Mining Open Government Data Used in Scientific Research

An Yan[1] and Nicholas Weber[1]

The Information School, University of Washington
yanan15@uw.edu, nmweber@uw.edu

Abstract. In the following paper, we describe results from mining citations, mentions, and links to open government data (OGD) in peer-reviewed literature. We inductively develop a method for categorizing how OGD are used by different research communities, and provide descriptive statistics about the publication years, publication outlets, and OGD sources. Our results demonstrate that, 1. The use of OGD in research is steadily increasing from 2009 to 2016; 2. Researchers use OGD from 96 different open government data portals, with data.gov.uk and data.gov being the most frequent sources; and, 3. Contrary to previous findings, we provide evidence suggesting that OGD from developing nations, notably India and Kenya, are being frequently used to fuel scientific discoveries. The findings of this paper contribute to ongoing research agendas aimed at tracking the impact of open government data initiatives, and provides an initial description of how open government data are valuable to diverse scientific research communities.

Keywords: Open Data, Literature Mining, Research Policy, E-Government

1 Introduction

The release of public sector information (PSI) has traditionally been motivated by democratic ideals related to representative governance, accountability, and transparency. Over the last decade, governments around the world have greatly expand the scope of PSI accessibility through their participation in e-democracy and e- government initiatives [1]. Most notably, there is a global movement towards publishing openly licensed, machine-readable data, known as Open Government Data (OGD), from transportation, budgeting, agricultural, and public health agencies [2].

While much of the public support for open government data remains focused on increasing government transparency, public officials have also begun to recognize the value of open data for the private sector [3]. Thus, much of the existing research into open government data initiatives has focused on two questions: First, how does open government data affect government accountability and transparency ? [1]; and, Second, in opening up government data to entrepreneurs, what economic impact does open government data create? [4]. In this paper, we seek to expand current efforts in tracking the impact and value of open data initiatives by exploring the use of OGD in scientific research. By mining citations, mentions, and links to open government data found in a broad collection of peer-reviewed literature, we present empirical evidence about which scientific research communities are using OGD, and

for what purposes. In the following section, we situate our work amongst previous studies of OGD users, and then present the research design for this study.

2   Related Work

Although OGD has sparked a growing level of interest among researchers, there is little empirical work that focuses explicitly on how these resources are used, in practice, by the public [5]. Notable exceptions include, Bright et al. [6] who analyzed downloads of data hosted by data.gov.uk. They found that a majority of OGD has never been accessed, and the most frequent use of OGD was for commercial purposes. Similarly, Young and Yan [7] examined the challenges and expectations of civic hackers using open government in developing new technologies. They found that this community expressed a desire for higher quality data, and that issues related to the functionality of an open data portal greatly limited OGD use.

There are few studies that have examined OGD and its potential use in scientific research. Safarov et al. [8] conducted a systematic literature review examining the utilization of open government data through peer-reviewed publications. These authors categorize open government data use based on the type, condition, effect, and users of open data. They argue that OGD could allow researchers to form new analysis based on OGD, but note that there is little evidence suggesting that scientific communities are interested in this data source. In one of the few targeted studies looking into how open government data is actually being used in scientific research, Martin et al. [9] examined the requirements for and potential uses of the New York State's open health data portal. Collecting data through surveys and focus groups, they argue that obstacles to OGD use in research include low awareness of data availability and limited engagement with government data producers.

This paper seeks to better understand the use of open government data for scientific research by answering the following research questions:

– How are OGD used in academic research?
– What sources of OGD do researchers use?
– What fields of academic research are using OGD?

In the following sections, we describe the methods used for identifying and analyzing citations, mentions, and links to open government data in peer reviewed literature published between January 2009 and July 2017. We summarize our findings through descriptive statistics, and preliminary observations about how open data are used by different scientific research communities.

3   Methods

We developed a systematic review protocol to identify, select, and categorize published literature where open government data was explicitly used as a research input. By following a systematic review protocol this study is replicable, in that our results can be reproduced by following the steps outlined below, and exhaustive, as

our search for relevant literature makes use of all potential sources currently indexed by three major publication databases [10]. In the following sections, we describe the systematic protocol used for identifying and selecting relevant literature, and the methods used for mining this literature for evidence of OGD use.

Search Terms: We first obtained a list of open government data portals from data.gov[1]. We chose to remove portals from this list that were not specific to one particular government or agency (e.g. The World Bank), and we also added well-respected international OGD sources that were absent from the initial list (e.g. Taiwan - data.gov.tw). In total, there were 302 unique URLs representing sources where researchers might obtain open government data. Table 1 provides a summary of the types of OGD portals included in this study.

Table 1. Summary of OGD portals used as search terms

Type	Count	Examples
International Regional	155	Victoria (http://www.data.vic.gov.au/)
International Country	54	Australia (http://data.gov.au/)
US City or County	46	Seattle (http://data.seattle.gov/)
US State	38	Hawaii (http://data.hawaii.gov/)
Other State Related	8	NY Department of Health(https://health.data.ny.gov/)
US	1	Data.gov (https://www.data.gov/)

Article Retrieval: We constructed a query combining the 302 data portal URLs described above, and used this query to search the Scopus, Springer, and IEEE databases in July, 2017. Queries were limited, by date, to January 2009 to July 2017 in order to align with the establishment of major open data initiatives, such as the launch of data.gov in 2009. In total, 2486 articles were identified.

Next, we removed duplicate publications that appeared in multiple databases. We then further refined the study sample by selecting only peer-reviewed full-text journal articles or conference papers (i.e. we did not include abstracts, chapters, or books). We only selected publications that were written in English, had full-text accessible to our research team, and in which OGD was actually used by the authors, and not simply mentioned as being a relevant topic or phenomena of interest. To determine which papers should remain in the sample using this selection criteria, the two authors of this paper individually examined 170 articles selected randomly from the initial sample of 2039 publications. Cohen's Kappa was calculated to determine selection agreement, and Landis and Koch's guideline was used to interpret the kappa statistics [11]. The authors achieved substantial agreement for both inclusion and exclusion of relevant papers. The first author then applied the selection criteria to the remaining papers in the initial sample. This process yielded a study sample of

---

[1] https://www.data.gov/open-gov/

1229 papers.
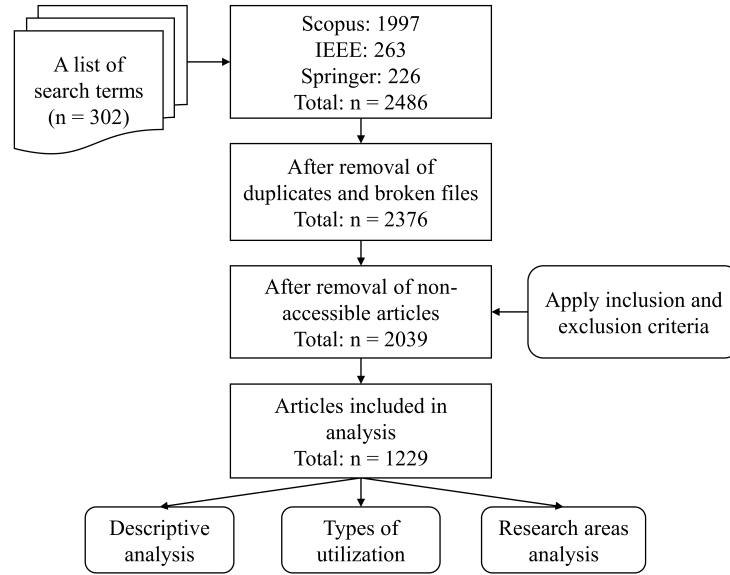
**Fig. 1.** The process of literature mining

Determination of Usage Types: Using the study sample (n=1229), the two authors read through the full text of a randomly selected subset of 63 papers. While reading the two authors inductively developed separate codebooks to categorize different ways that the publications described using OGD. The authors then met, combined their codebooks, and recoded an additional 200 articles. Seven usage types emerged from this inductive coding: OGD were used 1) As the main data source for new analysis, 2) As ancillary data source for new analysis, 3) For result evaluation, 4) To demonstrate the effectiveness of a proposed new method , 5) To develop information services or new platforms, 6) To create a combined or composite dataset, and 7) To provide a broader context for a study's subject or topic. We noted that the categories were not mutually exclusive; an author may use multiple OGD sources in a single publication, or may use a single OGD source for multiple purposes in the same publication. To ensure our shared understanding of the merged codebook, Cohen's Kappa was calculated in a final round of coding (n=50 articles). Nearly perfect agreement was achieved for usage types 1), 2), 5), 6), and 7), and substantial agreement was achieved for usage types 3) and 4). The first author then coded the rest of the samples based on the mutual understanding of the codebook.

Additional Analysis of Sample: We analyzed publications in our sample based on publication year, in order to understand the trend of OGD usage over time, as well as

the source of OGD, in order to understand which data portals were most popular. We also analyzed the publication outlets (e.g. journal, conference, etc) to gain an understanding of how various research fields engage with OGD. We associated each outlet with its research area using Scopus source title list [2] which provided information of each publication outlet indexed by Scopus. We discuss these results below.

4   Results

Descriptive Analysis: Figure 2 shows the number of articles that used OGD from January 2009 to July 2017. There is an upward trend of publications using OGD over time, implying the increasing attention to OGD in scientific research. This is consistent with previous findings that as open data initiatives mature, portals are more frequently used by diverse stakeholders[12].
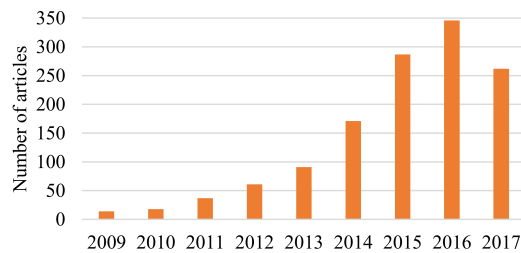

Fig. 2. Number of articles using OGD by year

The sources of OGD research are provided in Figure 3. Data from 96 different Open Government Data Portals were used by publications in our sample. Notably, the United Kingdom's OGD portal (data.gov.uk) was used by 25.5% of all papers in our sample. This is more than twice the number of articles (11.6%) that used the United States OGD portal (data.gov). It is not surprising that national portals are more frequently used than state, province, city or county portals due to the fact that both the UK and USA portals have more data available. National OGD portals, such as data.gov, also harvest data from sub-national portals at the city and state level. However, it is surprising that Chicago is ranked as the sixth most frequent source of OGD (3.6 %). Chicago is the most popular OGD portal among all US city or county level portals. This is also surprising as other major USA cities such as New York City, Seattle, and Boston have more data available for download, and have holdings which are more extensive than those of Chicago [13]. Among international regions, Amsterdam (3.8%) and Queensland (2.6%) received the most use. This result is in line with Safarov et al.'s [8] findings that OGD-related studies were most active in developed

countries with the Netherlands, the United States, the United Kingdom as the most researched countries. However, OGD in two developing countries, India (7.0%) and Kenya (6.3%), notably appear in the top 5 of all sources found in our study. This is a significant finding, as it suggests that although these two nations have relatively new OGD programs, they are a unique and valuable source of data for researchers working on topics related to these countries.
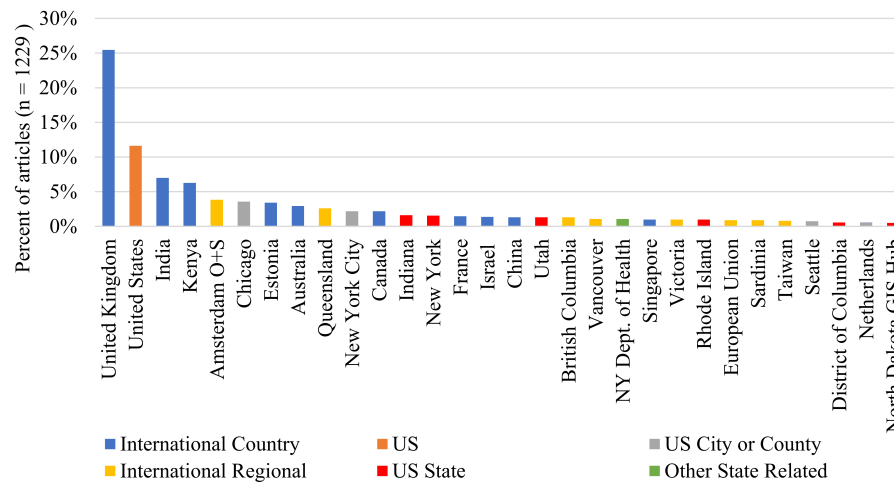
Fig. 3. Top 30 sources of OGD used in scientific research

Research Areas That Use OGD: We examined the publication outlets for the selected articles in our sample (n=1229). Among the 815 different publication outlets, PloS ONE produced the most (32) articles, followed by BMC Public Health (14), Science of the Total Environment(12), Sustainability (Switzerland) (11), BMJ Open (11), and Scientific Reports (9). Notably, all of these journals, with the exception of Science of the Total Environment, are open-access publications. This suggests that authors who use open data may be more likely to publish in open-access venues, or that these publication outlets have more strict standards about data source citation. However, further research is required to validate this observation.

We also associated the title of each publication in our sample with a corresponding field of research as defined by the Scopus database. Table 2 shows that Medicine, Environmental Sciences and Social Sciences occupy the top three fields that use OGD in publications, suggesting that OGD is playing the role of advancing both natural science and social science discoveries. Computer Sciences and Engineering also rank high in the list (No. 4 and No. 6), suggesting that OGD is also contributing to technical innovations. Noticeably, OGD was used by all available Scopus research fields, including Dentistry and Chemistry. We caution that the research area of a publication outlet does not necessarily represent the research areas of a paper published

in it, but we do argue that these labels are a strong proximate indicator of the topic of the article. Thus, the breadth and diversity of research fields using OGD demonstrates that these PSI are indeed a valuable source for scientific research.

Table 2. Top 10 research areas that use OGD

Research Areas	Count
Medicine	557
Environmental Science	448
Social Sciences	412
Computer Science	242
Agricultural and Biological Sciences	178
Engineering	145
Earth and Planetary Sciences	144
Biochemistry, Genetics and Molecular Biology	80
Energy	76
Business, Management and Accounting	45

OGD Usage Types: Through inductive coding, seven major OGD usage types were identified. OGD were used 1) As the main data source for new analysis, such as being a variable in regression or a parameter in a mathematic model; 2) As ancillary data source for new analysis, such as the base map upon which newly collected data are plotted; 3) For result evaluation, such as to validate or verify a research claim, or to benchmark the quality of an existing dataset; 4) To demonstrate the effectiveness of a proposed new method (e.g., an innovative statistical model); 5) To develop information services or new platforms (e.g., a semantic web browser); 6) To create a combined or composite dataset; and 7) To provide study context, such as describing the motivation for a study. Figure 4 illustrates the distribution of OGD uses from our study sample. Over half of the publications used OGD as data sources for new research: 33.4% used OGD as main source and 19.5% as an ancillary source. In our sample, 33.2% of the articles cited OGD to provide background context for the study, including demographic information about a population of study, or as validation for a chosen sampling method. Less then 5% of the articles belong to other usage types that were unclear from the text of the article.

Limitations: There are several limitations to this study. First, and most significantly, our search only retrieved articles that explicitly cite, mention, or link to an open government data portal. Many researchers likely use OGD, but do not explicitly document the source. We present these results as preliminary findings about how OGD are used, and discuss future ways in which implicit OGD use could be identified below. Second, we only considered OGD sources retrieved from an initial list of 302 portals. Although we believe this list to be extensive, there are certainly sources of OGD that were not included in our search. Third, we only included three databases
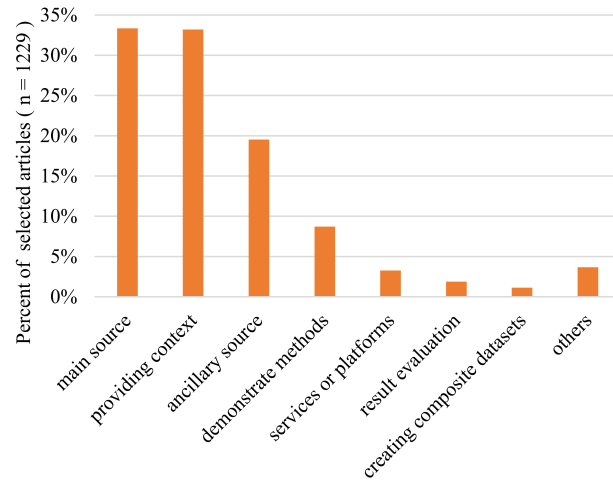
Fig. 4. Percentage of usage types by article

in our analysis and most selected articles were retrieved from Scopus, therefore our results relied heavily on the content coverage of the Scopus database. Our categorization of research areas is also constrained by the way that Scopus assigns research areas to each publication outlet. In the future, we hope estimate the impact of OGD on academic research by incorporating journal impact factors, and citation counts of selected articles into our analysis. We also hope to understand researchers' practices with OGD by conducting surveys and interviews. By combining the literature mining approach described here with survey and qualitative methods, we believe these results could be improved. Below, we restate the key findings of this initial work.

5 Discussion & Conclusion

This paper is one of the first empirical examinations of how OGD are being used in scientific research. By mining a total of 1229 research articles that used OGD, we found that 1. The use of OGD in research is steadily increasing from 2009 to 2016; 2. Researchers use OGD from 96 different open government data portals in their publications, with data.gov.uk and data.gov being the most frequent sources; 3. Contrary to previous findings, we provide evidence that OGD from developing nations, notably India and Kenya, are popular amongst OGD users; 4. OGD has been used by nearly all research fields, with Medicine, Environmental Science, and Social Sciences being the most active fields that engage with OGD; and 5. Researchers mainly use OGD as data sources for new analysis and to provide context for a research finding. OGD was also used for testing new methods, providing new services or systems, result evaluation, and creating composite datasets.

Popular OGD Sources: We discovered in our results that United Kingdom, Kenya, and Chicago ranked high in our sources of OGD. Looking further into the use of these sources, we found out that each of the three portals had one particular type of dataset that had been used most frequently. For example, in the United Kingdom, an "index of deprivation" [3] dataset was used frequently as an ancillary source for Medicine and Public Health research [14]. For Kenya, demographic data was often used to describe the study setting for research on poverty in developing countries [15]. For Chicago, crime data [4] was the most used source for social, urban, and computer science studies [16]. The frequent use of a particular dataset from a particular data portal requires further consideration: First, as official data published by a government entity, these datasets may be the only authoritative sources for certain types of information, such as demographic characteristics of a population. This would explain why, for example, the city of Chicago appears high in our rankings of ODG sources. Researchers interested in crime may find this the best, or only source to answer a research question. Second, these key sources of data can help answer questions that are of great local interest, but the remaining datasets published by a government may be of limited use to researchers outside of a particular topic. For example, if we removed poverty data from the Kenya portal we would get substantially different results as to the popularity of developing nations ODG source. But, these results imply that governments should seek to identify highly unique data, and promote their use amongst research communities. In turn, this could spur further engagement with a data portal. Further, governments should consider prioritizing the release of datasets that satisfy the above-mentioned conditions in order to maximize the utilization of their investment in open government data initiatives.

Research Areas: Finally, we highlight our finding that nearly all scientific fields are using OGD. Previous studies [17] have shown that OGD may be used by researchers to investigate education, policies, health-care, and government activities. But, our results suggest that OGD can also be applied to broad research areas such as chemistry [18] and Dentistry [19] that seem less likely to use OGD and that OGD has exhibited more potential use for research than reflected in exiting literature.

6 Acknowledgement

References

1. Jaeger, P.T., Bertot, J.C.: Transparency and technological change: Ensuring equal and sustained public access to government information. Gov. Inf. Q. 27(4), 371-376 (2010). doi: 10.1016/j.giq.2010.05.003

---

[3] https://data.gov.uk/dataset/index-of-multiple-deprivation

[4] https://data.cityofchicago.org/public-safety/crimes-2001-to-present/ijzp-q8t2

2. Davies, T.: Open data, democracy and public sector reform. A look at open government data use from data. gov. uk (2010)
3. Zuiderwijk, A., Marijn J.: Open data policies, their implementation and impact: A framework for comparison." Gov. Inf. Q. 31(1), 17-29 (2014). doi: 10.1016/j.giq.2013.04.003
4. Gruen, N., Houghton, J., Tooth, R.: Open for business: How open data can help achieve the G20 growth target. (2014)
5. Zuiderwijk, A., Marijn J., Yogesh K.D.: Acceptance and use predictors of open data technologies: Drawing upon the unified theory of acceptance and use of technology. Gov. Inf. Q. 32(4),429-440 (2015). doi: 10.1016/j.giq.2015.09.005
6. Bright, J., Margetts, H.Z., Wang, N., Hale, S.A.: Explaining Usage Patterns in Open Government Data: The Case of Data.Gov.UK, Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2613853, (2015). doi: 10.2139/ssrn.2613853
7. Young, M., and Yan A.: Civic Hackers' User Experiences and Expectations of Seattle's Open Municipal Data Program. In: Proceedings of the 50th Hawaii International Conference on System Sciences (2017). doi: 10.24251/HICSS.2017.324
8. Safarov, I., Meijer, A., Grimmelikhuijsen, S.: Utilization of open government data: A systematic literature review of types, conditions, effects and users. Inf. Polit. 1-24, Preprint (2017). doi: 10.3233/IP-160012
9. Martin, E.G., Helbig, N., Birkhead, G.S.: Opening health data: what do researchers want? Early experiences with New York's open health data platform. J. Public Health Manag. Pract. 21(5), E1–E7 (2015). doi: 10.1097/PHH.0000000000000127
10. Fecher, B., Friesike, S., Hebing, M.: What Drives Academic Data Sharing? PloS one, 10(2), e0118053 (2015). doi: 10.1371/journal.pone.0118053
11. Landis, J.R., Gary G.K.: The measurement of observer agreement for categorical data. Biometrics, 159-174 (1977). doi: 10.2307/2529310
12. Thorsby, J., Stowers, G.N., Wolslegel, K. and Tumbuan, E. : Understanding the content and features of open data portals in American cities. Gov. Inf. Q. 34(1), 53-61 (2017). doi: 10.1016/j.giq.2016.07.001
13. Kassen, M.: A promising phenomenon of open data: A case study of the Chicago open data project. Gov. Inf. Q. 30(4), 508-513 (2013). doi: 10.1016/j.giq.2013.05.012
14. Evans, J., Kaptoge, S., Caleyachetty, R., Di Angelantonio, E., Lewis, C., Parameshwar, K., Pettit, S. J.: Socioeconomic Deprivation and Survival After Heart Transplantation in England An Analysis of the United Kingdom Transplant Registry. Circ. Cardiovasc. Qual. Outcomes 9(6), 695–703 (2016). doi: 10.1161/CIRCOUTCOMES.116.002652
15. Okotto, L., Okotto-Okotto, J., Price, H., Pedley, S., Wright, J.: Socio-economic aspects of domestic groundwater consumption, vending and use in Kisumu, Kenya. Appl. Geogr. 58, 189–197 (2015). doi: 10.1016/j.apgeog.2015.02.009
16. Zheng, Y., Liu, T., Wang, Y., Zhu, Y., Liu, Y., Chang, E.: Diagnosing New York City's Noises with Ubiquitous Data. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 715–725. ACM, New York (2014). doi: 10.1145/2632048.2632102
17. Graves A., Hendler J.: A study on the use of visualizations for Open Government Data. Inf. Polit 19, 73–91 (2014). doi: 10.3233/IP-140333
18. Croft, B., Wentworth, G.R., Martin, R.V., Leaitch, W.R., Murphy, J.G., Murphy, B.N., Kodros, J.K., Abbatt, J.P., Pierce, J.R.: Contribution of Arctic seabird-colony ammonia to atmospheric particles and cloud-albedo radiative effect. Nature communications 7, 13444 (2016). doi: 10.1038/ncomms13444
19. de Silva, A.M., Gkolia, P., Carpenter, L., Cole, D.: Developing a model to assess community-level risk of oral diseases for planning public dental services in Australia. BMC Oral Health 16(1), 45 (2016). doi: 10.1186/s12903-016-0200-5