The Report Committee for Nicholas George Hadjigeorge
Certifies that this is the approved version of the following report:

Civic Engagement Analysis of Select Open Data Portals

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Kenneth Flamm

Sherri R. Greenberg

Civic Engagement Analysis of Select Open Data Portals

by

Nicholas George Hadjigeorge, B.A.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Public Affairs

The University of Texas at Austin

May 2015

Dedication

This report is dedicated to the public servants and volunteers who make the open data and open government movement possible.

Acknowledgements

This Report would not have been possible without the help and support of the following people: Professor Kenneth Flamm and Professor Sherri Greenberg, Frances Hadjigeorge, George Hadjigeorge, Fanny Trang, Mateo Clarke, Chip Rosenthal, Nathan Brigmon, Parham Daghighi, Brian O'Donnell, Luqmaan Dawoodjee, Thomas Levine, and Stuart Gano.

Abstract

Civic Engagement Analysis of Select Open Data Portals

Nicholas George Hadjigeorge, MPAff

The University of Texas at Austin, 2015

Supervisor: Kenneth Flamm

Governments and non-profit institutions increasingly are publishing data on the Internet in databases known as open data portals. The data is freely available, accessible, and used by citizens and application developers. However, there is little research about the relationship between open data portal characteristics and their effect on civic engagement. In addition to an overview of open data portals, user behavior, and civic applications, this report analyzes portal-level panel data across 36 months and 14 portals to estimate the effects of portal characteristics on civic engagement. Results show that portals with frequently updated data experience more civic engagement. These empirical findings validate the open data principle of timeliness, and are important for policy-makers designing open data policies to maximize the potential for users to add value to open data.

Table of Contents

List of Tables ix

List of Figures x

INTRODUCTION **1**

Background 1

Open Data 1

Open Data Portals 4

User Behavior 4

Data Discovery 4

Data Innovation 6

Portal Information Extraction 6

CIVIC ENGAGEMENT EXAMPLES **8**

METHODOLOGY **14**

Dataset Creation 17

Model Design 18

Dependent Variables 18

Independent Variables 23

FINDINGS **26**

Rows-Accessed Model Findings 26

Rows-Loaded Model Findings 27

Discussion 28

RECOMMENDATIONS **31**

Open Data Analytics 31

Open Data Quality 31

CONCLUSION **32**

Appendix A 33

Appendix B 35

References 37

List of Tables

Table 1: Open Data Portals Used in Analysis 17

Table 2: Rows Accessed Variables 33

Table 3: Rows Loaded Variables 34

Table 4: Data Quality and Data Quantity Variables 34

List of Figures

Figure 1: Austin Finance Online eCheckbook Dataset Example.....5

Figure 2: Instabus.....8

Figure 3: VoteATX.....9

Figure 4: ATXfloods.....10

Figure 5: Pet Alerts.....11

Figure 6: 311 vs 10-ONE District Map.....12

Figure 7: Data.austintexas.gov Analytics Dashboard.....15

Figure 8: API request for data.austintexas.gov Analytics Metadata in 2013.....16

Figure 9: Rows-Loaded, Disaggregated.....20

Figure 10: Rows-Accessed, Disaggregated.....21

Figure 11: Rows-Loaded, Aggregated.....22

Figure 12: Rows-Accessed, Aggregated.....23

Figure 13: Independent Variables.....25

Figure 14: Rows-Accessed-Total Model Summary (Model 1).....26

Figure 15: Rows-Loaded-Total Model Summary (Model 2).....27

Figure 16: Model 1 R Code.....35

Figure 17: Model 2 R Code.....36

INTRODUCTION

Open data is expanding the traditional notions of civic engagement to include new forms of citizen participation. This report explores the relationship between open data and a particular form of participation - civic engagement driven by innovative uses of open data. Using an econometric analysis, I explain the variation of civic engagement with open data across a sample of open data portals. The purpose of this analysis is to determine how a portal's characteristics affect this type of civic engagement. In addition, the findings in this report may be used to help governments better manage their portals to increase engagement and maximize the potential for value added to open data.

Background

This section will define the basic principles of open data, describe the characteristics of open data portals (henceforth referred to as "portals"), and describe portal user behavior. It will also serve as a review of open data literature.

OPEN DATA

Governments and not-for-profit institutions are publishing open data on the Internet using database platforms known as portals. Open data must adhere to certain openness principles in order to achieve positive outcomes. The principles include data completeness, timeliness, accessibility, and machine readability, among others.[1] Without these characteristics, a dataset may not be suitable for application development, data analysis, or general usability, and therefore may hinder civic engagement or value-adding activities.

[1] "Open Data Policy Guidelines," *Sunlight Foundation*, accessed April 25, 2015, http://sunlightfoundation.com/opendataguidelines/.

Governments bear the primary responsibility for the quality of their open data.[2] Historically, governments were mainly focused "on increasing the openness and provision of information or documents rather than on opening up the data underlying the production of information, documents, etc."[3]

Providing and maintaining open data is a shift away from simply curating information produced from government data. Now, governments "may often only play the role of arbiter, co-ordinator, funder and regulator for the activities of others in delivering public value through the use of public sector information and data."[4]

There are four groups that may benefit from the publication and use of open data: government, citizens, civil society, and the wider economy. The benefits generally include economic development, effective governance, civic engagement, and government accountability and transparency.[5][6]

Open data allows the government to improve service provision, increase resource allocation efficiency, and improve relations between the government and citizens.[7] The government is able to achieve these results since open data allows governments to "create, promote, and execute information-based policies."[8] Regarding citizens, open data has the potential to improve quality of life and improve decision-making by enabling

---

[2] Ubaldi, Barbara. "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives", OECD Working Papers on Public Governance, No. 22. OECD Publishing, 2013. http://dx.doi.org/10.1787/5k46bj4f03s7-en

[3] Ibid, 18.

[4] Ibid, 18.

[5] Emmy Tran and Ginny Scholtes. "Open Data Literature Review." Https://www.law.berkeley.edu. April 14, 2015. Accessed April 17, 2015. https://www.law.berkeley.edu/wp-content/uploads/2015/04/Final_OpenDataLitReview_2015-04-14_1.1.pdf.

[6] This report does not focus on the relationship between open data and government accountability and transparency.

[7] Supra, note 2.

[8] Supra, note 5.

civic and social engagement via applications using open data.[9] Civil society organizations such as Open Austin and the Sunlight Foundation may benefit by organizing citizens around open data issues.[10] Finally, the private sector benefits by creating services that add value to open data.[11]

Open data generates large amounts of economic value. Economic value may take the form of increased efficiency in government service provision, or as private sector profit generated by using open data. The U.S. Department of Commerce estimates that open data may generate somewhere between $24 and $211 billion in annual private sector revenues.[12] At the federal level, the U.S. government has spent an average of $3.7 billion annually on open data initiatives since 2004, which equates to .02 percent of the $17 trillion U.S. economy.[13]

More specifically, open data leads to economic value by increasing the potential for private sector innovation using information generated by government services and practices. Tim O'Reilly argues that open data is crucial for this innovation process to take place, since it allows "the private sector to build applications that government didn't consider or doesn't have the resources to create."[14]

By definition, open data does not include information that is personally identifiable. Privacy issues are a major component of open data policy and practice. However, this report does not focus on privacy issues.[15]

---

[9] Supra, note 2.
[10] Ibid, 12.
[11] Ibid, 12.
[12] "Fostering Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data." Economics & Statistics Administration. July 1, 2014, http://www.esa.doc.gov/reports/fostering-innovation-creating-jobs-driving-better-decisions-value-government-data (accessed April 17, 2015).
[13] Ibid.
[14] Tim O'Reilly, "Government as a Platform," *Innovations: Technology, Governance, Globalization* 6, no. 1 (January 1, 2011): 13–40, doi:10.1162/INOV_a_00056.
[15] For discussion and analysis of open data privacy issues, see Meijer et al (2013).

OPEN DATA PORTALS

Open data is generated through the process of an institution carrying out its various operations. These operations include government finance, planning, economic development, environmental services, law enforcement services, fire department services, parks and recreation services, and others. After open data is generated, it can be stored on an open data portal and accessed by the public. Before data is uploaded to a portal, it may be manipulated to ensure privacy and data quality standards.

For this report, I focus on portal analytics from a single portal provider, Socrata. Socrata is a software-as-a-service provider that currently hosts 136 portals.[16] Other portal platforms include Junar and CKAN.[17][18] These platforms are not included in the following analysis because analytics from their portals were not readily available at the time of analysis.

USER BEHAVIOR

I specify two categories of portal user behavior: data discovery and data innovation.

Data Discovery

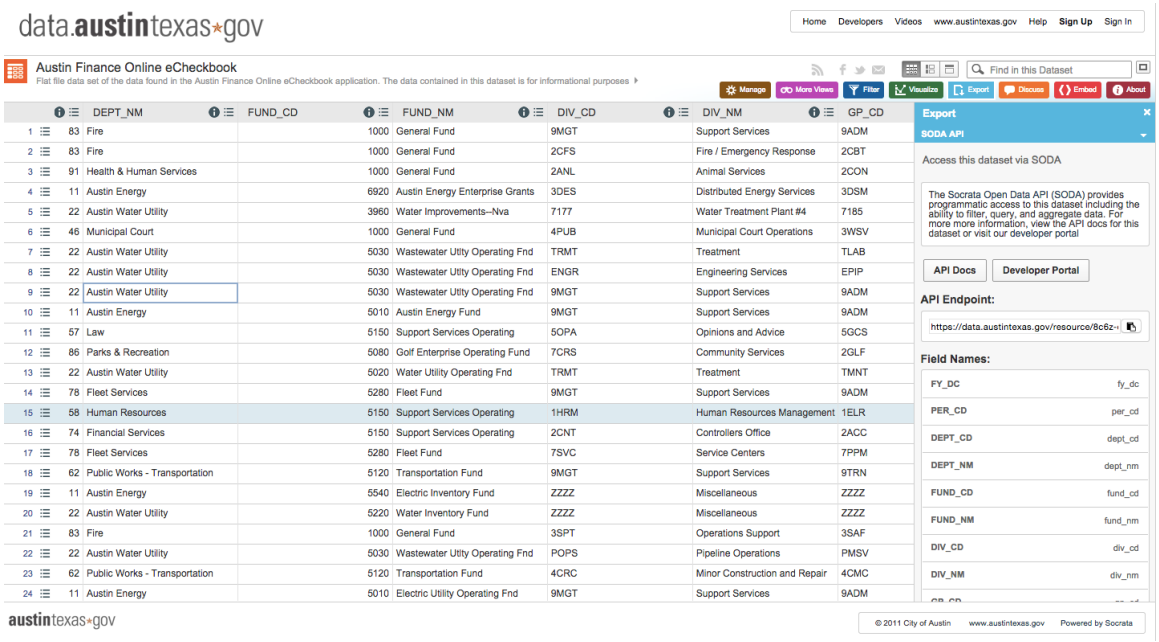
Data discovery may take the form of a user searching a portal for budget information. For users engaging in data discovery, the Socrata platform contains basic analysis functions via in-browser tools. These include filtering, sorting, and graphing tools.

The figure below displays a dataset from the City of Austin data portal describing expenditure data.

---

[16] "Socrata Customer Spotlights | Socrata." OpenData by Socrata. https://opendata.socrata.com/dataset/Socrata-Customer-Spotlights/6wk3-4ija (accessed March 18, 2015).
[17] "About | ckan – The open source data portal software," accessed April 25, 2015, http://ckan.org/about/.
[18] "The Open Data Platform · Junar," accessed April 25, 2015, http://www.junar.com/open-data.

Figure 1: Austin Finance Online eCheckbook Dataset Example[19]



The data contained within this dataset may be consumed using the in-browser tools, or may be extracted through several extraction methods. Data extraction allows the user to explore data in ways that are not limited by the built-in capabilities of the portal. Portals may facilitate government transparency and accountability by allowing users to discover data.

[19] This dataset can be found at https://data.austintexas.gov/Financial/Austin-Finance-Online-eCheckbook/8c6z-qnmj.

Data Innovation

The second category of portal use is data innovation. Data innovation may take the form of a user extracting geo-location data for use in a web application or for research purposes.[20] Data innovation often leads to the creation of civic applications by volunteer developers or profit-seeking firms. Civic applications may be used by the public, often as a supplement to a government service, or as a new type of service previously not provided by the government. Civic engagement associated with the use of civic applications plays a key role in generating value and realizing the potential benefits of open data.

Civic applications, in virtue of their use of portal data, may also serve as methods of data discovery for civic applications users. For example, a civic application using government expenditure data and the portal where the dataset is stored may provide a user with the same data and data discovery features. The distinction between data discovery and data innovation is the potential for adding value to open data. Data innovation is more likely to be the source of added value for open data, while data discovery is more likely to facilitate government transparency and accountability.

Both behavior types represent civic engagement with a portal and can be measured via the Socrata site metrics API.

PORTAL INFORMATION EXTRACTION

There are several ways to extract data from a Socrata portal. These include the Socrata Development API (SODA), direct download in select file formats, Rich Site Summary (RSS) feed, and other Socrata tools. It is possible for data discovery and data

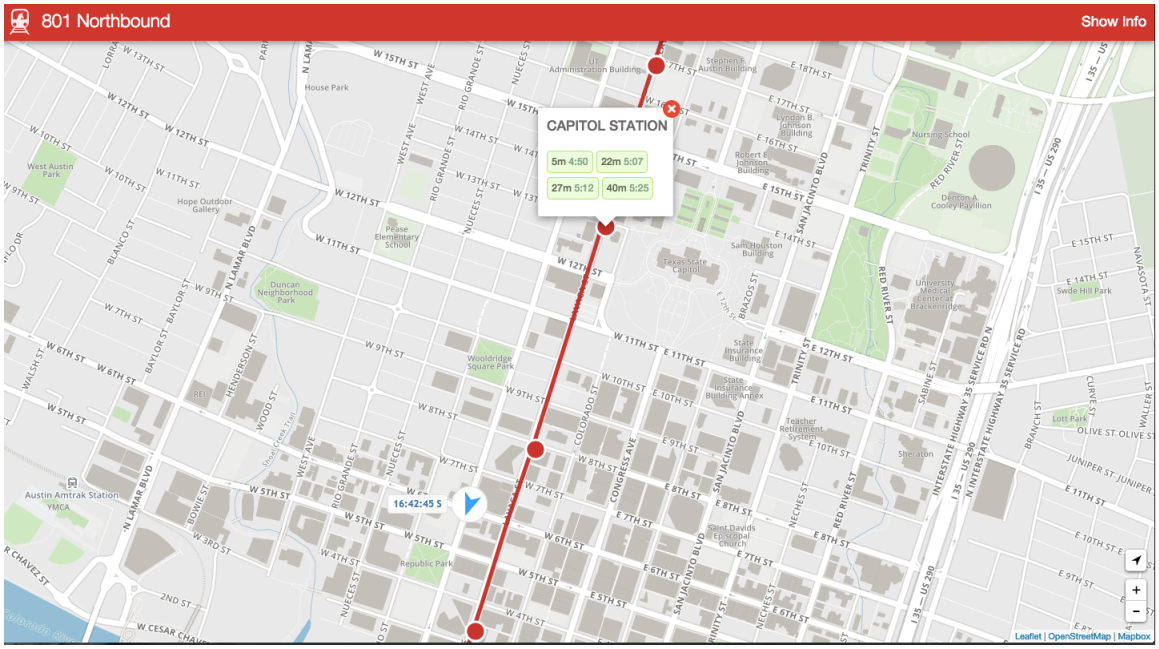
[20] I define a civic application as any Internet-based application or technology that extracts government open data to power any of its features.

innovation to power civic engagement or software application development through any combination of data extraction methods and Socrata portal features.

In the next section, I will discuss examples of data innovation relying on open data and representing civic engagement with the City of Austin portal and other sources of Central Texas open data.

CIVIC ENGAGEMENT EXAMPLES

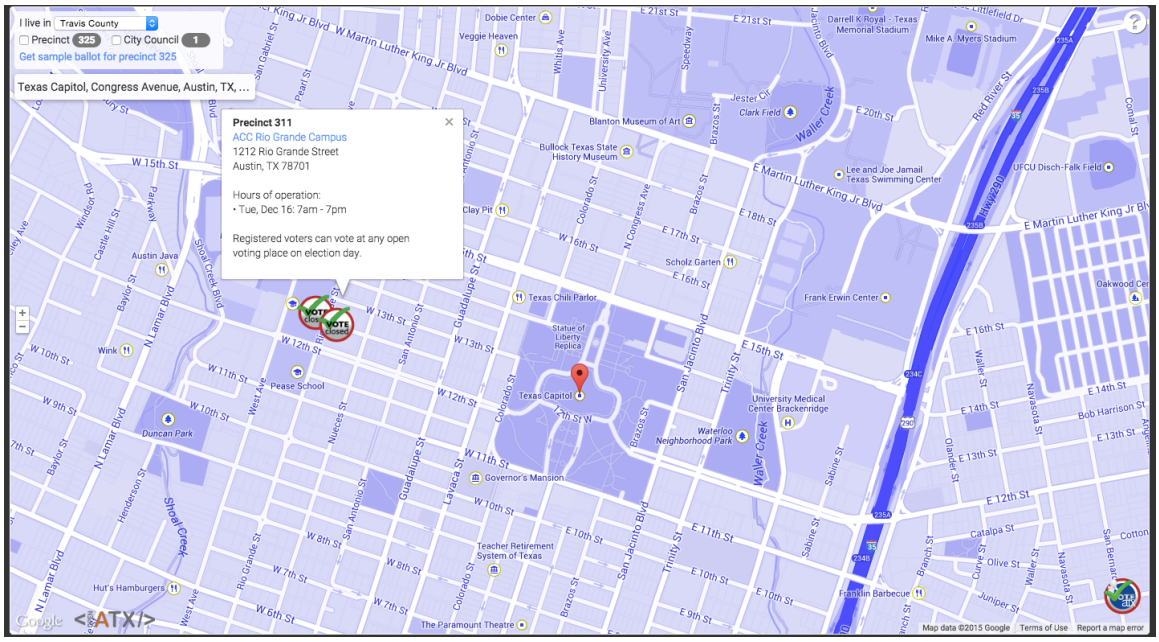
Figure 2: Instabus



Instabus.org is a civic application that uses open transit data from Austin's transit agency, Capital Metro. The data is updated in real time, and is accessible via Austin's Socrata portal. The application is device-agnostic, which allows users to access the application's features via a web browser on any internet-enabled device. The application displays the location of all transit vehicles for a given route and displays vehicle stop locations. The user may select a stop and receive information regarding expected vehicle arrival times. Luqmaan Dawoodjee and Darrell Maples developed the application in coordination with a local civic developer organization, Open Austin.[21]

[21] "luqmaan/Instabus," *GitHub*, https://github.com/luqmaan/Instabus (accessed April 25, 2015).

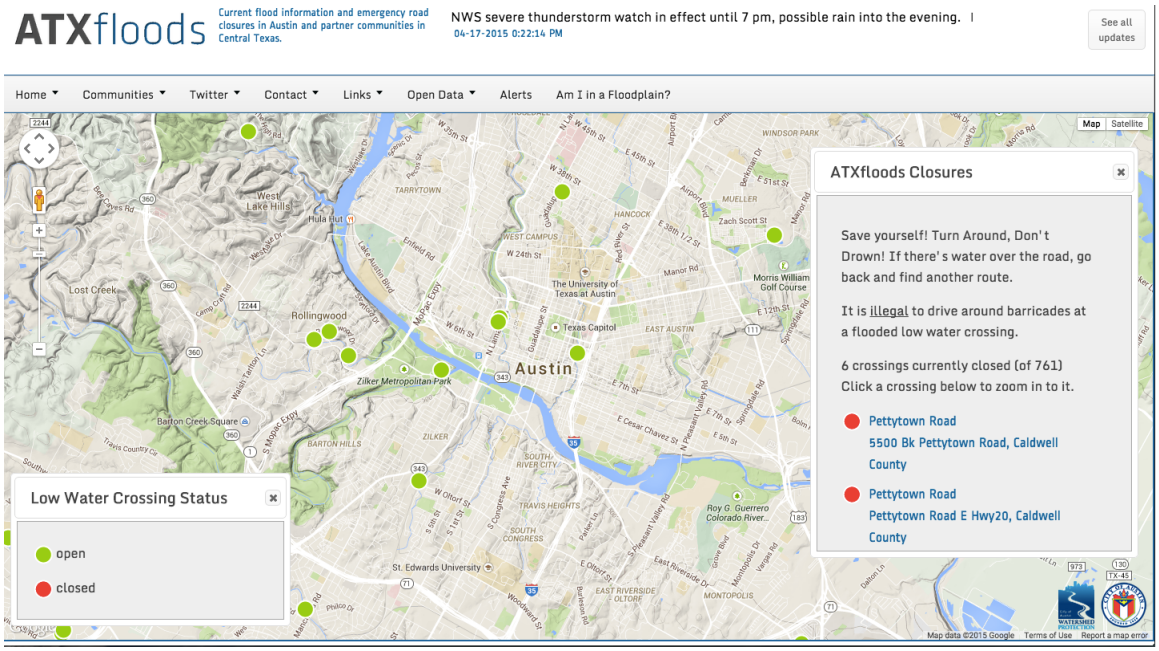
Figure 3: VoteATX



VoteATX.us is a device-agnostic civic application that displays polling locations, polling location hours of operation, and appropriate polling locations according to a user's home address. The application uses static polling location data provided by the Travis County Clerk and Travis County Tax-Assessor, which is then displayed using Google Maps. Chip Rosenthal developed the application in coordination with Open Austin.[22]

[22] "Open-Austin/voteatx-App," *GitHub*, https://github.com/open-austin/voteatx-app (accessed April 25, 2015).

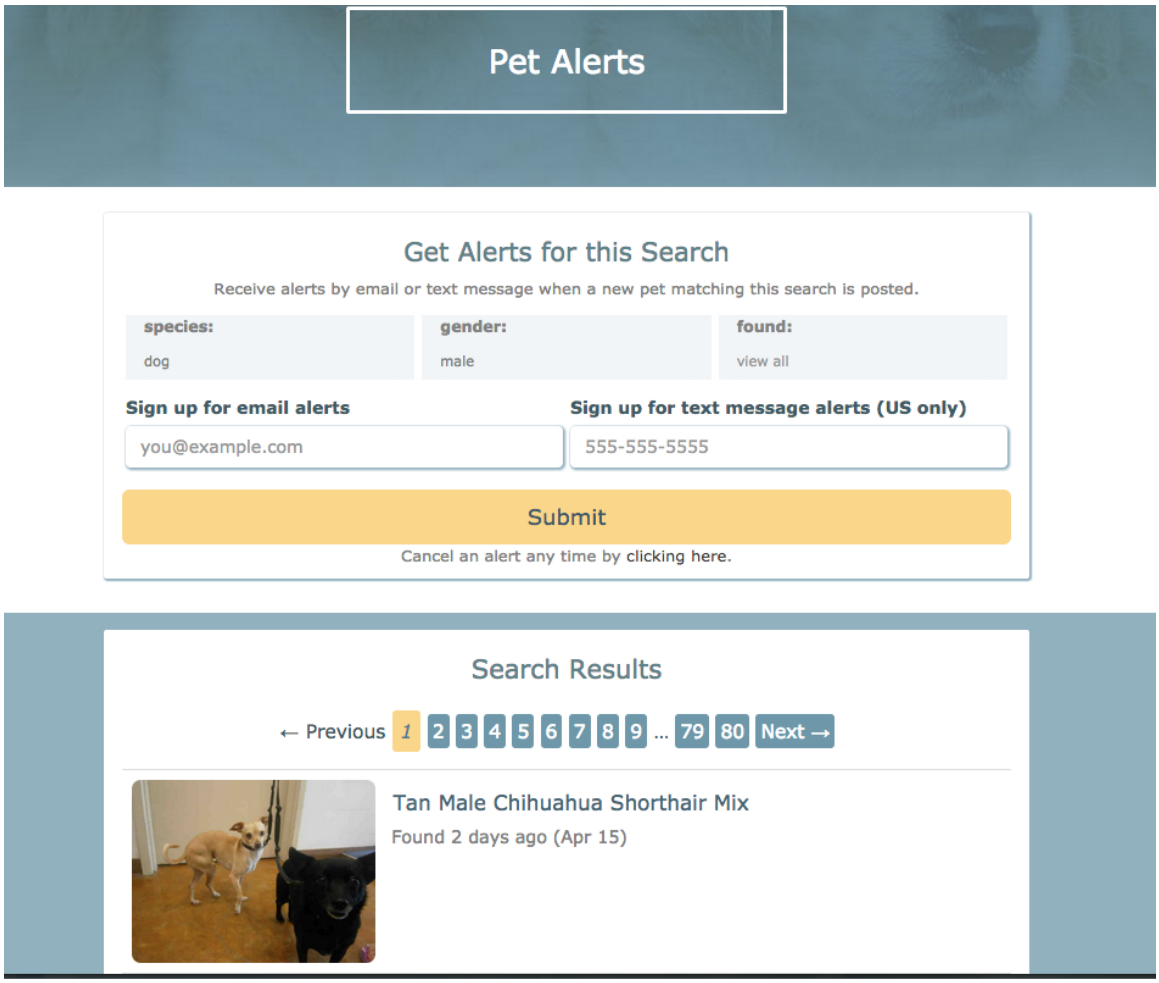
Figure 4: ATXfloods



ATX Floods is a device-agnostic civic application that uses real time low water crossing data to display the status of low water crossings on an interactive map. The application is used to alert drivers when low water crossings are open or closed. The application is unique in that it relies on a network of remote sensors to gather data at the location of each low-water crossing.[23] A team of Code for America fellows developed the application in 2012, and the City of Austin Flood Early Warning System (FEWS) team currently maintains the service.[24]

[23] For a discussion of real time open data, see Kitchin (2013)
[24] http://www.atxfloods.com/

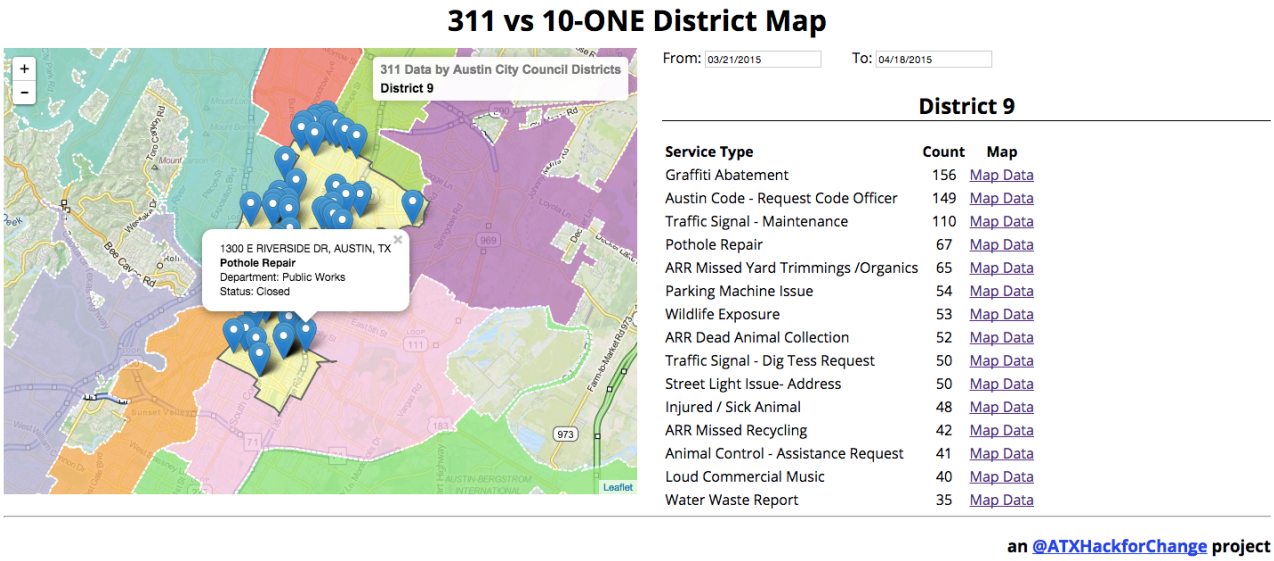
Figure 5: Pet Alerts



Pet Alerts is a device-agnostic civic application that uses data from Austin's portal to display information and alert users regarding Austin Animal Center intake data. The application's purpose is to reunite pet owners with their lost pets. When an animal arrives at the Austin Animal Center, the animal's information and photo are entered into a database that is uploaded to the portal on a daily basis. Users can view animal intake data according to the specifications of their lost animal, and receive alerts when an

animal matching the description arrives at the Animal Center. Natalia Shelburne and Tim Shelburne developed the application in coordination with Open Austin.[25]

Figure 6: 311 vs 10-ONE District Map



The 311 vs 10-ONE District Map is a civic application that displays 311 service request data for a given Austin City Council district and for a given time range. Both the district map data and the 311 service request data are housed on Austin's portal. Users can specify a district and a time range and receive geographic data for various service requests. Selecting a particular service request provides more details, such as street address and request status. Mateo Clarke developed the application during the June 2014 ATX Hack for Change hackaton.[26]

[25] See https://github.com/open-austin/pet-finder for the Pet Alerts source code.
[26] See https://github.com/mateoclarke/311vs10One for the 311 vs 10-ONE District Map source code.

The degree to which civic engagement can be measured as a result of the use of these applications is dependent on user analytic features built into the application itself, as well as the analytic features built into the portal.

Therefore, analyzing civic engagement with civic applications as the unit of analysis is difficult, given that analytic features may not exist for a given civic application or may use non-standard analysis methods. Studying user analytics at the application level could potentially provide more detailed information regarding civic engagement with open data, but nevertheless is not the focus of this report.

However, site metrics retrieved from data portals may provide insights into civic engagement with open data. Therefore, the remainder of this report focuses on portals as the unit of analysis. In the next section, I discuss my methodology for analyzing civic engagement with 14 Socrata portals.

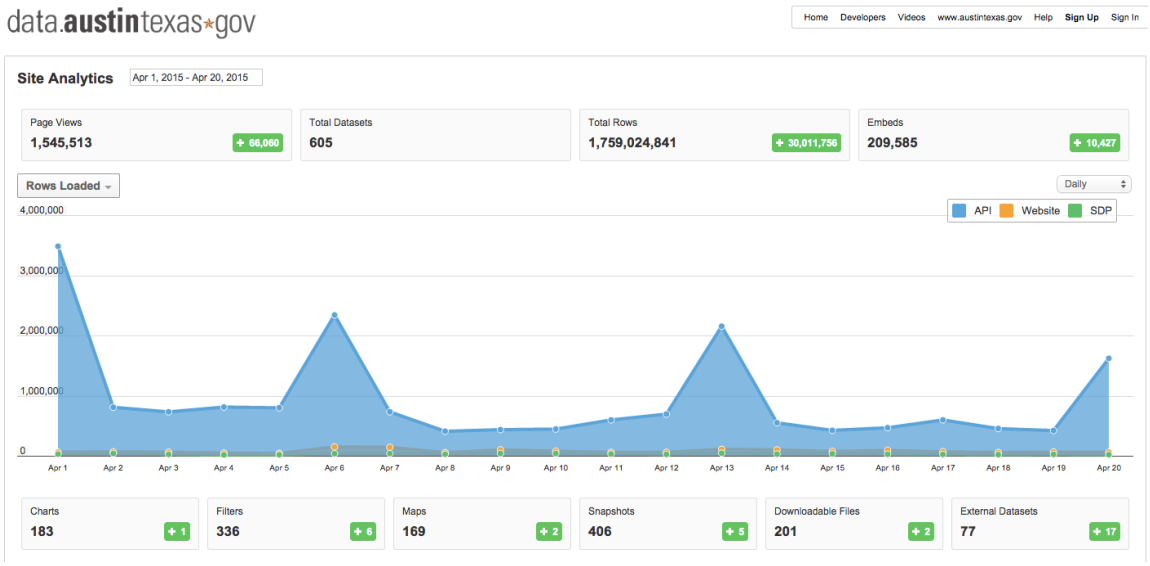
METHODOLOGY

For this report, I created a 36-month panel dataset containing publically available analytics data from 14 portals. The analysis period begins in January 2012, and ends in December 2014.[27] The metadata is count data measuring the characteristics of a given portal, including how many rows of data were created, deleted, accessed, and shared.[28]

Portal metadata is accessible through an online dashboard and an application program interface (API). The dashboard provides a curated visualization of portal analytics, while the API provides the raw data that is used to create the visualization. An API request for portal analytics returns a JavaScript Object Notation (JSON) file containing raw portal metadata for a given time range. The following figures illustrate the Socrata portal analytics dashboard and the information returned from a portal analytics API request.

[27] Not all portals in the sample contain data from the entire analysis period. For those portals, data collection starts when the portal begins publishing metadata.
[28] See the Appendix A for a comprehensive list of Socrata portal metadata variables and definitions.

Figure 7: Data.austintexas.gov Analytics Dashboard[29]



[29] The Austin Texas site analytics dashboard can be found at https://data.austintexas.gov/analytics.

Figure 8: API request for data.austintexas.gov Analytics Metadata in 2013[30]


```
{
  "datasets-total" : 605,
  "datasets-created-total" : 15310,
  "datasets-deleted-total" : 13739,
  "datasets-created-snapshot-total" : 8772,
  "datasets-deleted-snapshot-total" : 8366,
  "datasets-created-blobby-total" : 266,
  "datasets-deleted-blobby-total" : 65,
  "datasets-created-href-total" : 85,
  "datasets-deleted-href-total" : 8,
  "charts-created-total" : 5193,
  "charts-deleted-total" : 5010,
  "maps-created-total" : 13320,
  "maps-deleted-total" : 13151,
  "filters-created-total" : 21222,
  "filters-deleted-total" : 20886,
  "rows-created-total" : 3479935747,
  "rows-deleted-total" : 1720910906,
  "page-views-total" : 1545524,
  "embeds-total" : 209591,
  "embeds" : 43649,
  "datasets-deleted-blobby" : 19,
  "maps-created" : 3633,
  "page-views" : 247743,
  "bytes-out" : 597498458637,
  "datasets-created-blobby" : 29,
  "rows-loaded-print" : 5593162,
  "rows-loaded-api" : 25618634,
  "datasets-deleted-snapshot" : 1615,
  "datasets-deleted-href" : 1,
  "rows-accessed-website" : 661752,
  "rows-loaded-download" : 321626475,
  "rows-accessed-api" : 421565,
  "rows-loaded-website" : 9057712,
  "ratings-count" : 2,
  "datasets-created-snapshot" : 1704,
  "rows-accessed-rss" : 3951,
  "rows-deleted" : 534665653,
  "maps-deleted" : 3594,
  "datasets-created-href" : 5,
  "filters-created" : 4593,
  "rows-loaded-widget" : 1875517,
  "rows-accessed-widget" : 124857,
  "geocoding-requests" : 237106,
  "users-created" : 346,
  "js-page-view" : 602776,
  "datasets-created" : 4458,
  "rows-accessed-download" : 9761,
  "datasets-deleted" : 4150,
  "view-loaded" : 92990,
  "app-token-created" : 12,
  "charts-deleted" : 962,
  "rows-accessed-print" : 1926,
  "shares" : 172,
  "rows-loaded-rss" : 164027,
  "bytes-in" : 2468485404,
  "ratings-total" : 80,
  "filters-deleted" : 4521,
  "charts-created" : 1017,
  "rows-created" : 1103453534,
```


[30] This particular site metrics API request can be acquired at https://data.austintexas.gov/api/site_metrics.json?start=1357002061000&end=1388451661000.

The following table displays the 14 portals described in my analysis, along with their URLs and data collection starting dates.

Table 1: Open Data Portals Used in Analysis

Government/Institution	Portal URL	Starting Date
The City of Austin, TX	Data.austintexas.gov	01/01/2012
The City of New York, NY	Data.cityofnewyork.us	01/01/2012
The State of Hawaii	Data.hawaii.gov	04/01/2012
Lehman College	Bronx.lehman.cuny.edu	08/01/2012
The City of San Francisco, CA	Data.sfgov.org	01/01/2012
The City of Baltimore, MD	Data.baltimorecity.gov	01/01/2012
The City of Raleigh, NC	Data.raleighnc.gov	03/01/2012
The State of Oklahoma	Data.ok.gov	01/01/2012
The City of Seattle, WA	Data.seattle.gov	01/01/2012
Montgomery County, MD	Data.montgomerycountymd.gov	02/01/2012
World Bank Group Finances	Finances.worldbank.org	01/01/2012
The City of Boston, MA	Data.cityofboston.gov	05/01/2012
The City of Kansas City, MO	Data.kcmo.gov	10/01/2012
The City of Providence, RI	Data.providenceri.gov	03/01/20132

DATASET CREATION

To determine which Socrata portals have publicly available metadata, I used a shell script to check for the existence of a "Portal Analytics" page in each of the 136 portals.[31] Of the 136 portals, I determined that 14 portals had publicly available metadata. With the help of local civic developer Mateo Clarke, I designed a JavaScript application to extract metadata from the Socrata Open Data API for each of the 14 portals, and in monthly increments.[32] Using another online tool, I compiled the JSON metadata into CSV format.[33]

---

[31] "Tlevine/socrata-Nominate," *GitHub*, https://github.com/tlevine/socrata-nominate (accessed March 20, 2015).
[32] "Mateoclarke/socrata-Data-Portal-Analytics," *GitHub*, https://github.com/mateoclarke/socrata-data-portal-analytics (accessed March 20, 2015).
[33] "JSON to CSV," http://konklone.io/json/ (accessed March 20, 2015).

MODEL DESIGN

Portal metadata from a site metrics API request is returned in the form of non-negative count variables. The Poisson distribution is the nominal distribution for count data. It provides consistent and asymptotically normal estimators of the independent variables whether or not the underlying distribution is Poisson.[34] The count data used in this model exhibits heteroskedasticity given that the residual errors do not have equal variances across portals. To correct for heteroskedasticity, I use the HC2 estimation to calculate robust standard errors.[35]

I implement the model using the following R statistical packages: glm, sandwich, and lmtest.[36]

DEPENDENT VARIABLES

I define the following dependent variables as measurements of civic engagement. These variables fall into two categories: rows-accessed and rows-loaded. Rows-accessed is a count variable measuring access requests, and rows-loaded is a count variable measuring the number of rows of data loaded during access requests. For example, in January 2012, the City of Austin portal had 2,774,114 rows-loaded over the course of 1,852 rows-accessed requests. It is not possible to determine which rows were loaded, nor to determine the row's destination.

I aggregate four types of rows-accessed and rows-loaded variables to reflect the multiple methods of data extraction available on a Socrata portal. Aggregating the

---

[34] Wooldridge, Jeffrey M. "Limited Dependent Variable Models and Sample Selection Corrections." In *Introductory Econometrics: A Modern Approach*, 595-600. 4th ed. Mason, Ohio: South-Western Cengage Learning, 2009.

[35] Long, J. Scott, and Laurie H. Ervin, "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," *The American Statistician* 54, no. 3 (August 1, 2000): 217–24, doi:10.1080/00031305.2000.10474549.

[36] "nghadji/Socrata-Data-Portal-Analytics," *GitHub*, https://github.com/nghadji/Socrata-Data-Portal-Analytics (accessed April 25, 2015).

dependent variables serves to capture both data discovery and data innovation activity resulting from data extraction. For example, the API extraction method is more likely to be used by civic applications, rather than the print function, which exports a portable document file (PDF). Nevertheless, both methods may be used for data innovation and data discovery.

I did not include variables of the type "website," or "widget." These are measures of data extraction from Socrata managed applications, and do not necessarily reflect civic engagement or data innovation.

The following figures illustrate the eight dependent variables before aggregation for each portal, separated by category. The y-axis displays variable counts with a logarithmic transformation, and the x-axis displays monthly time units. Month 1 is January 1, 2012, and Month 36 is December 31, 2014. I used the ggplot2 package in R to create the following figures.[37]

[37] Supra, note 36.

Figure 9: Rows-Loaded, Disaggregated

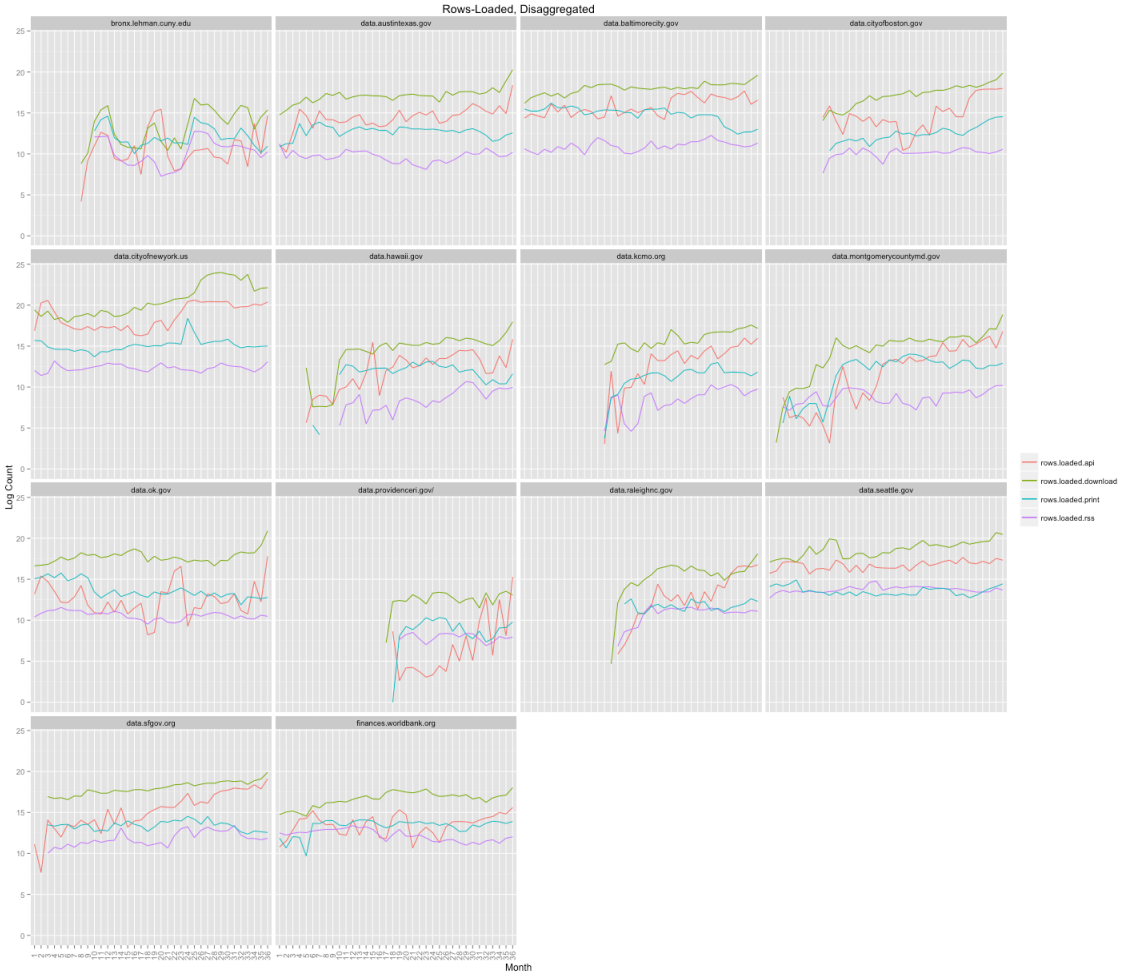


Figure 10: Rows-Accessed, Disaggregated

The following figures illustrate the dependent variables after aggregation. I use the aggregate totals of rows-loaded and rows-accessed in my model.

Figure 11: Rows-Loaded, Aggregated

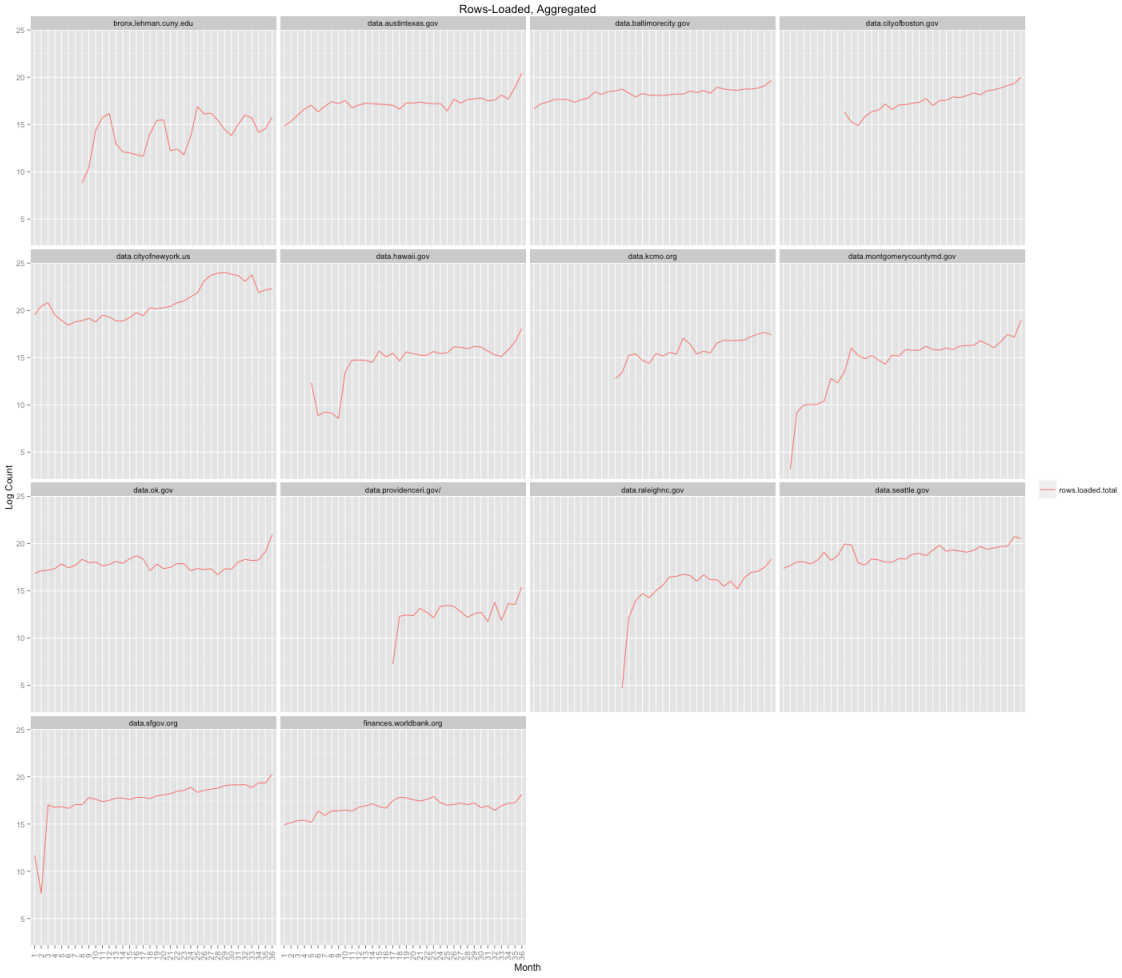
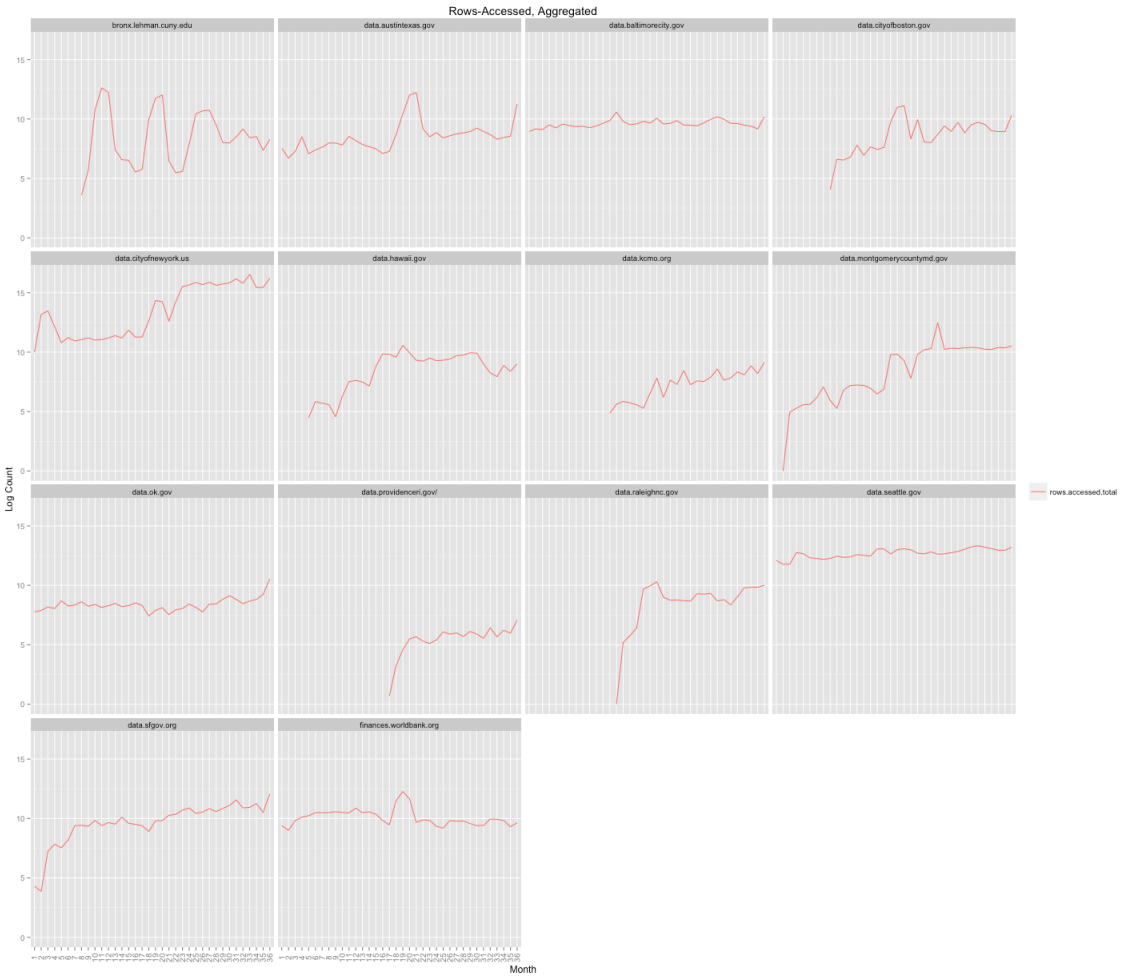


Figure 12: Rows-Accessed, Aggregated



INDEPENDENT VARIABLES

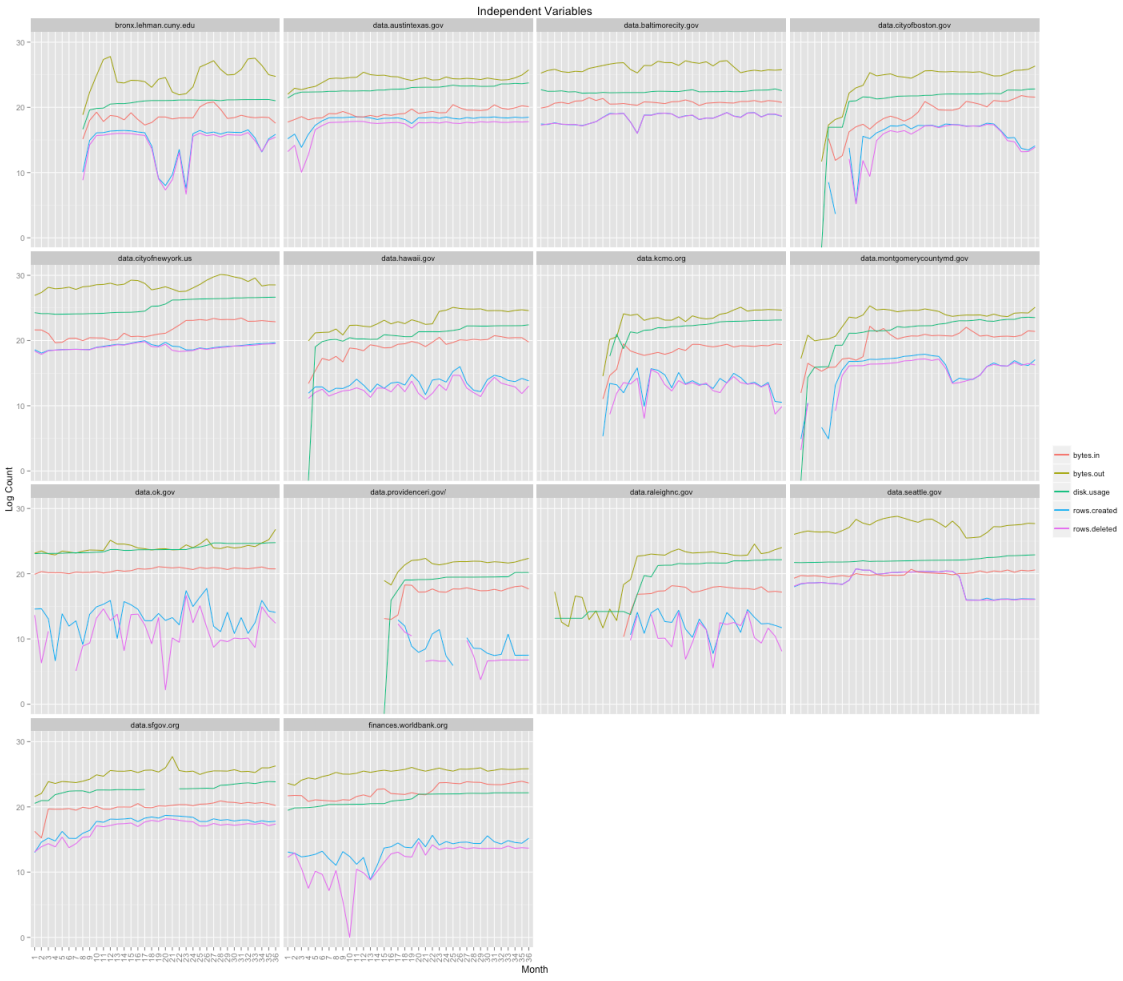
I define the following independent variables as indicators of data quality and data quantity. The independent variables measure the number of rows-created, rows-deleted, incoming bytes of data (bytes-in), outgoing bytes of data (bytes-out), and disk-usage. I interpret rows-created and rows-deleted as indicators of data quality. The net effect of row creation and deletion reflects the degree to which data on a portal is updated or refreshed. Data that is refreshed very often (every minute, hour, etc.) is known as real-time data.

I interpret incoming and outgoing bytes of data, and disk usage as a measure of portal data quantity. Bytes-in reflects the amount of data a portal receives in the form of rows, datasets, and all other types of data. Bytes-out reflects the amount of all outgoing data from a portal. Disk usage reflects the amount of data on the portal as a whole. Data quantity variables only measure the file size of data being transmitted to and from a portal, and do not measure any qualitative characteristics of portal data.

I include a two-month lag variable for each of the aggregated dependent variables. The lag variable accounts for the potential delay between portal activity and civic engagement activity. I also include factor variables for month and institute of higher education status. The time factor accounts for the effects of time across all portals. The institute of higher education factor accounts for one portal in my sample, Bronx.lehman.cuny.edu, and the effects of the school year calendar on its portal activity.

The following figure illustrates the independent variables, excluding the dummy and factor variables.

Figure 13: Independent Variables



In the next section, I will discuss the findings from this model.

FINDINGS

The following figure displays the results from the Rows-Accessed-Total model.

Figure 14: Rows-Accessed-Total Model Summary (Model 1)[38]

	Estimate	Robust SE	Pr(>|z|)
(Intercept)	11.3900814646431	0.6731069126446	0.0000000000000 **
rows.created	-0.0000000047481	0.0000000039837	0.2333058079841
rows.deleted	0.0000000078619	0.0000000039941	0.0490254876343 **
bytes.in	-0.0000000000106	0.0000000000221	0.6298701602116
bytes.out	0.0000000000001	0.0000000000001	0.0660269943455 *
disk.usage	0.0000000000172	0.0000000000017	0.0000000000000 **
rows.accessed.total.lag	-0.0000001722987	0.0000000523945	0.0010072677877 **

Note: ** = < .05 significance. * = < .1 significance.

ROWS-ACCESSED MODEL FINDINGS

Recall that rows-accessed is a measure of the number of access requests for a given portal during a given time period. According to the Rows-Accessed model (Model 1), the number access requests increases by 7.86e-09 for every row of data that is deleted, and decreases by 4.75e-09 for every row of data that is created. The net effect of rows-created and rows-deleted is an increase of 3.11e-09 access requests.

I interpret this result as evidence that data quality has a net positive effect on portal data access requests. In other words, if a portal deletes more rows than it creates, then that condition will have a positive effect on portal data access requests. Deleting rows indicates that data is being updated or refreshed.

Regarding data quantity, the number of access requests decreases by 1.06e-11 for each byte entering a portal, and increases by 1.25e-13 for each byte leaving a portal. I

[38] See Appendix B for the R code used to generate Model 1.

interpret this result as evidence that civic engagement decreases as data entering a portal increases.

For data exiting a portal, I interpret the obverse to be true. Civic engagement increases as data exiting the portal increases. However, the net effect of bytes-in and bytes-out results in overall decreased civic engagement.

For disk usage, which is the size of all data contained on a portal, the number of rows-accessed increases by 1.72e-11 for each additional byte of portal size. I interpret this as evidence that civic engagement increases as a portal's disk-usage increases.

ROWS-LOADED MODEL FINDINGS

The following figure displays the results from the Rows-Loaded-Total model.

Figure 15: Rows-Loaded-Total Model Summary (Model 2)[39]

	Estimate	Robust SE	Pr(>|z|)
(Intercept)	18.6975396447205	0.7069625379999	0.0000000000000 **
rows.created	0.0000000043124	0.0000000036483	0.2371872074618
rows.deleted	-0.0000000033659	0.0000000035368	0.3412604952607
bytes.in	-0.0000000000074	0.0000000000143	0.6027171556916
bytes.out	0.0000000000005	0.0000000000001	0.0000000000000 **
disk.usage	0.0000000000069	0.0000000000016	0.0000138898555 **
rows.loaded.total.lag	-0.0000000000126	0.0000000000200	0.5280623877036

Note: ** = < .05 significance. * = < .1 significance.

Recall that rows-loaded is a measure of the number of individual dataset rows loaded during a given number of access requests. According to the Rows-Loaded model (Model 2), the number of rows-loaded increases by 4.31e-09 for each row created, and

[39] See Appendix B for the R code used to generate Model 2.

the number of rows-loaded decreases by 3.7e-09 for each row deleted. The net effect of rows created and deleted on the number of rows-loaded is .95e-09 rows-loaded.

I interpret this result as evidence that data quality has a net positive effect on the number of portal data rows-loaded. Contrary to the findings in Model 1, Model 2 suggests that rows-loaded increases when more rows are created than deleted.

Regarding data quantity, the number of rows-loaded decreases by 7.4e-12 for each byte entering a portal, and increases by 4.6e-13 for each byte exiting a portal. I interpret this result as evidence that civic engagement decreases as data entering a portal increases.

For data exiting a portal, I interpret the obverse to be true. Civic engagement increases as data exiting the portal increases. However, the net effect of bytes-in and bytes-out results in overall decreased civic engagement.

For disk-usage, the number of rows-accessed increases by 6.8e-12 for each additional byte of portal size. I interpret this as evidence that civic engagement increases as a portal's disk-usage increases.

DISCUSSION

The results from Model 1 and Model 2 suggest that both measures of civic engagement are positively related to indicators of high quality data. The net effect of the number of rows-created and rows-deleted, when positive, leads to increased civic engagement in both the number of access requests a portal receives and the number of rows-loaded during those requests. In terms of rows-accessed, deleting rows has a stronger effect than creating rows. In terms of rows-loaded, creating rows has a stronger effect than deleting rows.

I will present a hypothetical example to illustrate my interpretation of the results. Let us use the Pet Alerts civic application as an example. The application uses an Animal Center dataset stored on Austin's portal. Whenever the dataset changes to reflect updated information, as it does every hour, the Pet Alerts application will make an API request to extract the updated information. I assume Pet Alerts extracts information as soon as the Animal Center dataset is updated.

The Socrata site metrics API reports this interaction as one rows-accessed-API. Suppose the dataset initially contains 100 rows, and the updated version contains 150 rows. The site metrics API reports the interaction as 150 rows-loaded. We can infer that some combination of rows-created and rows-deleted took place, resulting in 50 new or updated rows. Given the capabilities of the site metrics API, it is not possible to know which rows were deleted and updated, deleted and not updated, or created as new rows.

Since the dataset is updated hourly, Pet Alerts contributes one rows-accessed-API request every hour, and loads as many rows as the updated dataset contains. In this example, rows-accessed is a function of the number of time Pet Alerts accesses the API to extract updated information, and is determined by the update frequency of the dataset. Rows-loaded is a function of the data contained within the dataset at the time of extraction and a reflection of how the dataset changes over time.

Rows-deleted has a stronger effect on rows-accessed, since deleting and updating rows is a condition that leads to civic applications extracting updated information at intervals determined by a dataset's update frequency. The more frequently a dataset is updated, the more rows-accessed requests a civic application will add to the total. Therefore, the findings suggest that rows-accessed may be more accurately described as a measure of the connectedness of civic applications with data on a portal.

Rows-created has a stronger effect on rows-loaded, since the number of rows-loaded increases if there have been more rows-created, and decreases if there have been more rows-deleted. Therefore, the findings suggest that rows-loaded may be more accurately described as a measure of the quantity of data that civic applications are extracting.

In terms of byes-in, bytes-out, and disk-usage, both models suggest mixed results and small effects on civic engagement. As the amount of information entering a portal increases, civic engagement decreases. However, civic engagement slightly increases as the total size of the portal increases. These indicators, while measurements of portal quantity, may not provide sufficient information to study civic engagement.

RECOMMENDATIONS

OPEN DATA ANALYTICS

Government open data mangers should implement robust analytics practices in their open data policies to measure and understand civic engagement with open data. This includes analyses like the one found in this report, as well as detailed studies of individual civic applications. Whenever feasible, governments using the Socrata platform should always enable the Site Analytics feature.

In addition, civic application developers should implement standardized analytics features into their applications to allow for cross-application analyses of civic engagement with open data. In conjunction with portal analysis, civic application analysis has the potential to provide greater insights into civic engagement with open data.

OPEN DATA QUALITY

In terms of data quantity, this analysis suggests that having the right kind of data may be more important for civic engagement than having a large amount of data. Governments should ensure that data on their portals is not duplicative, redundant, or unnecessarily storage space-intensive. In addition, governments should implement open data standards to ensure their data is published in a timely manner and is appropriately updated to maintain timeliness.

CONCLUSION

Open data creates enormous opportunities for innovation in both the public and private sectors. These innovations can lead to more efficient government operation and large amounts of economic value. The value added to open data is the result of civic engagement.

It is necessary for citizens to engage with open data in order for the potential benefits to appear. However, government data must adhere to openness principles if citizens are to use open data effectively for innovative purposes. Governments are responsible for ensuring that open data is of sufficient quality to allow data innovation to occur.

Open data innovation often results in the creation of civic applications by volunteer developers or profit-seeking firms. These applications can supplement existing government services or create entirely new ones. Governments and citizens are thereby able to access services and information that would likely not have been created in the absence of volunteer or profit-seeking motivations. This form of data innovation is a key component of adding value to open data.

The results from this analysis suggest that indicators of open data quality have a positive effect on civic engagement with open data. The net effect of the number of rows-created and rows-deleted, an indicator of data freshness, is positively related to civic engagement – a necessary component of open data value creation. The findings validate the importance of the open data timeliness principle. Governments should strive to publish and maintain data that is timely, or else risk the loss of value associated with untimely data.

Appendix A

Table 2: Rows Accessed Variables[40]

rows-accessed-api	API calls returning or accessing rows.
rows-accessed-download	Download requests.
rows-accessed-print	Row data requests for PDF export.
rows-accessed-rss	Row data requests for RSS Feeds.
rows-accessed-website	Row data requests from Socrata managed websites.
rows-accessed-widget	Row data requests from Socrata managed "widgets."

[40] This information was provided by Socrata Data Analyst Stuart Gano via email correspondence.

Table 3: Rows Loaded Variables[41]

rows-loaded-api	Total number of rows loaded for API requests.
rows-loaded-download	Total number of rows loaded for CSV, Excel, etc. requests.
rows-loaded-print	Total number of rows loaded for PDF exports.
rows-loaded-rss	Total number of rows loaded by RSS feeds.
rows-loaded-website	Total number of rows loaded for Socrata managed websites.
rows-loaded-widget	Requests for row data for Socrata managed "widgets."

Table 4: Data Quality and Data Quantity Variables[42]

disk-usage	Disk usage for all data for the domain, in bytes
rows-created	Rows created, including republish cycles
rows-deleted	Rows deleted, including republish cycles
bytes-in	Total bytes-in for a domain for API, data, most Assets, and blobs
bytes-out	Total bytes-out for a domain for API, data, most Assets, and blobs

[41] Ibid.
[42] Ibid.

Appendix B

Figure 16: Model 1 R Code

```
## Model 1
model1 <- glm(rows.accessed.total ~ rows.created + rows.deleted
              + bytes.in + bytes.out + disk.usage + rows.accessed.total.lag
              + factor(month) + factor(school), family=poisson(),
              data = panel.data)


## Calculate Robust Standard Errors for Model 1
library(sandwich)
library(lmtest)

cov.m1 <- vcovHC(model1, type="HC2")
std.err <- sqrt(diag(cov.m1))
r.est <- cbind(Estimate= coef(model1), "Robust SE" = std.err,
               "Pr(>|z|)" = 2 * pnorm(abs(coef(model1)/std.err), lower.tail=FALSE),
               LL = coef(model1) - 1.96 * std.err,
               UL = coef(model1) + 1.96 * std.err)

print(r.est)
```

Figure 17: Model 2 R Code

```
## Model 2
model2 <- glm(rows.loaded.total ~ rows.created + rows.deleted
              + bytes.in + bytes.out + disk.usage + rows.loaded.total.lag
              + factor(month) + factor(school), family=poisson(),
              data = panel.data)

summary(model2)

## Calculate Robust Standard Errors for Model 2
library(sandwich)
library(lmtest)

cov.m2 <- vcovHC(model2, type = "HC2")
std.err <- sqrt(diag(cov.m2))
r.est2 <- cbind(Estimate= coef(model2), "Robust SE" = std.err,
               "Pr(>|z|)" = 2 * pnorm(abs(coef(model2)/std.err), lower.tail=FALSE),
               LL = coef(model2) - 1.96 * std.err,
               UL = coef(model2) + 1.96 * std.err)

print(r.est2)
```

References

"About | ckan – The open source data portal software," http://ckan.org/about/ (accessed April 25, 2015).

"ATX floods." http://www.atxfloods.com/ (accessed April 25, 2015).

Clarke, Mateo. "mateoclarke/311vs10One," *GitHub*, https://github.com/mateoclarke/311vs10One (accessed April 25, 2015).

Clarke, Mateo. "Mateoclarke/socrata-Data-Portal-Analytics," *GitHub*, https://github.com/mateoclarke/socrata-data-portal-analytics (accessed March 20, 2015).

Dawoodjee, Luqmaan. "luqmaan/Instabus," *GitHub*, https://github.com/luqmaan/Instabus (accessed April 25, 2015).

Gano, Stuart. "Dataset Meta Data Metrics." Email correspondence.

Hadjigeorge, Nicholas. "nghadji/Socrata-Data-Portal-Analytics," *GitHub*, https://github.com/nghadji/Socrata-Data-Portal-Analytics (accessed April 25, 2015).

Kitchin, Rob. *The Real-Time City? Big Data and Smart Urbanism*, SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, July 3, 2013), http://papers.ssrn.com/abstract=2289141.

Levine, Thomas. "Tlevine/socrata-Nominate," *GitHub*, https://github.com/tlevine/socrata-nominate (accessed March 20, 2015).

Long, J. Scott, and Laurie H. Ervin, "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," *The American Statistician* 54, no. 3 (August 1, 2000): 217–24, doi:10.1080/00031305.2000.10474549.

Meijer, Ronald, Peter Conradie, and Sunil Choenni, "Reconciling Contradictions of Open Data Regarding Transparency, Privacy, Security and Trust," *J. Theor. Appl. Electron. Commer. Res.* 9, no. 3 (September 2014): 32–44, doi:10.4067/S0718-18762014000300004.

Mill, Eric. "JSON to CSV," http://konklone.io/json/ (accessed March 20, 2015).

"Open Data Policy Guidelines," *Sunlight Foundation*, http://sunlightfoundation.com/opendataguidelines/ (accessed April 25, 2015).

O'Reilly, Tim. "Government as a Platform." *Innovations: Technology, Governance, Globalization* 6, no. 1 (January 1, 2011): 13–40, doi:10.1162/INOV_a_00056.

Rosenthal, Chip. "Open-Austin/voteatx-App," *GitHub*, https://github.com/open-austin/voteatx-app (accessed April 25, 2015).

Shelburne, Natalia, and Tim Shelburne. “Open-Austin/pet-Finder,” *GitHub*, https://github.com/open-austin/pet-finder (accessed April 25, 2015).

"Socrata Customer Spotlights | Socrata." OpenData by Socrata. https://opendata.socrata.com/dataset/Socrata-Customer-Spotlights/6wk3-4ija (accessed March 18, 2015).

“The Open Data Platform · Junar,” http://www.junar.com/open-data (accessed April 25, 2015).

Tran, Emmy, and Ginny Scholtes. "Open Data Literature Review." Https://www.law.berkeley.edu. April 14, 2015. https://www.law.berkeley.edu/wp-content/uploads/2015/04/Final_OpenDataLitReview_2015-04-14_1.1.pdf (Accessed April 17, 2015).

Ubaldi, Barbara. *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives*. OECD Working Papers on Public Governance, No. 22. OECD Publishing, 2013. http://dx.doi.org/10.1787/5k46bj4f03s7-en.

U.S Department of Commerce, Economics & Statistics Administration. *Fostering Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data*. 2014, Report. http://www.esa.doc.gov/reports/fostering-innovation-creating-jobs-driving-better-decisions-value-government-data.

Wooldridge, Jeffrey M. "Limited Dependent Variable Models and Sample Selection Corrections." In *Introductory Econometrics: A Modern Approach*, 595-600. 4th ed. Mason, Ohio: South-Western Cengage Learning, 2009.