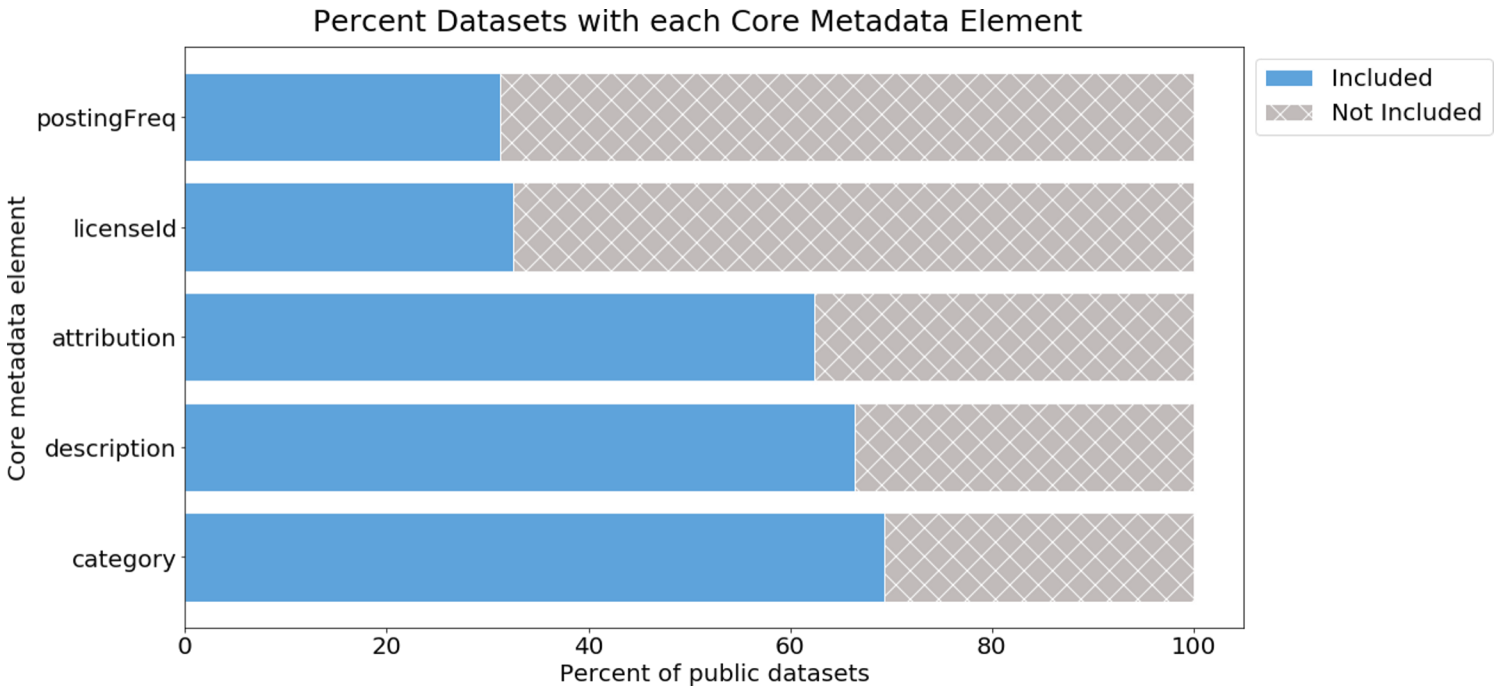
License and Posting Frequency (how often the dataset is updated) are two of the least filled out metadata fields but are both extremely important for users. The license lets a user know how they can use the data and the posting frequency provides an indication of how well maintained the dataset is.



CORE ELEMENTS

Five metadata elements are particularly helpful for finding, understanding, and using a dataset: Description, Category, Update Frequency, Data provided by, and License. These are core metadata elements.

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.



21%

Are missing all core elements

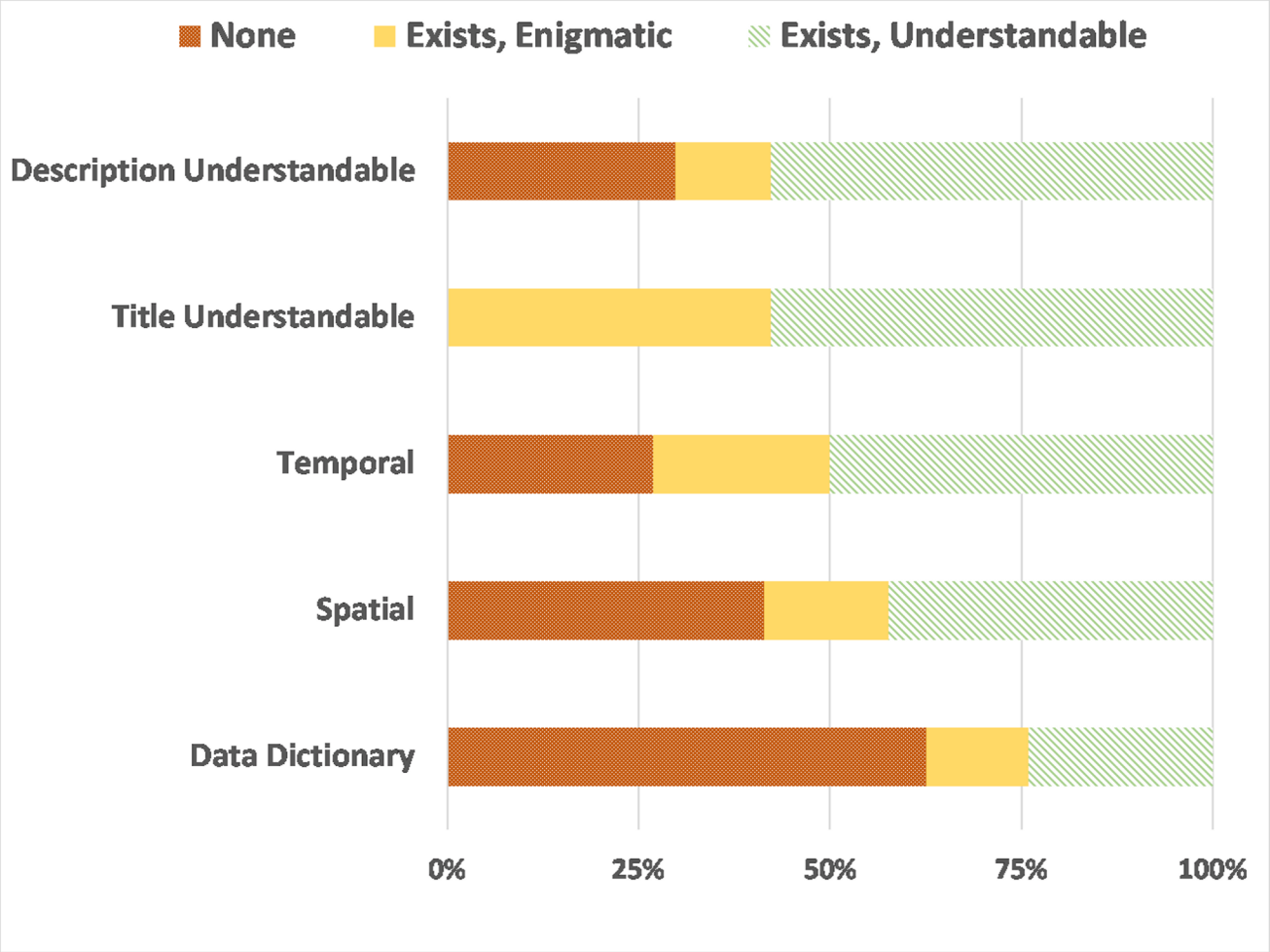
67%

Are missing license information

METADATA UNDERSTANDABILITY

Even when a metadata element is filled out for a resource, the content may not always be understandable. Examining a sample of 112 datasets, half have enigmatic information for at least one of the studied areas. This could affect the trustworthiness and ultimately the reusability of data.

Understandable titles help users searching for data efficiently find what they need but 42% of titles are difficult to interpret. Temporal information may be in multiple metadata elements (Description, Posting Frequency, etc) but in almost 25% of datasets these confilct or are enigmatic.



CORE METADATA 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.

7%

Excellent

All core metatadata undertandable

28%

Moderate

One core element needs improvement

RECOMMENDATIONS

Data.wa.gov is different than a centrally managed library collection

A curator does not have control over what is published

- Curation will involve working closely with agencies
- Likely increase in publishing volume makes curation even more important

Modify Metadata Entry

- Add controlled vocabularies to appropriate core metadata elements
- Remove redundant elements
- Add explanatory text to encourage publishers to fill out core elements.

Curate Existing Datasets

- Focus efforts on datasets missing one or two core elements
- Remove test/dummy datasets
- Work with agencies to combine datasets that are in a series.

Portal Visibility and Use

- Create a user feedback system that highlights the role of the librarian as curator
- Create stories using existing datasets to highlight portal use
- Advertise portal resource to potential users

The Washington State Library began in 1853 and has supported the State/Territory Legislature's information needs and advocated for library development throughout the state since then. It is growing as an open data leader and educator within the state.

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.

Resources

Full report, code, and notes available:

https://github.com/OpenDataLiteracy/WSL-AMF

ODL
OPEN DATA LITERACY

The Open Data Literacy Program is an Institute of Museum and Library Services grant funded program and the University of Washington Information School. The program partners with public institutions to design curricula and conduct research focused on open data.