The Complicated Problem of Dataset Removal

Open government data portals can be a terrific resource for diverse public sector information in a structured format —water quality test results, licensing data for contractors and healthcare providers, lobbyist spending, et cetera. However, these data sets often vary widely in quality and utility. This is especially true where agencies publish directly to the portal without a central entity curating each dataset (i.e. unmediated deposits), leading to lots of data that may or may not be what people are looking for. Although a decentralized approach to running an open data portal likely lowers costs and other barriers for data publishing, it also means that the steward of the open data portal needs to periodically remove datasets in order to maintain a more accessible collection for the public and other users. So, if some datasets need to be removed, how does a curator document the removal process in a transparent way??

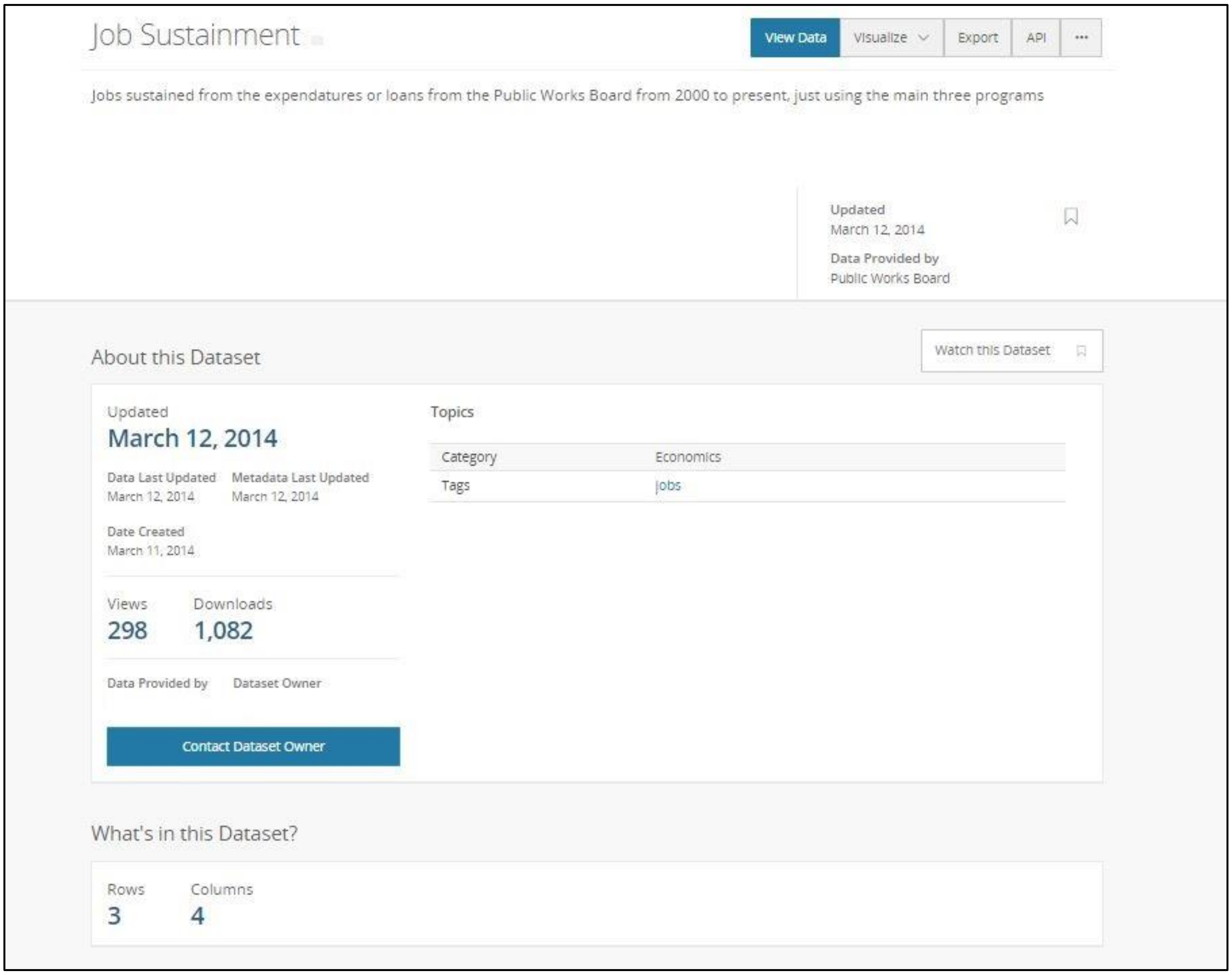
As an intern for the Open Data Literacy (ODL) program with the University of Washington iSchool [1], I have been working with the Washington State Library to assess the state's open data portal and offer recommendations for curating this important resource. After seven years of growth, the Washington State open data portal is now in need of curation and the State Library with its history of curating government documents and state related information, is the perfect institution for the task.

After years of successful application in hundreds of city, state, and international portals, the open government data movement now faces the problem of nearly drowning in oceans of poor-quality data and metadata. The open data goals of setting appropriately ambitious timelines and meeting an 'open by default' paradigm could be the reasons behind this.

A clear and broadly applicable policy on the criteria for removing a dataset to mitigate this data flood, as well as a system to track removals, has yet to emerge. This is not surprising considering one of the 10 Principles of Open Data is that data should be published with permanence in mind. It should be considered though: weeding out low-quality datasets may not only strengthens the reputation of quality for a data repository, but it may also make it easier for users of any type to find data that is current and relevant to their needs. However, in removing datasets, curators run the risk of cutting off sources of information that users may have grown accustomed to using in the past.

There is also a concern that removing data from an open government portal could be seen as a step away from transparency, no matter how rational the decision to remove it.

Here is a use case: A very small dataset of jobs and expenditure totals uploaded five years ago with low quality metadata might seem like an obvious candidate for removal. However, that old dataset populates some figures in a rarely used but important online government document. Removing the dataset would break links in that document and could also be seen as an effort to obscure information. A curator planning to remove this dataset needs a transparent and reversible process.



Job Sustainment

Jobs sustained from the expendatures or loans from the Public Works Board from 2000 to present, just using the main three programs

Updated
March 12, 2014
Data Provided by
Public Works Board

About this Dataset

Updated
March 12, 2014

Data Last Updated
March 12, 2014

Metadata Last Updated
March 12, 2014

Date Created
March 11, 2014

Views
298

Downloads
1,082

Data Provided by
Dataset Owner

Topics

Category	Economics
Tags	jobs

What's in this Dataset?

Rows
3

Columns
4

Removal of this dataset would improve overall data quality for the portal, but may have unintended consequences.

There has been very little focus on this topic in research or literature. New York City may have one of the clearest, and only publicly viewable, policies available. Their process works like this:

1. Data is identified for removal either by the dataset owner or portal curators. Candidates include data that do not meet the definition of data or are not updated and not analytically useful.
2. After confirmatory discussion with the dataset owner, the dataset is unpublished for three months and then deleted from the portal.
3. A small portion of that dataset's metadata and the reason for removal is added to a dataset available on the portal.
4. The dataset may be archived or completely deleted.

This is a thorough solution, but it requires quite a bit of human labor and only works when a portal is curated by a central department. Additionally, depending on the deletion process, deleted datasets may never be recoverable if they are ever needed again.

An ideal solution for a more decentralized portal would be an automated system that updates the metadata with a reason for deletion and compresses and archives each dataset upon deletion by the dataset owner. The results page for any search on the portal could prompt users to rerun their search and include records of deleted datasets if their original search did not provide expected results.

Socrata and CKAN, two popular open government data portal platforms, both allow the deletion and recovery of datasets but neither provides a user-friendly tracking process. Until an automated, user-focused solution appears, what can a portal steward do to maintain transparency and access while effectively removing or archiving datasets?

If storage is not a limiting factor, a portal curator could do the following:

1. Request feedback from the dataset owner about upcoming changes to their dataset
2. Replace the original dataset with an identical, but compressed version, or upload compressed format as a new version
3. Note in the description why the data is being archived and where better data is located
4. Recategorize the dataset to 'Archived' or something similar in the 'Category' metadata element.

This solution would indicate to a user the status of the dataset both by the category of 'Archived' and the compressed data format, while also indicating why the dataset was archived and possibly where better data is located.

As always, a system would need to work within the constraints of budgetary restrictions and portal curator time. Developing something like this system will likely be one of my recommendations for how the State Library can help curate the Washington State open data portal, hopefully to create a system that remains both functional and efficient for data administrators and users alike.

[1] The ODL program is funded by the Institute of Museum and Library Services and works towards creative and realistic ways to help government agencies and libraries create or work with open data.