Open Data Publishing by Public Libraries

Nicholas Weber
nmweber@uw.edu
University of Washington
Seattle, WA

Bree Norlander
norlab@uw.edu
University of Washington
Technology & Social Change Group
Seattle, WA

ABSTRACT

Public libraries in the USA are part of a broad civic information ecosystem that is rapidly adopting transparency legislation aimed at publishing structured open data for public reuse. In this preliminary results paper we look specifically at the open data publishing practices of 85 public libraries in the USA. We find that less than half of these libraries have published any open data, and that there is no relationship between revenue nor staff size and open data publishing practices. Categorizing public library open data by type we find overwhelmingly the most frequent type of open data published by libraries are geospatial (map) information. We use these findings to develop a proposal for public libraries to engage in publishing a core set of open data, and conclude by discussing the potential for reuse of open public library data.

CCS CONCEPTS

• **Information systems → Digital libraries and archives**.

KEYWORDS

conceptual models, digital libraries

ACM Reference Format:
Nicholas Weber and Bree Norlander. 2019. Open Data Publishing by Public Libraries. In *JCDL '19: Joint Conference on Digital Libraries, June 2-6, 2019, Champaign-Urbana, Illinois.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Over the last decade open government initiatives have spurred the release of valuable public sector information at the federal, state, and municipal levels [4]. Many of these initiatives include the development of digital library infrastructures, such as repositories or information portals, that enable the publication of and public access to structured data [11]. An initial motivation for adopting open government and open data initiatives is to increase the transparency of government operations, and in turn provide greater political accountability [7]. A secondary motivation of open government data programming is the potential for public services to make gains in efficiency through academic and private sector reuse - such as the developers who reuse open data in mobile or web applications that connect citizens to relevant public services offerings [9, 20]. Public libraries are beginning to explore service offerings around the use and publication of open data to meet the needs of an increasingly engaged polity [1]. As most public libraries in the USA are financed through state, county, and city taxes [15], these institutions are also beholden to the same transparency and accountability ordinances around 'open government' as other public service agencies (e.g. Police Departments, Health and Human Services, etc).

In this preliminary results paper we investigate the practices of public libraries as not just services for, but publishers of open government data. We focus in particular on public libraries in 85 USA cities that have either passed open government laws, or established open data infrastructures to publish structured public sector information. Our goals in this research project are to establish a baseline for documenting trends in public libraries publishing open government data. As government funded and provisioned entities we argue that public libraries have rich and valuable public sector information that can contribute to policy making decisions, and can, similar to other public institutions, spur innovation from end-user applications that encourage engagement with unique library resources and service offerings. We hypothesize, however, that as a result of well-documented shortages in technological expertise and data practices more generally [6], public libraries will be immature in their open data publishing efforts. This study will therefore document the publishing practices of USA public libraries at an early stage of broader open government compliance.

A second goal of this work is to develop recommendations for what might be a core set of data that all public libraries in the USA can and should publish as open government initiatives mature. We recognize that a number of reporting mechanisms for aggregating data about public libraries already exist. For example, the Institute for Museum and Library Services (IMLS) conducts an annual survey of public libraries to capture general statistics such as library funding levels, circulation trends, program offerings, etc [8]. However, self-reported data from the IMLS annual survey only provides summary statistics, and this type of aggregate data are necessarily limited for meaningful reuse. By describing how USA public libraries can publish full and complete records of their activities to local open data repositories, we argue for the potential to federate public library data into reports that provide more granular information, and in turn compliment existing data sources such as the IMLS's annual statistical reports.

In the following sections we describe the methods used to gather information about the open data publishing practices of 85 public libraries in the USA. We then present preliminary findings from this analysis and answer two broad questions: 1. Are public libraries in the USA publishing open data to city portals?; and if so, 2. What types of open data are public libraries publishing? We conclude with

JCDL, 2019, June 2-6, 2019, Champaign-Urbana, Illinois

a discussion of the potential for increased open data publishing, and the value that these activities can bring to public library policy development and program planning.

2 METHODS

To gather a sample of public libraries publishing open data we first randomly selected 85 USA cities from a curated index of 2600 open government data portals from around the world [13]. For each of the 85 cities listed in the index we recorded the initial url of the data portal, and then tried to verify whether or not the indexed portal link was still valid. If the portal was not still in operation we then attempted to find the city's current open data publishing site. Even though the indexed list of data portals was developed in 2015, we found that 11 of the cities original data portal urls were either incorrect (pointing to an incorrect location), wouldn't resolve, or were forwarded to a different url.

For each city's data portal that was still in operation or that we were able to find via a Google search, we gathered descriptive information about the repository software used to publish the data, the total number of datasets accessible in each portal (we limited Socrata-hosted sites to "type = Datasets"), and searched specifically for data published by or about public libraries. Within each portal we also searched specifically for library-related datasets using queries such as "Library" or "Libr*", and also made use of faceted browsing features such as "categories" or "Data Owner" to locate relevant public library data.

For library related datasets to be considered relevant to our analysis we examined metadata, contents (e.g. variable names), and dataset descriptions. We excluded datasets that contained the keyword "library" but were, in reality, about broader topics such as Boston's "CityScore" dataset which provides "metrics on overall city health based on work done across all facets of the City of Boston" [16]. Other false positives were removed based on the identified publisher. For example, the City of Seattle has a number of datasets that are published to the portal by users who have performed a specific analysis of that public library's data, but were not officially published by the Seattle Public Library.

In addition to searching data portals for relevant library data we also looked to broader statistical indexes and data census information to understand the open data publishing practices of the cities in our sample. From the Open Knowledge Foundation's (OKFN) 2016 USA City Open Data Census - a project that inventoried the types and quality of open data published in over 150 USA cities - we recorded each city's overall openness score [1]. An overall openness score is calculated by combining a 100-point evaluation for 19 different types of open data. For example, a city will receive a 100-point score for "open crime data" if the city publishes this information in a machine-readable format, is free to download, etc. Combining the 19 different scores by type , a city can at best score a 1900 in the OKFN census. We hypothesize that cities with high overall openness scores in the census will be more likely to publish public library data, as it speaks to the city's commitment to making public sector information accessible and the maturity of supporting data infrastructures.

We also gathered statistical information about each of the libraries in our sample from the 2016 IMLS Public Library survey [2]. Several cities in our initial dataset had multiple entries in the IMLS survey (e.g. Austin which has four different library systems within the city limits: Austin Public, Lake Travis Community, Westbank Community, and Wells Branch Community). For this preliminary study, rather than investigate each occurrence and disaggregate the statistical information, we simply removed cities with multiple entries from our sample. We also removed cities with insufficient data or broken links to data (Reno, Lubbock, Jacksonville), cities with custom portals that made it difficult to search for and/or total datasets, and Minneapolis which has no corresponding library in the IMLS data due to a county-wide library system. This resulted in a final dataset that included 62 cities, including information we gathered from our own inventory as well as supplementary IMLS survey statistics.

3 FINDINGS & DISCUSSION

Below we present findings from our analysis, which are addressed at answering two broad questions: 1. Are public libraries in the USA publishing open data to city portals?; and if so, 2. What types of open data are public libraries publishing?

3.1 Libraries Publishing Open Data

All 62 cities that remained in our sample were classified by the IMLS as City (Large, Midsize, or Small) or Suburban (Large) library systems. This designation in the IMLS Public Library survey is based on the geocoded latitude and longitude values of a library's main branch street address. The size subcategory (Large, Midsize, or Small) is based on the National Center for Education Statistics (NCES) locale coding system [8]. That all library systems in our sample were either categorized as a City or Suburb system should be interpreted as a result of the adoption of open government initiatives and open data infrastructure development. Previous research has shown that throughout Europe [5] and the USA [3] large municipalities, including cities and well funded suburbs [14], are the earliest adopters of the transparency initiatives that result in open data publishing. Our sample, gathered from an indexed list of existing open data portals necessarily reflects this systemic bias in open data publishing practices.

Of the 62 cities that remained in our sample less than half (n = 27) have published any library related data. In figure 1 we show the frequency distribution of all cities in our sample that published more than 1 library related dataset (n=10). The library that published the most datasets (n= 55) was Chicago, a large metropolitan city that represents one of the longest running and most robust open data programs in the USA [10]. That less than half of the cities in our sample have published library related open data, and only ten had published more than one dataset, confirms our hypothesis that public libraries conformance with open government initiatives is immature.

After confirming the frequency of public library open data publishing we attempted to better understand how these practices related to the overall open data publishing of a particular city. For

[1] http://us-city.census.okfn.org/year/2016

[2] https://data.imls.gov/Public-Libraries-Survey/Public-Libraries-Survey-2016-Library-Systems-Admin/grpq-tgei

example, we would expect that Chicago, with a long-standing open data program, might simply be publishing more library related data as a spillover over effect of the existing human and technical infrastructure that is available to support a library's publishing efforts. We also suspected that cities with high overall 2016 USA City Open Data Census Scores would be more likely to publish public library data. From the IMLS annual survey data there are a number of potentially relevant variables that can shed light on what influences the practices of public libraries as open data publishers - namely, the total revenue of a city's public library system (as a proxy to available funding), and the number of full time staff (as a proxy to labor size). One would expect that with a larger annual revenue or a larger staff a public library could invest more time and effort into complying with transparency initiatives, and overall managing data for public release.

To investigate these hypotheses, we ran a simple linear regression with the "count of library datasets" as the dependent variable against the the independent variables (each in turn), "total annual revenue," "total staff," and "Open Data Census Score." We found no significant relationship between library datasets and revenue (though "total overall datasets available in a city" does have a statistically significantly positive correlation to library revenue with a p-value of $2.35e^{-12}$). Regression analysis of "count of library datasets" and "total staff" produced a nearly statistically significant correlation with a p-value of 0.08. While this value does not demonstrate a strong enough correlation to reject the null hypothesis (that there is no relationship between the variables "staff size" and "total library datasets published") we do believe that the IMLS statistical data may be relevant for future exploration. Finally, simple linear regression showed no statistically significant relationship between library datasets and their corresponding open data census scores. While the largest provider of public data, Chicago, does appear as our 5th highest census score city, the second largest provider, Durham, does not even make the top ten in census scores.

In summary, given the small sample size of libraries in this analysis we find no significant correlation between either staff size, nor the total annual revenue of a public library as indicating a propensity to engage open data publishing. Staff and revenue are likely confounding variables (in fact they are highly correlated with a Pearson correlation coefficient of 0.95). However, that they have no correlation with publishing practices further emphasizes the fact that public libraries are part of, and embedded in, a broader civic information ecosystem that often evolves piecemeal through slow and steady maturation [2].

3.2 Types of Public Library Open Data

To better understand and describe the types of open data that were being published by public libraries we further analyzed the contents and formats of each dataset in our sample (n=27 libraries that had actually published 1 or more open datasets). We began by listing the title of all datasets, and where possible the portal (repository) categories that these data were published under. From these general open data categories, we created a categorization to describe the type of library data that were being published. This categorizations included the following 10 types of structured public library data: Geospatial, Facilities, Catalog & Circulation, Patrons,

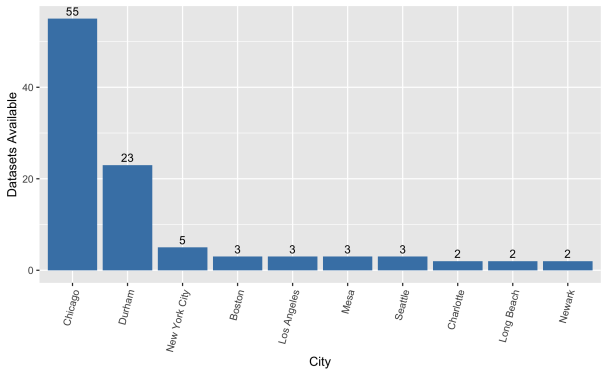

Figure 1: Frequency distribution of all cities the sample that published more than 1 library-related dataset.

Public Records, Events & Calendar, Human Resources, Financial, Utilities, Technology Offerings (for a complete definition of each category please see the full dataset).

Applying this categorization to the 27 public libraries that published open data we find that the most frequent type of data published are Geospatial (n= 27) followed by Catalog & Circulation (n=7) and information about Patrons (n=7) (see Figure 2). In almost every instance of Geospatial data published by a library there was an accompanying shapefile. It is not surprising that geospatial data appears to be the most frequent category of data published by libraries in our sample; In fact, geospatial information is one of the most popular, by format, of all open government data published in the USA [19].

Catalog & Circulation data are published second most frequently in our sample, but it is still somewhat surprising that this type of data is not more frequently a part of the publishing practice of a library engaged in open data programs. Circulation data in particular are already well structured and often described as the most valuable asset for data-driven collection development and evaluation [12]. We hypothesize that libraries not publishing this data is due to a lack of skills in anonymizing what might be patron records containing personally identifiable information (PII) - which is a recognized hurdle and future challenge for releasing and making use of circulation data [17]. .

Overall, facilities information, patron data (e.g. de-identified circulation records by branch), and technology offerings (e.g. number of laptops available for lending, computer terminals, etc.) were rarely published by the libraries in our sample. We expect that this type of data is readily available to library staff in structured (or structurable) formats, and therefore represent candidate data types for public libraries to easily and efficiently publish as their first open data contribution to a city portal. We further discuss this potential in the following section.

4 CONCLUSION & FUTURE WORK

Through an empirical study of the open data publishing practices of 85 cities in the USA we have demonstrated that: 1. Less than half of the libraries in our sample are actively publishing open data (n=27), and a much smaller number are publishing more than one

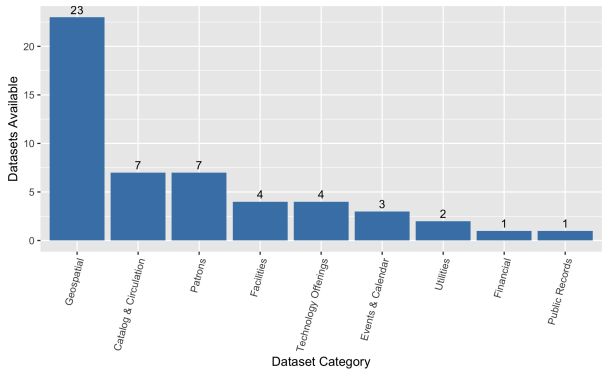

Figure 2: Total library datasets in each category.

dataset (n=10); 2. We found no relationship between the size of a library's annual revenue and their propensity to publish open data, but did find a weak correlation between staff size and propensity to publish open data; 3. By examining specific types of published open library data we find that most datasets contain geospatial information. These findings demonstrate that open data publishing by public libraries is relatively immature compared to other city entities (e.g. Fire departments, health and human services, etc). However, we believe that public libraries are also well positioned to make their data more accessible due to the fact that circulation, patron, technology offerings, and event data are already broadly collected and likely well structured.

In future work we believe there is a need to better understand not just the open data publishing practices, but also the data collection efforts of local public libraries. Future steps might include working with libraries to inventory their current holdings, and understand what types of data might be most useful in alignment with local and national measurement systems for evaluation of public services [18]. We also see a great potential to use the type categories of library open data developed in this paper to suggest structured data that can easily be published by public libraries. Further, there is the potential to share best practices in publishing certain types of open public library data, and develop professional development and outreach materials to help public library staff understand challenges in releasing this type of public sector information.

Finally, there may be questions as to the utility of public libraries investing scarce resources in publishing open data. We return to the second motivation of open government initiatives described in the Introduction section - there are many unanticipated reuses of structured open data that can spur public and private sector innovation, and lead to gains in public engagement. As evidence, we point to the platform Kaggle where datasets are published, curated, and analyzed by a large social network of data scientists. On Kaggle one can currently find analysis and conversations around public library data made available by San Francisco, Seattle, Oakland, New York City, and a number of other cities [3]. This community is building models to understand patron behavior, and identify resource consumption patterns in public library systems. The online collective action that could facilitate insights for libraries and cities is indeed a potential value already being realized, and we believe this wisdom of the crowd insight will only improve with greater public library data accessibility.

[3] e.g. https://www.kaggle.com/bengin/mastering-the-sf-library-data

5 DATA AVAILABILITY

The data described in this paper, further documentation about the methods used, and the images included are openly accessible at: https://github.com/OpenDataLiteracy/JCDL-2019- OpenDatainPublicLibraries

ACKNOWLEDGMENTS

This research was supported in part by IMLS grant RE-40-16-0015-16.

REFERENCES

[1] Laurie Allen, Claire Stewart, and Stephanie Wright. 2017. Strategic open data preservation: Roles and opportunities for broader engagement by librarians and the public. *College & Research Libraries News* 78, 9 (2017), 482.
[2] Judie Attard, Fabrizio Orlandi, Simon Scerri, and Sören Auer. 2015. A systematic review of open government data initiatives. *Government Information Quarterly* 32, 4 (2015), 399–418.
[3] Domonic A Bearfield and Ann OâĂŹM Bowman. 2017. Can you find it on the web? An assessment of municipal e-government transparency. *The American Review of Public Administration* 47, 2 (2017), 172–188.
[4] Soon Chun, Stuart Shulman, Rodrigo Sandoval, and Eduard Hovy. 2010. Government 2.0: Making connections between citizens, data and government. *Information Polity* 15, 1, 2 (2010), 1–9.
[5] Peter Conradie and Sunil Choenni. 2012. Exploring process barriers to release public sector information in local government. In *Proceedings of the 6th international conference on theory and practice of electronic governance*. ACM, 5–13.
[6] Kevin Curran, Michelle Murray, and Martin Christian. 2007. Taking the information to the public through Library 2.0. *Library Hi Tech* 25, 2 (2007), 288–297.
[7] Tim G Davies and Zainab Ashraf Bawa. 2012. The promises and perils of Open Government Data (OGD). *The Journal of Community Informatics* 8, 2 (2012), 1–8.
[8] Lisa Frehill, Kim Williams, Carol Wan, and Evan Nielsen. 2018. Public Libraries Survey Fiscal Year 2016. https://www.imls.gov/sites/default/files/fy2016_pls_data_file_documentation.pdf
[9] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. 2012. Benefits, adoption barriers and myths of open data and open government. *Information systems management* 29, 4 (2012), 258–268.
[10] Maxat Kassen. 2013. A promising phenomenon of open data: A case study of the Chicago open data project. *Government Information Quarterly* 30, 4 (2013), 508–513.
[11] Rob Kitchin. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.
[12] Jennifer E Knievel, Heather Wicht, and Lynn Silipigni Connaway. 2006. Use of circulation statistics and interlibrary loan data in collection management. *College & Research Libraries* 67, 1 (2006), 35–49.
[13] Remi Mercier. [n. d.]. We listed 2600+ Open Data portals around the world! See how! https://bit.ly/1O0mStt
[14] Ines Mergel, Alexander Kleibrink, and Jens Sörvik. 2018. Open data outcomes: US cities between product and process innovation. *Government Information Quarterly* 35, 4 (2018), 622–632.
[15] Redmond Kathleen Molz and Phyllis Dain. 1999. *Civic space/cyberspace: The American public library in the information age*. Mit Press.
[16] City of Boston. [n. d.]. CityScore. https://data.boston.gov/dataset/cityscore
[17] John Renaud, Scott Britton, Dingding Wang, and Mitsunori Ogihara. 2015. Mining library and university data to understand library use patterns. *The Electronic Library* 33, 3 (2015), 355–372.
[18] John L Smith, Joseph Matthews, Mike Crandall, Sandy Nyberg, and Timothy Cherubini. 2017. Landscape of Major US Public Library Data Collection Efforts: A Working Paper for the Measures that Matter Initiative. (2017).
[19] Anne Washington and David Morar. 2016. Open for Whom? An Overview of Data. Gov File Formats. (2016).
[20] An Yan and Nicholas Weber. 2018. Mining Open Government Data Used in Scientific Research. In *International Conference on Information*. Springer, 303–313.