Contextualizing Open Data Quality:

Institutional Factors and the Openness of City Data Portals

Pranathi Siva Posa

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Arts in Public Policy

at the University of Chicago

Preceptor: Rachel Dec

Abstract

With municipal data portals becoming more commonplace among big cities, the ten Open Government Data Principles principles provide a benchmark for the effective release of government data, possibly serving as a goal for cities to aspire towards achieving in the name of increasing transparency, participation, and innovation. This paper examines the data portals of 32 cities to assess their adherence with the ten principles, and conducts an investigation into what institutional factors may have significant effect on a city's overall data quality and its adherence to specific principles. It discovers that cities have strong OGD foundations, but that much improvement is still required to achieve full adherence, and that there are no consistent associations between institutional factors and performance on individual data principles. Going forward, it is critical to establish guidelines for what data should be published, and for cities to take greater ownership over data catalog construction.

Table of Contents

Contextualizing Open Data Quality:

Institutional Factors and the Openness of City Data Portals

---

1. Introduction

Cities are no strangers to generating data, it's a critical part of daily city function. But in the last twelve years we've started to see cities take the step from just generating data to releasing it.

Releasing data is heralded as a way for cities, big and small, to be transparent, participatory, & innovative. It has resulted in a movement for, specifically "open" government data. Open Government Data (OGD) is the idea that data generated by citizens and collected by the government should be available for anyone to view, use, and manipulate (Sunlight Foundation 2010). The notion of open data is viewed as the next logical step to give citizens access to public information they can privately use to understand their government and community, without jumping through loopholes to request it.

These data sources are often centralized by governments into data portals, webpages that allow citizens to access datasets and any related information. Datasets are often uploaded and updated directly by city agencies, and then utilized by interested citizens to whatever end they may wish. These may include things like creating apps that aggregate transport information (Citymapper 2015), or something as simple as being able to compare changes in community property values over a ten-year period by using datasets for each year.

But municipal data portals are by no means standardized in the quality or breadth of information they make available to their data-generating communities. And a lack of standardization

in quality means it is far more difficult for these cities to reach their OGD goals and really serve their citizens with open data that is actually useful and accessible. As such, it is critical to evaluate municipal data portals and understand the quality of the information that is being provided and where the release of data can be improved.

1.1. Open Data Standards

In 2007, a group of OGD advocates came together in California to establish the first set of OGD principles. These were eight principles that were meant to guide what good OGD practices looked like, and included the following:

1. **Complete:** All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.
2. **Primary:** Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
3. **Timely:** Data is made available as quickly as necessary to preserve the value of the data.
4. **Accessible:** Data is available to the widest range of users for the widest range of purposes.
5. **Machine processable:** Data is reasonably structured to allow automated processing.
6. **Non-discriminatory:** Data is available to anyone, with no requirement of registration.
7. **Non-proprietary:** Data is available in a format over which no entity has exclusive control.
8. **License-free:** Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

(Open Government Data Principles, 2007)

Two additional principles were added to this list in 2010 by the Sunlight Foundation, and were:

9. **Permanence**: Data is available over time, such that older data is archived and accessible.
10. **Usage Costs:** Data is accessible without users having to pay a fee.

(Sunlight Foundation, 2010)

The aim of these standards is to ensure that open data functions in a manner that's actually useful to the citizen, and to the government. Not complying with any one of these standards would make access or use of open data difficult, antithetical to the goals of OGD advocates. This study will utilize these standards as part of its method to evaluate the quality of data in municipal open data portals.

While most OGD frameworks adopted by governments around the world are rooted in these ten principles, actual compliance with these principles is an ongoing endeavor. This may be due to a variety of factors, such as misalignment between available data and citizen needs, lack of institutionalized processes that reduce the time necessary to publish data, and lack of consideration for reusability—there is quite a way to go before governments have the kinds of mechanisms in place to make full compliance commonplace (Sayogo 2014).

1.2. National & State Open Data Initiatives

In 2009, the Obama administration began the push for open publication of federal government data with the release of the Transparency and Open Government Memorandum. This memorandum announced the Obama administration's intention to spearhead an open government initiative and called for the creation of the Open Government Directive. This directive set deadlines for the publishing of government datasets, the improvement of government standards for data quality, the development of strategies to institutionalize open government, and built a policy framework for enabling the continued evolution of open government (White House 2009). In calling for the publishing of government datasets, the directive called for the creation of a federal open data portal, and in May of 2009, Data.gov was launched with an initial 47 datasets.

Shortly after the creation of Data.gov, many other countries followed in establishing their own data portals, with the UK launching Data.gov.uk in September of 2009, and New Zealand launching data.govt.nz in November of 2009. These countries also established their own frameworks for the goals of engaging in OGD practices (UK 2009; New Zealand 2011). Currently, over fifty-three countries around the world have established their own national data portals (Data.gov 2022).

Despite the early creation of Data.gov, it wasn't until 2018 that federal standards for data release were codified into statute with the passage of H.R.4174 - Foundations for Evidence-Based Policymaking Act. Title II of this act established the OPEN Government Data Act, under which "government data is required to be made available in open, machine-readable formats, while continuing to ensure privacy and security." After the passage of this act, it became a requirement that

federal agencies adhere to specific publication and metadata practices when releasing data through their own websites or Data.gov.

Currently, Data.gov hosts over 340,000 datasets, and also has integration with various state, municipal, and international data portals, allowing users to access a catalog of data beyond just the U.S. federal government's repository.

Among these integrated portals are those of the 46 states that have their own open data portals. Of these 46 states, only sixteen have legislation on the books that have established a requirement for the publishing of data, while four states have had governors issue executive orders to establish open data initiatives (National Council for State Legislatures 2022). Granularity of data is critical, therefore meaning that as much as the increasing breadth and depth of Data.gov is an achievement, further investment into state data portals and data collection is critical, in addition to taking that one step further by building open data infrastructure on the municipal level.

1.3. Municipal Open Data

The first municipal open data initiatives were established from 2009 through 2012, with most early-adopting cities such as Chicago, Seattle, etc., launching their portals between 2011 and 2012. Most cities began by releasing spatial data that they already had available through their planning departments, and any commonly requested datasets (Johnson 2016). In the years since, open data portals in these early adopting cities have grown to make a variety of data available to their citizens, and

in the most recent example of cities releasing relevant data, many cities have released COVID-19-related data on their portals.

1.4. Municipal Data Quality

With the growth of municipal open data initiatives comes an increase in opportunity for cities to engage with their citizens, promote transparency, and forward economic innovation. However, actually doing so requires the release of data in a manner that is actually useful to a city's constituents—and a good shorthand for understanding whether or not the data being released is usable is by evaluating its quality.

Cities that have open data initiatives are wide and varied in their characteristics, and the data that they collect and release is ultimately a function of their community and the clarity that city government has regarding the process to release this data.

Which leads us to the question: how do institutional factors affect data quality in municipal data portals? By evaluating a set of 32 municipal data portals on their adherence to OGD standards and determining their relationship with institutional factors unique to each city, this study will be able to determine what kinds of cities are currently leading the pack in the quality of data that they are providing to their citizens.

2. Literature Review

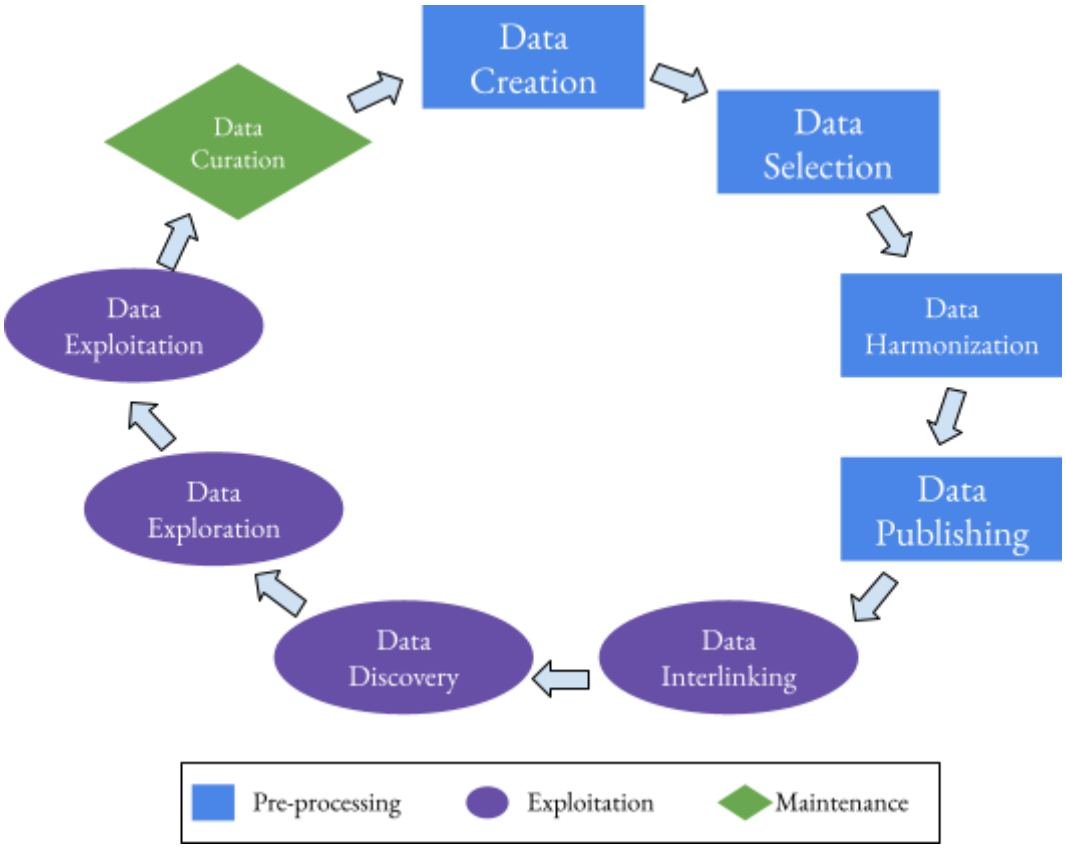
2.1. Data Governance

When discussing recent trends towards open data governance, researchers often cite three motivations for why governments have begun to make more data available to the public: transparency in an effort to hold themselves more accountable or at least appear as though they are doing so, cost-effectiveness in the development of solutions to public problems, and increasing the level of citizen participation in the government (Johnson & Robinson 2014; Attard et al. 2015; Johnson 2016; McNutt et al. 2016).

However, surveying municipal officers regarding why they believe that maintaining open data is important has revealed that in reality, the economic facet of open data governance is not in fact as prominent a motivating factor as early literature assumed (Johnson 2016). Municipal officers view the main goal of open data as "transparency and service to citizens and businesses", and often make decisions in respect to achieving these goals, but at the same time lack evaluatory mechanisms to determine the degree to which they fulfill those aims (Johnson 2016). As such, the question then becomes how effective these governments are in achieving those stated goals, which requires an understanding of the current process involved in data being released by the government and how useful the data they release is.

Attard et al. (2015), outlines three distinct phases of the open government data life cycle: pre-processing, exploitation, and maintenance. Within these phases are steps, as follows:

Figure 1: The Government Data Life-Cycle, adapted from Attard et al. (2015)



The two steps that are of large interest to the focus of this paper are data publishing and data curation, as these are acts largely engaged in by the agencies and governments that are publishing the data. Attard et al., defines the two steps as follows:

> "*Data Publishing* - This is the actual act of opening up the data by publishing it on government portals.

Data Curation - While not necessarily occurring at a fixed stage, data curation is vital in ensuring the published data is sustainable. This involves a number of processes, including updating stale data, data and metadata enrichment, data cleansing, etc."

While many governments use OGD principles as the framework for their open data policies, the degree to which these principles are followed vary widely, resulting in data publishing and curation that are not standardized. This lack of standardization in the way data is released poses a variety of problems, including the creation of a gap between what the government is willing to make available and what data constituents actually want or need access to. This is an issue that is seen around the world amongst national data portals and within the U.S. amongst municipal data portals (Sayogo 2014, Zhu 2019). Many governments release neutral data for the sake of demonstrating some level of baseline commitment to OGD, but this data does not have much use when it comes to demonstrating transparency, fostering economic innovation, or encouraging citizen participation (Nahon 2015).

Generally speaking, despite fifteen years having passed since the publication of the eight principles for OGD, governments still have much to catch up on in their adherence to the principles. And since most municipal governments and data managers don't have their own evaluatory mechanisms in place for determining the quality, accessibility, and usefulness of the data that they are releasing to their constituents, it becomes critical for researchers to consider how they can qualify the data initiatives of these governments so that areas of improvement can be determined.

2.2. Evaluating Data Governance

Despite early determination of municipal data initiatives being the best forum for the investigation of open data governance, most studies in the field focus on national governments (Kassen, 2013; Conradie and Choenni 2014). Nahon (2015) suggests that the reason for this is due to the number of differences that exist across municipalities not only in their treatment of open data, but in the broad variety of factors that affect their application of open data policies, ranging from socioeconomic differences to the wide range of municipal services that a city may be collecting data on that are not necessarily comparable to any other city's service structure.

However, pulling from studies on national governments can provide useful information regarding the kind of pitfalls that all forms of government may generally be susceptible to in releasing open data, while also providing methods that can be built upon and adapted as tools to evaluate local data initiatives.

2.2.1. National Governments

Studies approach the evaluation of various national government initiatives for open data around the world from different directions (Sayogo, 2014; Bogdanović et al., 2014; Veljković et al., 2014; Nugroho et al., 2015; Milić et al., 2022). While some create benchmarking methods, others utilize the eight principles as their tool for determining adherence, while others yet focus on open data policy more so than the data itself.

Sayogo (2014) views evaluating data governance as a matter of adherence to the eight principles, much as this paper does. In evaluating the portals of 35 national governments, Sayogo was an early identifier of reasons for concern regarding the ability of governments to provide access to the datasets that their citizens wanted access to, and for developing the kind of internal infrastructure necessary to make data curation and publishing part of how government agencies collected information.

Nugroho et al. (2015) compares national open data policies and echoes these concerns regarding internal infrastructure for quality data release, and also notes that some of this can be alleviated by assigning the responsibility of implementing open data policies to specific task forces or agencies that are created for the purpose of coordinating the release of open data. And though the study focuses on the effectiveness of national government OGD initiatives, Nugroho also highlights how encouraging open data work at the local-level can increase the quality and granularity of data that is made available on the national level.

When it comes to metrics for evaluating the data and data catalog itself, Veljković (2014) created the OpenGovB method for benchmarking national open data initiatives using indicators that evaluate factors such as basic dataset availability, data openness based on the eight principles, data transparency, and user-involvement along the levels of participation and collaboration constituents engage in with their government. Bogdanović et al. (2014) focuses on the data openness indicator of the OpenGovB model, while Milić et al. (2022) focus on the transparency indicator. In focusing on these specific indicators, the two papers are able to demonstrates that the OpenGovB metric is a valid

way of automate the assessment process, however, at the same time the method leaves a few questions, as it doesn't take into account factors such as the size of data categories available within an open data catalog, or how long the portal has been established for—as such, it is difficult to draw a true comparison between countries using the OpenGovB metric, though it is quite useful for a country to understand its own year-over-year progress. Additionally, the metric is not one that can be directly adapted to city governments, as some of its indicators, such as data transparency, utilize existing national metrics for estimating factors, such as the Corruptions Perception Index (CPI) and Control of Corruption (CC), which do not have analogous municipal metrics. However, as will be seen later in this paper, some elements of the OpenGovB metric are useful as starting points for a municipal methodology to evaluate data availability and quality.

2.2.2. Government Departments and Agencies

Other papers approach the topic of data governance from the perspective of different government departments and agencies, both at the national and local levels (Vetrò et al., 2016; Zuiderwijck et al. 2014; Kučera et al. 2013). Most of these studies evaluate European governments but are useful, as data release done on the level of departments and agencies can closely mimic the decentralized manner in which data is often aggregated and released on municipal data portals (Conradie and Choenni 2014).

Vetrò et al., 2016 conducted a look into Italian data portals at both the national and municipal level, with the intention of comparing what it defined as a more "centralized" portal with a

"decentralized" portal. In doing so, they were able to identify that the national, more-centralized portal, was better suited to releasing higher-quality data, due to its ability to aggregate more sources and implement a far more extensive quality evaluation process.

While the specific issues found with the portals' data quality are not highly generalizable and are quite specific to the two data portals in question, the paper provides a few possible methods for dealing with some of the issues that it found. Some of these include the ability to version datasets, or view past iterations of it to better understand the timeliness of the data that is available; introducing feedback mechanisms to gain a user perspective on the accuracy of data, as opposed to just relying on internal quality control; and utilizing external tools to take on the work of data cleaning and the integration of metadata.

Other than the feedback mechanism for accuracy, the other two recommendations can be related to underlying OGD principles of timeliness, permanence, and completeness. In determining whether or not a city's data complies with these OGD principles, it would therefore be useful to keep in mind what the ideal results would be if these potential issues were taken into account, e.g. seeing that cities have information on when the data was last updated, continue to have previous years of data available, and provide detailed metadata alongside clean data.

Mitigating these sorts of the issues from the get-go is easier when there is foundational policy to support an OGD initiative—not necessarily just legislative or executive work, but a set of guidelines, best practices, and goals for what an OGD initiative is meant to accomplish. Strong OGD policy serves as a useful starting point for building out the infrastructures and methods necessary for effectively

releasing data (Zuiderwijk et al. 2014). If a government has properly outlined the goals of its OGD initiatives and concrete steps for achieving those goals, such as through evaluating for data release, useability and/or usage, or other mechanisms, it become much easier for cooperating agencies to comply with and implement policies that allow for the achievement of those goals. Zuiderwijk (2014) found that organizations often approached their open data policies in one of two ways: (1) genuine interest in implementing open data policies to become more transparent, participatory, and accountable, or (2) an obligation that they have to fulfill for the sake of appearing more transparent, participatory, and accountable, resulting in a higher level of wariness regarding legal liability and other risks, and therefore less effective open data initiatives.

The second approach is often a result of organizations not critically considering their own individual contexts in implementing open data policies, resulting in policies that may be ill-suited to the needs of the organization and poor at adapting to new needs for open data. Genuinely open organizations benefit from an institutionalized culture of opening data that is created by focusing on the potential impact of a successful open data initiative, as opposed to just publishing and legislation. As such, organizations that adhere to the first approach are more likely to be publishing better quality data on a consistent basis, as opposed to simply providing lip-service (Zuiderwijk 2014).

Kučera et al. (2013) investigates the quality of OGD catalog records and emphasizes that it is important not just that the datasets available in an OGD catalog are high quality, but that the information about the dataset itself must be helpful in informing the user what the dataset contains, where it came from, and how up to date it is—essentially advocating for more robust metadata. In

doing so, the paper echoes the conclusions of Vetrò et al. (2016) and provides some suggestions for processes that neatly align with Zuderwijk et al. (2014)'s conclusions regarding the benefits of strong, tailored OGD policy.

In particular, Kučera et al. (2013) notes a distinction between *data-driven techniques* which "directly modify the values of data and thus they are used to improve the quality of existing data" versus *process-driven techniques*, which "aim at redesign of the data creation and modification processes in order to identify and eliminate the root cause of quality issues." Implementation of solely the latter form of technique will likely result in issues that are replicated across datasets, and therefore make themselves evident as a consistent issue within an OGD portal, whereas utilizing the latter form as well would eliminate some of this inconsistency and result in a stronger implementation of OGD policy that ultimately leads to a higher quality data catalog.

2.2.3. Municipal Governments in the U.S.

Comparative studies for evaluating data portals have included the development of a metric for evaluating the commitment of municipal governments to OGD, and evaluation of local data portals from a user-perspective by utilizing the eight principles (Nahon 2015; Zhu and Freeman 2019).

Cities in the U.S. vary in their degree of commitment to OGD initiatives due to a variety of external factors. According to Nahon (2015)'s OGD Heartbeat metric, cities can be categorized into three degrees of commitment: high, where OGD commitment is integrated into existing processes; low or inconsistent, where OGD might be an afterthought or external to most agencies' information

release; or barely any commitment, where no active work has been done to release data or adhere to OGD principles. To categorize U.S. cities in this manner, Nahon (2015) utilizes four factors: rhythm, coverage, feedback, and categorization. Rhythm refers to the number of datasets uploaded and the regularity with which datasets are uploaded, coverage refers to the breadth of datasets made available, feedback refers to the ability to contact municipal officials about the data, and categorization refers to metadata and other markers that can be used to identify and describe datasets.

While it is shown that the OGD Heartbeat is a useful and valid metric for understanding a city's commitment to OGD, the metrics utilized do not really consider the importance of data quality as a component of its own. Without considering data quality, it is difficult to really understand how open the data being provided is, even if it is being regularly uploaded with helpful metadata across a multiplicity of topics, and municipal officers are responsive to queries. Cities can upload large amounts of data without necessarily adhering to open data standards, thereby meaning that they wouldn't actually be making available high quality data on the topics that citizens want access to.

Zhu and Freeman (2019) evaluated 34 municipal data portals utilizing a User Interaction Framework that was developed by considering six dimensions of whether a given user could access, trust, understand, engage-integrate, and participate in the use of open data. Parts of these six dimensions are drawn from the eight original OGD principles, while other elements are drawn from previous literature and information systems theory. The Zhu and Freeman framework for analysis is useful in determining the effectiveness of open data portals and demonstrates that there is a large

degree of variability in the types of datasets made available by cities, but does not do much in depth analysis regarding what institutional factors may result in this variability.

2.3. Factors Affecting Data Governance

Various factors have been identified in previous literature as possible predictors of data portal or dataset quality. Some of these factors are echoes of those that have already been identified in literature that focuses on national governments, whereas others are specific to municipalities.

Thorsby et al. (2016) developed two indices, the Open Government Data Portal Index (OGDPI) to evaluate portal features, and the Dataset Content Index (DCI) to evaluate the content areas available within a given data portal, in order to evaluate 37 data portals sourced from Data.gov. The paper then goes on to determine whether or not various city characteristics are predictors of either index. Based on previous literature, the identified predictors were: population, level of education in the city, type of government, degree of innovation, age of the portal, and participation in a regional data consortium. Of these predictors, only population and participation in a regional data consortium had an effect on the index scores, with population having a significant positive effect—such that the higher the population, the higher the index scores—while participation in regional data consortiums had a negative effect.

Thorsby et al. (2016) specifically decided to not use the content areas identified by the Open Knowledge Foundation's City Census for its DCI, electing to use broader categories to determine the breadth of the data portal. This decision makes sense, as the focus of the paper was to generate an

understanding of the quality of data portals, as opposed to the quality of the data within the portals, which is the aim of this paper.

Identification of population as a significant factor in the quality of a data portal is also echoed by Zhu and Freeman (2019). As mentioned previously, this paper utilized a User Interaction Framework to access the usability of municipal data portals and developed this framework by integrating the eight OGD principles into its criteria. Along with acknowledging the role of population size, the paper also identified that portal platform may be a critical factor in the useability of open data portals, and suggested that further research was necessary along the lines of socioeconomic and political drivers in order to fully understand this variability. Nahon (2015) originally identified this need for further research into socioeconomic and political factors as necessary for gaining understanding of what causes specific types of OGD behavior by cities.

The evaluation of national data portals conducted by Máchová and Lněnička (2017) similarly identified portal platforms as an important indicator of data portal quality, specifically noting that of the national portals looked at, those that utilized CKAN as their platform performed better.

Young (2020)'s investigation of municipal data portals from the perspective of public administration is most like this paper in the focus that it places on institutional factors over the evaluatory mechanism it utilizes. This paper evaluated data portal content on a departmental level, however, as opposed to a city-government level. The response variable in this paper was the total number of datasets released by each department in a municipal government. The predictors considered were specific to this department-level approach, in addition to city characteristics such as city size

(number of full time employees for every 1,000 residents), population, population density, median income, percent unemployed, existence of a CDO, and demographics, among other variables.

While Young's paper is a useful foundation for choosing institutional city characteristics to investigate, the lack of nuance to the response variable (total number of dataset files) means that it is not possible to really understand how well a city is adhering to open data principles—as mentioned before, cities can always release lots of datasets without necessarily releasing high quality OGD. The paper ultimately determined that department-level characteristics are the strongest predictors of data availability, compared to city-wide characteristics, and specifically identified variables such as "organization size, managerial capacity, whether the data are useful for economic development, and demand-side pressures from technologically adept residents" as having relationships with the number of total files released by a city's departments.

3. Methodology

This study examines 36 cities with populations over 100,000 across the U.S.. Eighteen cities were randomly chosen from Data.gov's list of municipalities that have been integrated into the Data.gov catalog, while the other eighteen were randomly chosen from a list of municipalities with data portals that was put together by Forbes in 2018. While it is not anticipated that the Data.gov cities will be in any way significantly different from those sourced from the Forbes list, how the portal was identified has nonetheless been included as a variable in the data.

Data will be analyzed using an Ordinary Least Squares Regression (OLS). I will be using indicators that I have developed based on the OpenGovB method for benchmarking data quality to evaluate each city's data portal (Veljković et al. 2014)—the weighted average value of these indicators will make up the outcome variable. The indicator will then be compared across various institutional factors that have been collected on each city.

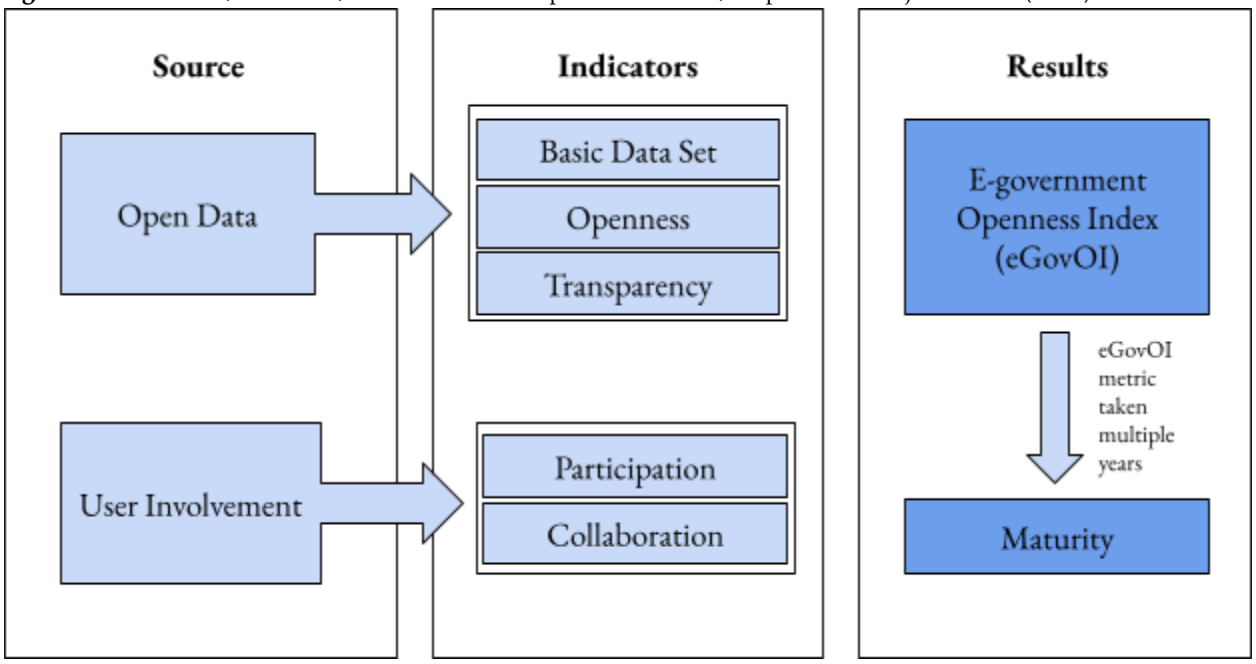
Table 1: Cities, portal source

Data.gov			**Forbes**		
Albuquerque	Las Vegas	Portland	Boston	Minneapolis	San Jose
Austin	Lexington	Providence	Charlotte	Nashville	Syracuse
Baltimore	Louisville	Raleigh	Cincinnati	Newark	Tampa
Chicago	Madison	San Francisco	Dallas	Pittsburgh	Tucson
Denver	New York City	Scottsdale	Detroit	Phoenix	Tulsa
Houston	Philadelphia	Seattle	Milwaukee	San Diego	Washington D.C

3.1. OpenGovB

OpenGovB is a method utilized to evaluate and compare national data portals, therefore meaning that some of the indicators in the method utilize metrics and factors that are not applicable to the specificity of municipal open data portals. As such, the method for this paper will be drawing from the OpenGovB rubric to rate data openness, but will be utilizing an independently developed method for determining which datasets are evaluated and what conclusions can be drawn from that analysis.

Figure 2: Full sources, indicators, and results of the OpenGovB Model, adapted from Veljković et al. (2014)



3.2. Determining Data Principle Score and Average Data Quality

The Data Principle Score or DPS metric will be determined by averaging all of a city's scores across each of the ten OGD principles for the Core Datasets (CDS) that it has, generating average scores for each principle.

The Average Data Quality or ADQ metric will be determined by adding up all of a city's Data Openness (DO) scores and dividing it by the number of CDS that it has.

The calculation of the CDS and DO will be described further in the following sections.

3.2.1. Calculating the Core Dataset Indicator

The Core Dataset indicator (CDS) is calculated through the determination of whether certain types of datasets are or are not present within a data portal. This study will be looking for 20 core different datasets in every municipal portal: Budget, Crime Reports, Construction Permits, Parcels, Zoning, Service Requests, Spending, Code Violations, Employee Salaries, Business Listings, Traffic Crashes, Restaurant Inspections, Property Assessment, Public Facilities, Emergency Calls, Procurement Contracts, Police Use-of-Force, Website Analytics, Lobbyist Activity, Property Transfers. These datasets have been identified by the Open Data Census—a collaboration between the Open Knowledge Project and the Sunlight Foundation—as key datasets for determining how open a city's municipal data portal is. The expectation is that in a data portal that provides a comprehensive set of datasets, all of these categories will be available for exploration.

While the Open Data Census collects survey responses provided by users of the site regarding whether a city's data portal has a particular dataset and the degree to which that dataset is open, this study only used the Census as a starting point. Data regarding the existence of a dataset, its openness, and its quality were all collected by going to each city's data portal and conducting keyword searches for the core datasets.

Each category will be evaluated as either 0 or 1, with 0 meaning that there is no dataset available for the category, while 1 means that there is at least 1 dataset published under the category. From there, the CDS indicator will be calculated by adding up the number of categories present in the data portal (n) out of the total number of categories that the portal is being checked for (N = 20), resulting in an

indicator score within the range of [0, 20]. A score of 0 means that none of the categories are available, while a score of 20 means all of the categories of data are available.

3.2.2. Calculating the Data Openness

While the original OpenGovB method for calculating Data Openness (DO) only considers the original eight OGD principles, this study will be modifying the indicator to include the additional two developed by the Sunlight Foundation. As such, the DO calculated in this study will be the sum of scores for adherence to the ten principles that were described earlier in the paper.

The DO will be calculated by evaluating each core dataset that the municipal data portal contains. Evaluation will occur using the ten OGD criteria, as follows (while Veljković et al. have defined the criteria in their own way, this study has opted to utilize the definitions provided by the Sunlight Foundation to establish the criteria).

Completeness: Completeness of a dataset will be characterized by four factors: availability of a data description, the ability to download the data, whether the data is in a machine-readable format, and whether the dataset is linked to other datasets in the catalog.

Primary: Data is published in raw formats, even if the website presents it in charts or graphs. Data that is only published in charts or analyzed formats are not primary, as the user cannot extract the raw data for their own analysis.

Timely: Datasets contain information regarding the time period the data spans, how often the data is updated, and when the last update was.

Accessible: Datasets are accessible without having to go to a physical office or submit a form before gaining access. Datasets can be downloaded in bulk or accessed through an API.

Machine Processable: The datasets are available in machine processable formats, with there being three-levels of processability: not processable (e.g. PDF), structured formats that can be processed (e.g. CSV), and structured formats that have the capability to contain meta-description and/or semantics (e.g. XML, RDF).

Non-Discriminatory: Datasets are available without registering for access.

Non-Proprietary: Datasets are released in formats that do not require specific, commercial software to open them, such as XLS, which requires Microsoft Excel. Instead the data is released in formats such as CSV, XML, or RDF, which can be opened by non-commercial software.

License-Free: Datasets are published under an open license.

Permanence: Datasets from as far back as when the portal was first launched, or the dataset was first made available are still accessible..

Usage Costs: There are no fees associated with accessing the datasets.

Table 2: DO Scoring Rubric, adapted from Veljković et al. (2014)

					Max Score
Complete	Description Available 0.25	Downloadable 0.25	Machine-readable 0.25	Linked Data 0.25	1
Primary	Are datasets available as raw data?				1
Timely	Time Period 0.3		Update Frequency 0.4	Last Update 0.3	1
Accessible	Is the dataset available to download in bulk or through an API?				1
Machine-processable	PDF, XLS, etc. 0.2		CSV 0.5	XML, RDF 1	1
Non-discriminatory	Is the data available without registering for access?				1
Non-Proprietary	Is data available in formats that do not require commercial software?				1
License-free	Is data available under an open license?				1
Permanence	Is older data from data portal launch/before launch still available?				1
Usage Costs	Is data available for no cost/fees?				1
				DO	10/10

3.3. Institutional Factors

The second part of analysis involves identifying institutional factors (independent variables), which will be compared amongst different cities to determine their impact on the overall DQ of a city's data portal.

Year Established: The year the data portal was established provides us important information regarding how long the city has had to develop and refine its catalog, in both depth and breadth of datasets available. It is expected that younger portals will have fewer datasets or lower dataset quality due to their infancy, however it is also possible that older portals were established with an initial rush to publish datasets that later died down and resulted in lower-quality data.

Population: The current population of each city will be considered. Previous literature has shown that cities with larger populations are more likely to have more datasets in their data portals (Thorsby et al. 2016; Zhu and Freeman 2019). We will see if this holds true for having higher data quality as well.

Percent White: Cities with more heterogeneous populations are more likely to record and release more data, and also make more of an effort to share data with third-parties. This has been traced to the idea that cities with more heterogeneous populations require data to be applied in more decentralized ways, and therefore benefit from generating large amounts of data that can be outsourced to third-parties to put that data into practice for policies or programs. Additionally, cities with more heterogeneous populations have a greater stake in demonstrating overall transparency by releasing data, as they have multiple, distinct populations to serve (Fusi & Feeny, 2020).

Gini Index: The Gini coefficient measures income inequality, with 1 = total equality and 0 = total inequality. It has been shown that cities with lower median incomes actually have more datasets made available, and Young (2020) suggests that this may be because larger cities have a wider distribution of income levels across a larger population. By using the Gini Index, we will be able to

actually test the effect of that income distribution on the quality of datasets available in a city's portal, without inadvertently conflating this socioeconomic indicator with population differences.

Percent College Graduates: This metric is meant to help determine how much of the population could potentially have the knowledge to explore and utilize open data made available by a city's data portal, which may determine the number of datasets made available, how complete those datasets are, and how suited they are to manipulation by users (Thorsby et al. 2016).

Percent Unemployed: This has been shown to be negatively correlated with the number of datasets made available, similarly to median income, and therefore is important to investigate in regards to data quality as well. Since we will be testing the Gini Index in place of median income, this variable will allow us further understanding of how the socioeconomic environment affects data quality (Young 2020).

City Government Type: Most city governments operate in either Mayor-Council or Council-Manager formats. Mayor-Council formats are generally driven by executive agenda-setting power, whereas in Council-Manager formats, the legislative has the primary agenda-setting power (National League of Cities 2022). While the Council-Manager format is the most popular type of city government across the U.S. at large, amongst bigger cities, the Mayor-Council format prevails. The third, less common form is a commission format, wherein commissioners are elected to a small board which has legislative and executive functions. Knowing which type of government a city has may provide insight into whether the establishment of municipal data portals is a more executive- or legislative-driven effort in these larger cities (Thorsby et al. 2016).

Political-Lean: Most city governments are nonpartisan, however, no city operates in isolation from broader trends in politics. It is possible that looking at the margin between the percent of votes cast for the Democratic and Republican candidate in the 2020 election may provide evidence of a relationship between political lean and commitment to open data, as it could serve as some proxy for the constituents' willingness and/or interest to have this data collected.

Policy: Whether the city's open data initiative has (or has not been) codified into policy may be an indicator of how strictly or loosely the city commits to the initiative and, consequently, has high-quality data (Kučera et al. 2013; Zuiderwijk 2014).

Type of Policy: How the city's OGD initiative was codified into policy, and by what level of government, e.g. executive action, legislation, internal policy, etc.

Chief Data Officer: In establishing OGD initiatives and data portals, some cities created the position of Chief Data Officer (CDO), who is responsible for coordinating the aggregation and release of data across city departments. In cities where CDOs are not established, it is often the case that existing departments, such as Information & Technology, take on the bulk of OGD work, and data release may also occur in a decentralized fashion, with each department being responsible for its own aggregation and release (Young 2020).

Portal Platform: There are four major software platforms utilized by city governments to host their data portals: ArcGIS, Socrata, CKAN, and DKAN. Each of these platforms are slightly different in the way that they present information, and the kinds of information they are most optimized for. For example, use of ArcGIS usually builds out of existing GIS infrastructure that a city has for zoning,

planning, and/or parcels—which may result in a data catalog that is skewed towards spatial data (Máchová and Lněnička 2017; Zhu and Freeman 2019).

3.4. Analysis & Limitations

Data will be analyzed through a series of Ordinary Least Squares (OLS) regressions to determine the relationship between the institutional factors and the DPS and the ADQ. I will be using a 95% significance level to identify any relationship between the variables. The biggest limitation of this approach is the relatively small sample size compared to the number of variables being tested.

4. Data Collection

Data was collected across multiple sources and initially input into Microsoft Excel. Before collection officially began, each data portal of the cities that had been selected was swept to ensure that I would be able to conduct my research. In this sweep, four of the 36 cities were removed from the set. Portland was removed due to low useability and difficulty in ascertaining whether the site had been recently updated, while Tampa and Tulsa were removed as the portals were actually ArcGIS GeoHubs which are geared towards hosting spatial data, therefore meaning they had no CDS and could not be assessed for data quality. Lexington was removed due to misidentification as a city, when in fact it is the Lexington-Fayetteville county in Kentucky.

4.1. Investigating Data Portals

To collect data for CDS and DO, I went to the website of each municipal data portal. The websites visited are as follows:

Table 3: Cities, portal websites

City	Website	City	Website
Albuquerque	http://www.cabq.gov/abq-data/	Nashville	https://data.nashville.gov/
Austin	http://data.austintexas.gov/	New York City	http://www.nyc.gov/data/
Baltimore	http://data.baltimorecity.gov/	Newark	https://data.ci.newark.nj.us/
Boston	https://data.cityofboston.gov/	Philadelphia	http://www.opendataphilly.org/
Charlotte	https://data.charlottenc.gov/	Phoenix	https://www.phoenixopendata.com
Chicago	http://data.cityofchicago.org/	Pittsburgh	https://data.wprdc.org/dataset?organization=city-of-pittsburgh
Cincinnati	https://data.cincinnati-oh.gov	Providence	https://data.providenceri.gov/
Dallas	https://www.dallasopendata.com	Raleigh	http://www.raleighnc.gov/open/
Denver	http://data.denvergov.org/	San Diego	https://data.sandiego.gov/
Detroit	https://data.detroitmi.gov	San Francisco	http://www.datasf.org/
Houston	http://data.houstontx.gov/	San Jose	https://data.sanjoseca.gov
Las Vegas	https://opendataportal-lasvegas.opendata.arcgis.com/datasets	Scottsdale	http://data.scottsdaleaz.gov/
Louisville	https://data.louisvilleky.gov/	Seattle	http://data.seattle.gov/
Madison	https://data.cityofmadison.com/	Syracuse	https://syrgov.net
Milwaukee	https://data.milwaukee.gov/	Tucson	https://gisdata/tucsonaz.gov
Minneapolis	https://opendata.minneapolismn.gov	Washington D.C.	https://opendata.dc.gov

4.1.1. Core Datasets (CDS)

Collecting information on the datasets in each portal involved conducting searches for each one within each city's catalog. To do this, I compiled a list of possible search terms before beginning the collection. As collection proceeded, terms were added to this list, resulting in the final set of search terms that are listed in the table below.

Table 4: Dataset types and search terms

Dataset	Search Terms
Budget	"Budget"
Crime Report	"Crime", "Incident", "Police"
Construction Permits	"Construction", "Permit", "Building"
Parcels	"Parcel"
Zoning	"Zoning"
Service Requests	"Service Requests", "311"
Spending	"Spending", "Expenditures", "Checkbook"
Code Violations	"Code Violation", "Inspection", "Building"
Employee Salaries	"Salary", "Compensation", "Wages"
Business Listings	"Business Listing", "Business Registration", "Business Licenses"
Traffic Crashes	"Traffic", "Crashes", "Accidents"
Restaurant Inspections	"Restaurant Inspection", "Food"
Property Assessment	"Property Assessment", "Property Value", "Property Tax"
Public Facilities	"Public Facilities", "City Property", "Government Property"
Emergency Calls	"Emergency Calls", "911", "Police"
Procurement Contracts	"Procurement Contract", "Vendor", "Bid"
Police Use-of-Force	"Police Use of Force", "Use of Force", "Right to Retaliation"

Table 4 (cont.): Dataset types and search terms

Website Analytics	"Website Analytics", "Web", "Google"
Lobbyist Activity	"Lobbyist", "Lobby"
Property Transfer	"Property Transfer", "Property Sale", "Real Estate"

If a data portal linked to a dataset external to the portal, I did not consider the portal as having that particular CDS, as this study is specifically looking at data available *within* municipal data portals, as opposed to data made available by municipalities in general.

During collection of CDS data, information on the explanatory variables of *portal host* and *total datasets* were also collected.

Table 5: Portal platform host

ArcGIS (11)		**Socrata (8)**	**CKAN (8)**	**DKAN (1)**	**Other (4)**
Baltimore	Raleigh	Austin	Boston	Louisville	Albuquerque
Charlotte	Scottsdale	Chicago	Houston		Denver
Detroit	Syracuse	Cincinnati	Milwaukee		San Diego
Las Vegas	Tucson	Dallas	Newark		
Madison	Washington D.C	Nashville	Pittsburgh		
Minneapolis		New York City	Philadelphia		
		Providence	Phoenix		
		San Francisco	San Jose		
		Seattle			

4.1.2. Data Openness (DO)

Every component of the DO rubric was scored as either 1 or 0. For Complete and Timely, this meant that the components of the principles were plugged into the formula for scoring after the fact.

Complete; Description: If the page for the dataset either had metadata describing what the dataset contained, the names of columns, and any other pertinent information, this component was scored as a 1. On some portals, this metadata was available directly on the web page itself, on others, it was available in the form of an additional file that could be downloaded. For the latter method of providing metadata, some of these files were in proprietary formats, however, it was not possible to incorporate that information into the scoring. Future methods of evaluation should incorporate this.

Complete; Downloadable: If the dataset was downloadable, this component was scored as a 1.

Complete; Machine-Readable: If the dataset was in a machine-readable format, this component was scored as a 1. It can and should be viewed in relation to the scoring for the Machine-Readable principle.

Complete; Linked: If the page for the dataset provided information on what other data could be analyzed alongside the given dataset, this component was scored as a 1. This component was also scored as a 1 if clicking on a dataset link would bring up the datasets from previous years as well. Example, clicking on the "Annual Budget" link and being able to access all budget datasets from previous years under that same link.

Primary: If the page allowed for the download of the raw data, this principle was scored as a 1. There is the occasional, low chance that a city may provide visualized data without providing access to the underlying information, in which case, this component would be scored as a 0.

Timely; Period: If the page included information on how long the dataset spanned, this component was scored as a 1.

Timely; Update Frequency: If the page included information on how often the dataset was updated, this component was scored as a 1.

Timely; Last Update: If the page included information on when the dataset was last updated, this component was scored as a 1. Future methods of scoring should consider incorporating a method of keeping track of whether the data has been recently updated, however this may be difficult if information regarding update frequency is regularly not provided.

Accessible: If full access to the data was available through the website without having to contact a government office or submit a request, this principle was scored as a 1.

Machine-Readable: This principle was scored according to the given rating system in the rubric for each type of data format. For some portal platforms, such as Socrata, even if the original dataset upload was in a proprietary format, users are able to download converted non-proprietary formats from the platform. In these cases, the principle was scored as a 1 for availability in a non-proprietary format.

Non-Discriminatory: If the page was accessible and the data was downloadable without any form of registration, this principle was scored as a 1.

Non-Proprietary: If the machine-readable format that the data was available in did not require proprietary software, this principle scored a 1.

License-Free: If the page provided information that the data was publicly licensed or under a Creative Commons attribution, this principle scored a 1. Portals where the license was not specified on

the page, or where the specific dataset's license did not line up with the usage terms that were outlined elsewhere in the portal were scored as a 0.

Permanence: If datasets from previous years were available, this principle scored a 1.

Usage Cost: If the datasets were available free of cost, this principle scored a 1.

4.2. Explanatory Variables

After recording all data on the data portals, values for the explanatory variables were then collected.

Current population and *percent white* were collected from the U.S. Census Bureau's 2019 American Community Survey's 5-Year Estimates. Of the 32 cities, Houston did not have demographic data available, but did have population data available through the Census Bureau's Quick Facts page.

Bachelor's degree attainment and *Gini Index* were also collected using the Census Bureau's Quick Facts tables and are derived from 2015-2019 data. Louisville did not have data available.

Unemployment rate was found on the U.S. Labor Bureau website and collected from the 2017 annual averages for metropolitan areas in the U.S. As such, the *actual* unemployment rate of individual cities may be higher or lower than that of the metropolitan area that they are in. However, this was considered permissible as the unemployment rate is being considered alongside median income, bachelor's degree attainment, and the Gini Index to understand the socioeconomic characteristics of the city.

Year established, *type of city government*, the *existence of a CDO* were all collected from city websites or documents (press releases, PDFs, etc.). In some cases, data like year established was collected from news articles that referenced the creation of the portal, as information could not be found through official government materials.

Table 6: Cities, year portal established

Year	City
2009	San Francisco
2010	Seattle
2011	Austin, Baltimore, Louisville, Philadelphia
2012	Albuquerque, Boston, Chicago, Denver, Madison, New York City,
2013	Las Vegas, Lexington, Providence, Raleigh
2014	Charlotte, Houston, Minneapolis, Nashville
2015	Cincinnati, Detroit, Newark, Pittsburgh, Tucson
2016	Dallas, San Diego, San Jose, Scottsdale
2017	Phoenix, Syracuse
2018	D.C., Milwaukee

Table 7: Government type

Mayor-Council (24)				Council-Manager (9)	
Albuquerque	Denver	Milwaukee	Philadelphia	Austin	San Jose
Baltimore	Detroit	Minneapolis	Providence	Charlotte	Scottsdale
Boston	Houston	Nashville	San Diego	Dallas	Tucson
Chicago	Lexington	Newark	San Francisco	Las Vegas	
Cincinnati	Louisville	New York City	Seattle	Phoenix	
D.C.	Madison	Pittsburgh	Syracuse	Raleigh	

Table 8: Chief Data Officer (CDO)

Has CDO (18)			**Does Not Have CDO (15)**		
Baltimore	Denver	Pittsburgh	Albuquerque	Lexington	Phoenix
Boston	Detroit	San Diego	Austin	Madison	Providence
Charlotte	Louisville	San Francisco	Dallas	Milwaukee	Raleigh
Chicago	Minneapolis	San Jose	Houston	Newark	Scottsdale
Cincinnati	Nashville	Seattle	Las Vegas	Philadelphia	Tucson
D.C.	New York City	Syracuse			

Determining the *existence and type of policy* that outwardly established OGD initiatives in each city was done by first using the Open Data Policy Hub's (ODPH) list of existing policies, and then by searching city websites for the cities that did not have policies listed. Some cities, like Dallas, did not establish their data portals through a policy but later created a team to evaluate the portal (Dallas Open Data). Other cities followed up initial policies with further directives, or executive or legislative action—in these cases, the initial policy was the one that was used for this categorization. Executive policy covers executive orders and directives. Legislative policy covers resolutions, ordinances, and legislation. Internal policy covers administrative actions and internal department policies. If no policy could be found online through the ODPH, or searching through the city website, it was determined that the city had no outward-facing OGD policy.

Table 9: Type of policy

Executive (10)		Legislative (14)		Internal (4)	None (5)
Boston	Louisville	Austin	New York City	Charlotte	Albuquerque
Chicago	Nashville	Baltimore	Pittsburgh	Houston	Dallas
D.C.	Philadelphia	Cincinnati	Providence	Phoenix	Denver
Detroit	Seattle	Lexington	Raleigh	Scottsdale	Newark
Las Vegas	Syracuse	Madison	San Diego		Tucson
		Milwaukee	San Francisco		
		Minneapolis	San Jose		

(Open Data Policy Hub)

4.3. Limitations

The biggest limitation of the data collection is that it could not be conducted as an aggregation of multiple ratings. As this study was conducted independently and data was collected over the period of a week, there is quite a bit of space for human judgment error in where there were slight variations in how each municipal data portal was rated. Were this study to be conducted again, it would be ideal for it to be done with a team of multiple researchers who start at different points in the list of cities and each determine their own ratings before ultimately reconciling those ratings to create the final set of DO scores.

5. Analysis

Analysis of the data was two-fold—first the DPS and ADQ were calculated, and then regressions were run to determine the relationship between the DPS and ADQ and the information collected on various institutional factors. Through this, it was possible to determine whether there is a

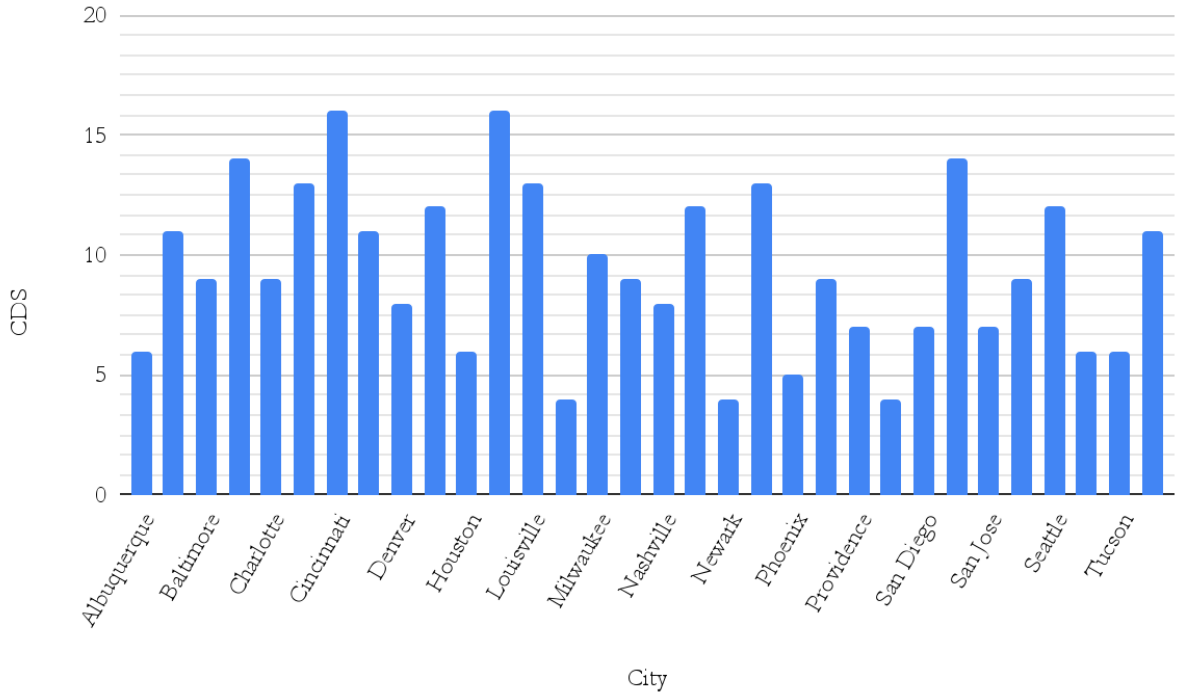
relationship between the institutional characteristics of a city and the quality of data that they provide in their municipal data portals.

Since data was recorded in Microsoft Excel, the preliminary calculations were also done in Excel. After preliminary analysis was completed, the data was imported into Google Sheets and RStudio. Google Sheets were used to create some of the visualizations seen in this report. RStudio was used for the remaining visualizations that were outside of the capabilities of Google Sheets, and for running more in-depth analysis, such as generating summary statistics and running the final regression. While using R, the dplyr, ggplot2, and tidyverse were heavily used to generate graphics and run analyses.

5.2. Core Datasets

Across the 32 cities, the lowest CDS was 4, while the highest was 16. The mean number of CDS per city was 9.406 and the median was 9.00—this is less than half of the maximum CDS possible, and indicates a serious lack of data availability in municipal data portals.

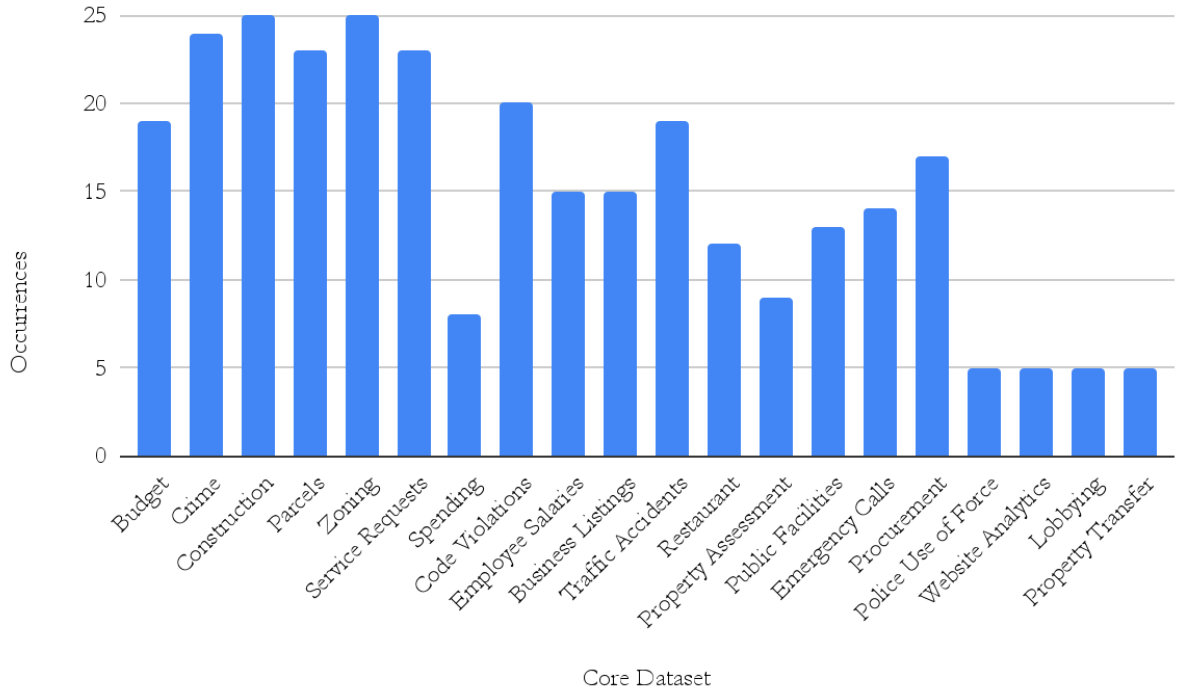
Figure 3: Number of Core Datasets by City



The most frequently available datasets were Construction Permits and Zoning, showing up 25 times across the thirty two cities. As most open data initiatives build off of existing planning and development information, this abundance of occurrences is unsurprising, and may also point to the economic development focus of some cities' OGD initiatives (Johnson 2016).

Data on Police Use-of-Force, Web Analytics, Lobbyist Registration, and Property Transfer were the least commonly available, only appearing five times each across all thirty two cities.

Figure 4: Occurrences of Each Core Dataset Type



5.3. Data Openness

In terms of DO by each CDS, Lobbying had the best average DO score across all 32 cities, with 9.4, while Parcels had the lowest, with 8.29. These scores should be considered in context of the frequency with which the CDS appeared in the data portals.

Table 10: Average DO Score for each Core Dataset

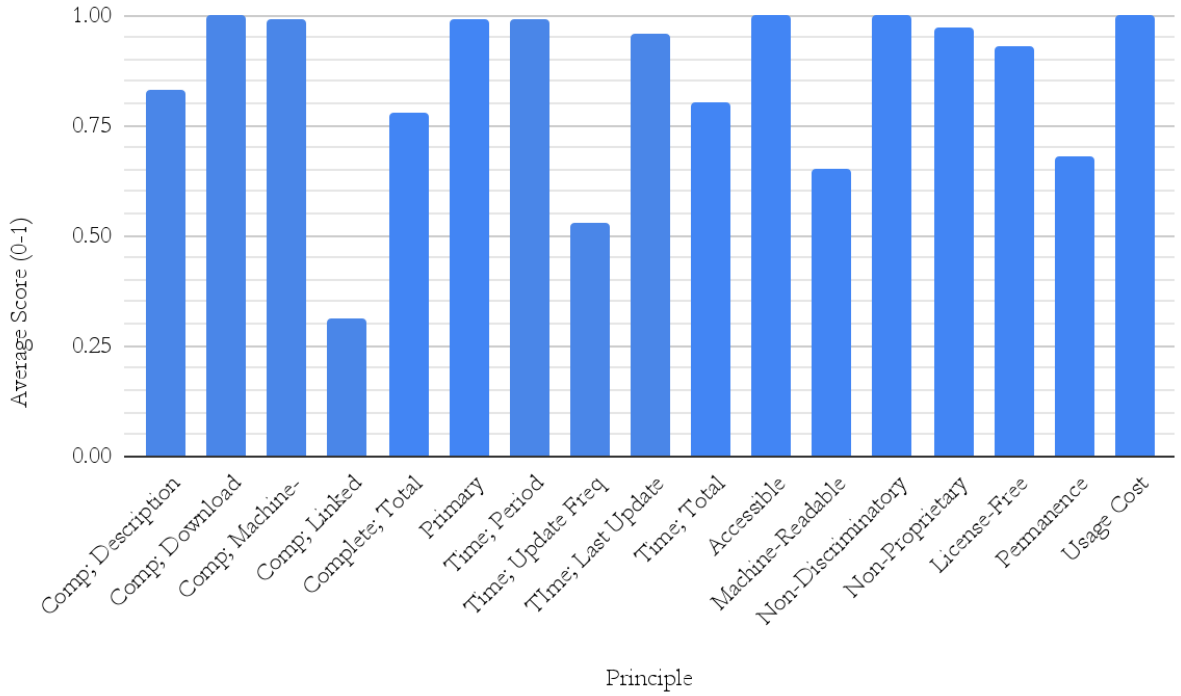
CDS	DO	CDS	DO
Budget	8.81578947	Traffic Accidents	9.11052632
Crime	9.14791667	Restaurant Inspections	9.0625
Construction Permits	9.146	Property Assessment	8.97222222
Parcels	8.29130435	Public Facilities	8.93076923
Zoning	8.314	Emergency Calls	8.92142857
Service Requests	9.03478261	Procurement Contracts	9.02352941

Table 10 (cont.): Average DO Score for each Core Dataset

CDS	DO	CDS	DO
Spending	9.15	Police Use of Force	9.24
Code Violations	8.835	Website Analytics	9.16
Employee Salaries	9.1	Lobbying	9.4
Business Listings	8.71	Property Transfer	8.83

Of the ten principles, cities performed best on Accessibility, Non-Discriminatory and Usage Costs, due to the fact that the existence of a free, accessible online data portal for which data could be collected from would automatically satisfy these principles. The next best principle performances were for Primary, Complete; Machine-Readable, and Time; Period. Cities performed worst on Machine-Readable and Permanence, with only nine cities scoring over 0.5 for Machine Readable, and only five consistently maintained permanent records of previous datasets. Madison and Newark in particular earned scores of 0 on Permanence.

Figure 5: Average Score for each Data Openness Principle



5.4. Data Principle Scores

DPS were calculated for each city at both the principle and the component level, though only the principle level is included here for the sake of space. As mentioned previously, DO scores for Accessible, Non-Discriminatory, and Usage Cost were 1 across all datasets and cities, and therefore are not included in the below table.

Table 11: Cities, Data Principles Scores

City	Complete	Primary	Timely	Machine-Readable	Non-Proprietary	License-Free	Permanence
Albuquerque	0.83	1.00	0.70	1.00	1.00	1.00	0.17
Austin	0.89	1.00	0.93	1.00	1.00	1.00	0.64
Baltimore	0.97	1.00	0.82	0.47	0.89	0.78	0.22

Table 11 (cont.): Cities, Data Principles Scores

City	Complete	Primary	Timely	Machine-Readable	Non-Proprietary	License-Free	Permanence
Boston	0.82	1.00	0.87	0.61	1.00	1.00	0.86
Charlotte	0.83	1.00	0.64	0.50	1.00	0.33	0.22
Chicago	0.87	1.00	0.98	1.00	1.00	0.92	1.00
Cincinnati	0.77	1.00	0.93	0.91	1.00	1.00	0.94
Dallas	0.77	1.00	0.89	1.00	1.00	1.00	0.64
Denver	0.75	1.00	0.78	0.50	1.00	1.00	0.38
Detroit	0.75	1.00	0.60	0.50	1.00	0.08	0.67
Houston	0.79	0.83	0.80	0.33	0.33	1.00	1.00
Las Vegas	0.78	1.00	0.85	0.50	1.00	0.88	0.81
Louisville	0.88	1.00	0.80	0.43	0.77	1.00	0.77
Madison	0.75	1.00	0.70	0.50	1.00	1.00	0.00
Milwaukee	0.63	1.00	0.92	0.69	1.00	1.00	0.90
Minneapolis	0.75	1.00	0.78	0.56	1.00	1.00	0.89
Nashville	0.75	1.00	0.60	1.00	1.00	1.00	0.75
New York City	0.85	1.00	1.00	1.00	1.00	1.00	0.75
Newark	0.56	1.00	0.60	0.43	1.00	1.00	0.00
Philadelphia	1.00	1.00	0.66	0.50	1.00	1.00	0.69
Phoenix	0.55	1.00	0.68	0.50	1.00	1.00	0.60
Pittsburgh	0.78	1.00	0.91	0.50	1.00	1.00	0.89
Providence	0.75	1.00	0.94	1.00	1.00	1.00	0.86
Raleigh	0.88	1.00	0.80	0.50	1.00	1.00	1.00
San Diego	0.86	1.00	0.60	0.50	1.00	1.00	1.00
San Francisco	0.82	1.00	1.00	1.00	1.00	1.00	1.00

Table 11 (cont.): Cities, Data Principles Scores

City	Complete	Primary	Timely	Machine-Readable	Non-Proprietary	License-Free	Permanence
San Jose	0.64	1.00	0.94	0.50	1.00	1.00	0.86
Scottsdale	0.81	1.00	0.69	0.50	1.00	1.00	0.22
Seattle	0.79	1.00	0.90	0.88	1.00	1.00	0.58
Syracuse	0.71	1.00	0.67	0.40	1.00	0.67	0.83
Tucson	0.75	1.00	0.60	0.50	1.00	1.00	0.83
Washington D.C.	0.77	1.00	0.89	0.50	1.00	1.00	0.91
Average	**0.78**	**0.99**	**0.80**	**0.65**	**0.97**	**0.93**	**0.68**

No city was able to achieve full scores across all principles, but San Francisco was very close, with full scores in all categories except for Complete. Only Philadelphia achieved a full score for Complete, and only two out of thirty-two cities achieved full scores for Timely.

5.5. Average Data Quality

The minimum ADQ across the 32 cities was 7.42 for Charlotte's data portal. The maximum ADQ was earned by San Francisco, with 9.82. The median ADQ was 8.79, with a mean of 8.95.

The generally high scores can part can be contributed to the fact that among the data portals being looked at, all cities scored full marks across the OGD principles of Accessible, Non-Discriminatory, and Usage Costs.

Figure 6: Average Data Quality by City

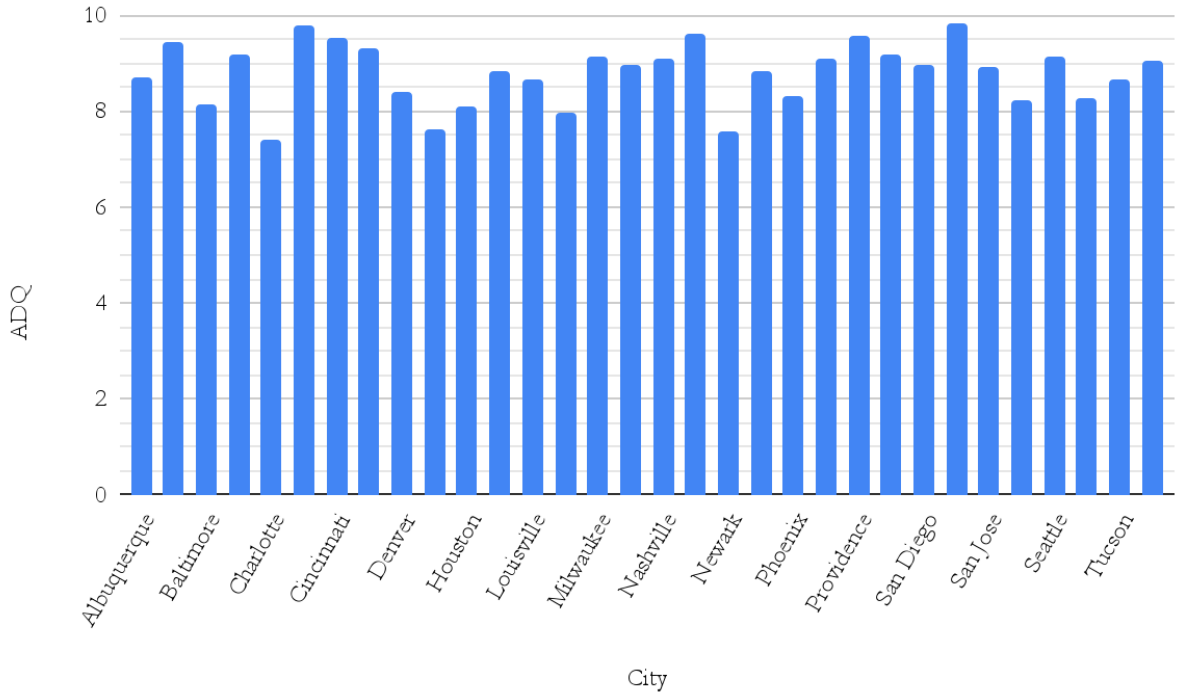


Table 12: Cities, Average Data Quality Scores

City	ADQ	City	ADQ	City	ADQ	City	ADQ
San Francisco	9.82	Boston	9.16	San Jose	8.94	Syracuse	8.28
Chicago	9.77	Seattle	9.15	Philadelphia	8.85	Scottsdale	8.22
New York City	9.60	Milwaukee	9.14	Las Vegas	8.82	Baltimore	8.15
Providence	9.55	Nashville	9.10	Albuquerque	8.70	Houston	8.09
Cincinnati	9.53	Pittsburgh	9.08	Tucson	8.68	Madison	7.95
Austin	9.45	Washington D.C.	9.07	Louisville	8.65	Detroit	7.60
Dallas	9.30	Minneapolis	8.97	Denver	8.40	Newark	7.59
Raleigh	9.18	San Diego	8.96	Phoenix	8.33	Charlotte	7.42

5.6. Relationship between ADQ and Institutional Factors

Running a regression analysis between ADQ and institutional factors reveals a few relationships between the average level of data quality within a given municipal data portal and the environment which the city's OGD initiative is situated within.

Cities with Council-Manager forms of government were likely to have an ADQ score that was on average 0.6899 ($p = 0.0258$) higher than cities with a Mayor-Council form of government. Among the various types of policies that could be utilized to implement OGD initiatives, cities that utilized legislative policies scored significantly higher on average than those cities which had no policy to implement their OGD initiatives (0.8009, $p = 0.0445$).

Model 1; ADQ

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	0.1960	0.3392	0.6282
Year Established	0.08093	0.05937	0.2000
Current Population	-0.000003073	0.000001948	0.1430
Government Type: Council-Manager	0.6899	0.2678	**0.0258 ***
Vote Margin	1.163	0.8216	0.1844
Host			
ArcGIS	-0.4985	0.3781	0.2142
ckan	-0.1550	0.4517	0.7379
Socrata	0.6383	0.3796	0.1208
Total Datasets	-0.0001726	0.0001129	0.1545
Current Voting Age Population	0.000004898	0.000003053	0.1369
Gini Index	3.356	3.544	0.3639
Percent College Graduates	0.0006383	0.01315	0.9622
Percent Unemployed	0.1965	0.1527	0.2244
CDO: Yes	-0.1491	0.2814	0.6067
Policy			
Internal	-0.5485	0.4631	0.2612
Legislative	0.8009	0.3531	**0.0445 ***
Executive	0.5040	0.3649	0.1946
Percent White	0.01567	0.009261	0.1187

5.7. Relationship Between DPS and Institutional Factors

Linear models between each DPS and institutional factors reveals a few relationships between the level of data quality attained by municipal data portals and the environments which the city OGD initiative is situated within. All models within this section can be found in the **Appendix**.

Where the data portal was sourced from, either Forbes or Data.gov, had significant associations with DPS for *Complete; Description* (0.4585, $p = 0.003274$) and *Time; Period* (0.09009, $p = 0.00518$). However, since the source variable was largely included to ensure that there was no statistically significant difference between the two sources, this does not actually provide us with much information regarding the municipal data portals themselves, other than the fact that of the portals chosen, the ones that were selected from the Forbes list had higher average scores for the two principles.

The newer the municipal data portal was, the lower the scores were for DPS such as *Complete; Description* (-0.09066, $p = 0.001423$), unsurprising, as having more robust descriptions would require a greater devotion to an OGD initiative, which would be difficult if the data portal is in its infancy. However, this difference is not statistically significant for the *Complete* principle as a whole.

Current population had a very slight negative association with *Complete; Total* (-0.0000007779, $p = 0.03537$). When the municipality had a Council-Manager form of government *Complete; Linked Data* (0.4080, $p = 0.0072$) had significantly higher scores on average. The same goes for *Complete; Total* (0.1137, $p = 0.02695$), and *Permanence* (0.4886, $p = 0.0160$). Further investigation into how different government types affect the implementation of OGD initiatives would likely provide insight into why these differences exist.

Vote Margin resulted in statistically significant differences in DPS for *Non-Proprietary* (-0.1004, $p = 0.0471$), and *License-Free* (0.6984, $p = 0.03088$). With the way the Vote Margin variable was calculated, this means that greater the difference between the percentages of Democrat and Republican votes cast within a city, the *Non-Proprietary* DPS decreases, while the *License-Free* score increases. Vote Margin is the only statistically significant variable for the *Non-Proprietary* DPS and suggests that cities with a higher vote margin and therefore a higher percent of Democrats over Republicans, are associated with utilizing proprietary platforms (such as Microsoft Excel, etc.) to make their data available. The higher the voting age population, the higher the scores for *Complete; Total* (0.000001222, $p = 0.03491$) and for *License-Free* (0.6984, $p = 0.03088$. A city with a higher population of white constituents is slightly associated with higher *License-Free* DPS (0.009983, $p = 0.00845$), but does not have associations with any other DPS.

Higher Gini Index scores are are associated with significantly higher *Complete; Description* (6.340, $p = 0.000435$) and *Complete; Total* (1.945, $p = 0.00709$) scores, meaning that cities with a greater degree of income inequality are associated with having better data descriptions and general fulfillment of the *Complete* principle. Drawing from Young (2020)'s conclusions regarding negative associations between median income and number of datasets, the higher degree of inequality might speak to an underlying indicator of the city being larger and a greater degree of funding being available to the government to implement OGD initiatives. Higher Gini Index scores are also associated with a dip in average *Time; Period* (-0.6064, $p = 0.4716$) scores.

Percentage of college graduates was not statistically significant for any of the DPS. Higher rates of unemployment were positively associated with *Time; Period* (0.05332, $p = 0.000802$), and negatively associated with *Time; Last Updated* (-0.1625, $p = 0.0487$). The varying association between unemployment and the DPS may be a result of the aforementioned association between inequality within cities and their size, and consequently, the resources available to them to implement OGD.

The host of the data portal had significant associations across multiple DPS, as compared to cities that hosted their portals on platforms that weren't Socrata, CKAN, or ArcGIS. CKAN was associated with slightly lower average scores on *Complete; Description* (-0.5934, $p = 0.03922$), meaning that datasets hosted on CKAN were less likely to have complete descriptive metadata that was available on the site or downloadable along with the dataset, and also on *Complete; Total* (-0.01966, $p = 0.02399$). ArcGIS was associated with lower scores on *Complete; Linked Data* (-0.4825, $p = 0.0187$). Hosting on Socrata had a significantly negative association with *Complete; Linked Data* (-0.4271, $p = 0.03304$), but an overall positive association with *Complete; Total* (0.1530, $p = 0.03391$), which makes sense as its interface provides recommendations to related datasets and also consistently has space for cities to include all three types of Time information.

Cities with Chief Data Officers have lower scores on *Complete; Description* (-0.3204, p = 0.009222), *Time; Period* (-0.07486, p = 0.005128) compared to cities without Chief Data Officers. This is interesting, as the existence of a CDO would hypothetically result in more centralized implementation of OGD initiatives and consequently, higher DPS due to standardization across datasets. Further investigation should be conducted into the role that CDOs do or don't play in setting

standards for data quality in OGD initiatives, and what other responsibilities they might hold alongside the updating of the municipality's data portal.

Among the three policy types, internal policies were associated with slightly higher *Time; Period* (0.09146, p = 0.025255) scores, similar to legislative policies (0.07782, p = 0.14855).

6. Findings and Policy Recommendations

6.1. Cities Need to Expand the Breadth of Data Availability

CDS scores revealed that cities are seriously lagging behind in providing access to datasets that allow them to fully fulfill OGD promises of transparency. With the majority of cities not even providing ten out of the twenty CDS, there is a long way to go for cities to make this information available to their constituents so that they can be informed and/or use the information for their own purposes.

Permanence DPS scores also speak to this area of improvement. Incorporating a culture of keeping older datasets and information is critical not just for archival purposes, but also to contextualize current data. Most of the data generated by a city is done over time, and citizens must be able to access the data of previous years in order to truly be able to understand what they can and cannot expect from their government, and what trends exist. Keeping older datasets also increases accountability for city government, and allows for more data-oriented policy making.

Increasing the breadth of data availability requires much more cooperation across city agencies to aggregate and release these datasets. It also requires cities to reckon with their own internal privacy

and liability concerns regarding releasing this data and allowing citizens to make use of it. Despite the fact that OGD is not a new concept at this point in time, most cities are still in the early stages of creating concrete, robust policy that will allow them to develop the kind of guidelines necessary for them to facilitate the safe release of a broad variety of data. Lack of this type of guidance is what leads to the over-reliance on releasing neutral data for the sake of accomplishing OGD goals.

Most cities had very large data catalogs but were unable to achieve full CDS scores. This study considered CDS scores to be analogous with data availability, but the breadth of city data catalogs despite the neglect of CDS does pose the question of how cities are approaching the question of determining what datasets are critical to include within their data catalogs. Further investigation into this topic by speaking with municipal data managers would be critical for understanding the thought process behind the determination of what data should and should not be included, and whether we can consider an OGD initiative to be successful simply for having a large breadth of data without necessarily focusing on these critical CDS.

On the local level, cities ought to increase the depth and breadth of their datasets by continuing to make datasets from previous years available, and by critically considering what data would increase their transparency, be useful to their citizens, or spur potential innovations. Nationally, it may be possible to encourage this development by establishing a set of guidelines regarding the types of data that should be released that cities can draw from when determining what they should try to include in their data catalogs. It could also be done by advocating for greater use of existing portal

evaluation mechanisms, such as the Open Data Census, which may result in cities working to achieve full scores on such existing publicly available surveys/benchmarks.

There is a risk that this latter suggestion would result in third-parties having undue influence on what data should and should not be included within city data portals. However, since there currently are *no* widely-used guidelines for what types of datasets should be made available in a successful portal, it may be useful to first institute such guidelines and leave space for later revision.

6.2. Cities Have Solid OGD Foundations

ADQ scores across the thirty-two cities revealed that most cities have a solid foundation of compliance with OGD principles. This is in part due to the nature of their OGD initiatives existing online and utilizing portal platforms that automatically aid in and prompt them for the information necessary to adhere with OGD principles. However, this also means that making up the distance between current compliance and full compliance is most often based on some of the most critical OGD principles, namely, characteristics like Machine Readability and Permanence. These principles are also those that might require the most work to achieve full compliance.

Data analysis has revealed that a platform like Socrata, which is able to convert proprietary formats into non-proprietary formats, might be key to achieving full machine readability scores. Analysis also reveals that cities need to consider what *more* their data catalog can do, as most use ArcGIS as a matter of default, as opposed to considering the specific needs of how releasing spatial data might be different from releasing data for the purpose of achieving OGD.

Of the fourteen cities that scored over a 9.0 ADQ, nine used Socrata as their portal host. Additionally, across the DPS, use of different hosts were significantly associated with *Complete* scores for both the principle at large and for its subcomponents. This is interesting to note, as CKAN and Socrata specifically provide space for data managers to include meta descriptive information, whereas for ArcGIS, metadata is often uploaded as a separate file or might be found listed on a secondary webpage linked to the primary page for the dataset.

One improvement municipalities can make is utilize hosts that already allocate space for such meta descriptions within their templates for uploading datasets. But it is important to note that some cities must have the adequate resources not just to utilize Socrata as their catalog host, but also for transferring their existing data catalog to a new platform.

There is also a general necessity for policies that standardize the kinds of metadata information made available for each dataset—municipalities will find that the data they provide will be much more useful not just ot their constituents but also to themselves if they take the time to create institutional templates for uploading metadata that can simply be edited with information specific to each dataset.

The ultimate ideal would be for each city to build its own data portal from the ground up in accordance with OGD principles, as opposed to relying on portal platforms to make their platforms adherent to the principles. Building their own portal would allow each city to provide information in a manner most useful for its government and for its constituents, while ensuring standardized adherence to OGD principles would allow for greater information transfer between cities or for citizens to more easily put together data from multiple cities.

6.3. Laying Effective Groundwork for OGD Implementation

Though centralization of OGD activities is often viewed as a critical part of effective data management, this paper's analysis has shown that it might not be as critical to quality data as has been theorized in the past, at least in the sense that having a CDO does not correlate with a significant increase in ADQ scores, and in some cases, is actually associated with a dip in DPS. CDOs are often created as a position either heading a separate data department, or as one that acts within an existing technology or information department, and these decisions are usually made based upon what would be most appropriate for a given city government.

However, this, combined with the lack of consistent significant average differences in DPS among cities with OGD policies and cities without OGD policies, may suggest that the way OGD policy implementation is being carried out within cities does not actually do much beyond the baseline of creating an OGD initiative with a data portal—that is, the city and its agencies are not necessarily considering what types of structural changes need to be enacted to reach full OGD compliance. While legislative policies have a positive association with higher ADQ scores compared to cities without policies, the fact that only *Time; Period* has positive association with having a policy instead of no policy, and internal policy has a negative association with *Complete; Description* means that there is not evidence that the existence of an OGD policy will result in better compliance.

It is also possible that cities are basing the success of their OGD initiatives on a different set of metrics than that of the ten OGD principles, in which case it would be important to investigate what their individual metrics are. Johnson (2016) provides a good roadmap for such research, and its

expansion to more cities, with specific queries targeted towards understanding how municipal officers prioritize different aspects of data quality, would provide us with greater understanding regarding what roadblocks cities face in achieving full compliance.

In the future, further analysis into how the specificities of OGD policies implemented by municipalities will be critical for understanding to what degree the differences in compliance are rooted in a lack of concrete policy and to what degree concrete policy actually results in significant increases in ADQ scores. While that is outside the scope of this paper, previous research such as that of Zuiderwijk et al. (2014) provide a useful starting point for beginning such investigation.

7. Conclusion

This paper investigates the relationship between the availability and quality of datasets in municipal data catalogs and the institutional factors that might influence those characteristics. In conducting this research, it has been revealed that cities have a long way to go in regards to compliance with the ten OGD principles, but that the past ten years have resulted in the creation of a solid basis upon which these cities can build out their OGD initiatives. I have also found that institutional factors such as Government Type and the type of policy used to implement OGD initiatives have significant effect on both the overall average compliance of a city to OGD principles. On the other hand, there are no institutional factors that are shown to consistently have an association with a city's ability to perform on each individual principle.

OGD has been given a lot of credit over the last decade in terms of its potential to increase transparency, participation, and innovation in cities. However, this research reveals that municipal data catalogs are lacking in providing the information most critical to achieving those goals. Bridging this final gap will require more coordinated effort on the part of cities than the initial implementation of OGD initiatives. Cities must tailor their policies to their own specific situations, consider the impact that decisions such as portal host may have on their ability to provide information about the data that is being published, and examine how they make decisions regarding what data to publish and what successful OGD looks like if they are not aspiring to adhere closely to the ten principles.

This paper is a starting point for additional qualitative research into municipality's OGD initiatives, and further papers should take the time to gain insights directly from city data managers. This paper also outlined several avenues for additional research in previous sections, including investigating what metrics cities use to determine compliance with OGD, and how much the existence of concrete policy actually affects data quality.

Having data portals that provide accurate, useful information is critical to municipalities' ability to function as transparent, accountable entities that can foster spaces for data-driven innovation. In a policy-making environment that is increasingly focused on drawing conclusions and creating solutions supported by data, encouraging cities to invest time and energy into increasing the quality of the datasets that they release can be a driving force in identifying policy issues and their potential solutions, or at the very least, give constituents the information needed to hold their governments accountable to the changes that they promise.

References

Attard, Judie, Fabrizio Orlandi, Simon Scerri, and Sören Auer. "A Systematic Review of Open Government Data Initiatives." *Government Information Quarterly* 32, no. 4 (October 2015): 399–418. https://doi.org/10.1016/j.giq.2015.07.006.

Bogdanović-Dinić, Sanja, Nataša Veljković, and Leonid Stoimenov. "How Open Are Public Government Data? An Assessment of Seven Open Data Portals." In *Measuring E-Government Efficiency*, edited by Manuel Pedro Rodríguez-Bolívar, 5:25–44. Public Administration and Information Technology. New York, NY: Springer New York, 2014. https://doi.org/10.1007/978-1-4614-9982-4_3.

City of Dallas. "Who Are We?." *Dallasopendata.com.* https://www.dallasopendata.com/stories/s/HomePage-About-Story/i62h-y4af/.

Congress.gov. "Text - H.R.4174 - 115th Congress (2017-2018): Foundations for Evidence-Based Policymaking Act of 2018." January 14, 2019. https://www.congress.gov/bill/115th-congress/house-bill/4174/text.

Conradie, Peter, and Sunil Choenni. "On the Barriers for Local Government Releasing Open Data." *Government Information Quarterly* 31 (June 2014): S10–17. https://doi.org/10.1016/j.giq.2014.01.003.

Data.gov. "Open Government." *Data.gov.* https://data.gov/open-gov/.

Data.govt.nz. "Open Data Policies." *New Zealand Government.* August 20, 2020. https://www.data.govt.nz/toolkit/policies/open-data-policy/.

Brown, Meta S. "Free Data Sources: Municipal Open Data Portals For 85 US Cities." *Forbes.* June 30, 2017. https://www.forbes.com/sites/metabrown/2017/06/30/quick-links-to-municipal-open-data-portals-for-85-us-cities/.

Fusi, Federica, and Mary K. Feeney. "Data Sharing in Small and Medium US Cities: The Role of Community Characteristics." *Public Administration* 98, no. 4 (December 2020): 922–40. https://doi.org/10.1111/padm.12666.

HM Government. "Putting the Frontline First: Smarter Government." December 2009. https://ntouk.files.wordpress.com/2015/06/smarter-government-final.pdf

Johnson, Peter A. "Reflecting on the Success of Open Data: How Municipal Government Evaluates Their Open Data Programs." *International Journal of E-Planning Research* 5, no. 3 (July 2016): 1–12. https://doi.org/10.4018/IJEPR.2016070101.

Johnson, Peter, and Pamela Robinson. "Civic Hackathons: Innovation, Procurement, or Civic Engagement?." *Review of Policy Research* 31, no. 4 (July 2014): 349-57. https://doi.org/10.1111/ropr.12074.

Kassen, Maxat. "A Promising Phenomenon of Open Data: A Case Study of the Chicago Open Data Project." *Government Information Quarterly* 30, no. 4 (October 2013): 508–13. https://doi.org/10.1016/j.giq.2013.05.012.

Kučera, Jan, Dušan Chlapek, and Martin Nečaský. "Open Government Data Catalogs: Current Approaches and Quality Perspective." In *Technology-Enabled Innovation for Democracy,*

Government and Governance, edited by Andrea Kő, Christine Leitner, Herbert Leitold, and Alexander Prosser, 8061: 152–66. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. https://doi.org/10.1007/978-3-642-40160-2_13.

Máchová, Renata, and Martin Lnenicka. "Evaluating the Quality of Open Data Portals on the National Level." *Journal of Theoretical and Applied Electronic Commerce Research* 12, no. 1 (2017): 21–41. https://doi.org/10.4067/S0718-18762017000100003.

McNutt, John G., Jonathan B. Justice, James M. Melitski, Michael J. Ahn, Shariq R. Siddiqui, David T. Carter, and Angela D. Kline. "The Diffusion of Civic Technology and Open Government in the United States." *Information Polity* 21, no. 2 (July 1, 2016): 153–70. https://doi.org/10.3233/IP-160385.

Milić, Petar, Nataša Veljković, and Leonid Stoimenov. "Using OpenGovB Transparency Indicator to Evaluate National Open Government Data." *Sustainability* 14, no. 3 (January 26, 2022): 1407. https://doi.org/10.3390/su14031407.

Nahon, Karin, Alon Peled, and Jenniver Shkabatur. "OGD Heartbeat: Cities' Commitment to Open Data." *JeDEM - EJournal of EDemocracy and Open Government* 7, no. 2 (December 14, 2015): 116–36. https://doi.org/10.29379/jedem.v7i2.410.

Nugroho, Rininta Putri, Anneke Zuiderwijk, Marijn Janssen, and Martin de Jong. "A Comparison of National Open Data Policies: Lessons Learned." *Transforming Government: People, Process and Policy* 9, no. 3 (August 17, 2015): 286–308. https://doi.org/10.1108/TG-03-2014-0008.

Obama, Barack. "Transparency and Open Government." *The White House.* January 21, 2009. https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government.

Open Knowledge Foundation and Sunlight Foundation. "U.S. City Open Data Census." *Open Knowledge Foundation.* http://us-cities.survey.okfn.org/.

Orszag, Peter R. "Open Government Directive." *The White House.* December 8, 2009. https://obamawhitehouse.archives.gov/open/documents/open-government-directive.

Public.Resource.Org. "Open Government Data Principles." *Public.Resource.Org.* December 7, 2007. https://public.resource.org/8_principles.html.

Sunlight Foundation. "Open Data Policy Collection." *Open Data Policy Hub.* https://opendatapolicyhub.sunlightfoundation.com/collection/alpha/.

Sayogo, Djoko Sigit, Theresa A. Pardo, and Meghan Cook. "A Framework for Benchmarking Open Government Data Efforts." In *2014 47th Hawaii International Conference on System Sciences*, 1896–1905. Waikoloa, HI: IEEE, 2014. https://doi.org/10.1109/HICSS.2014.240.

Sunlight Foundation. "Ten Principles For Opening Up Government Information." *Sunlight Foundation.* August 11, 2010. https://sunlightfoundation.com/wp-content/uploads/sites/2/2016/11/Ten-Principles-for-Opening-Up-Government-Data.pdf.

Thorsby, Jeffrey, Genie N.L. Stowers, Kristen Wolslegel, and Ellie Tumbuan. "Understanding the Content and Features of Open Data Portals in American Cities." *Government Information Quarterly* 34, no. 1 (January 2017): 53–61. https://doi.org/10.1016/j.giq.2016.07.001.

Veljković, Nataša, Sanja Bogdanović-Dinić, and Leonid Stoimenov. "Benchmarking Open Government: An Open Data Perspective." *Government Information Quarterly* 31, no. 2 (April 2014): 278–90. https://doi.org/10.1016/j.giq.2013.10.011.

Vetrò, Antonio, Lorenzo Canova, Marco Torchiano, Camilo Orozco Minotas, Raimondo Iemma, and Federico Morando. "Open Data Quality Measurement Framework: Definition and Application to Open Government Data." *Government Information Quarterly* 33, no. 2 (April 2016): 325–37. https://doi.org/10.1016/j.giq.2016.02.001.

Young, Matthew M. "Implementation of Digital-Era Governance: The Case of Open Data in U.S. Cities." *Public Administration Review* 80, no. 2 (March 2020): 305–15. https://doi.org/10.1111/puar.13156.

Zhu, Xiaohua, and Mark Antony Freeman. "An Evaluation of U.S. Municipal Open Data Portals: A User Interaction Framework." Journal of the Association for Information Science and Technology 70, no. 1 (January 2019): 27–37. https://doi.org/10.1002/asi.24081.

Zuiderwijk, Anneke, and Marijn Janssen. "Open Data Policies, Their Implementation and Impact: A Framework for Comparison." *Government Information Quarterly* 31, no. 1 (January 2014): 17–29. https://doi.org/10.1016/j.giq.2013.04.003.

Appendix

Model 2: Complete; Description

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	0.4583	0.1226	**0.003274 ****
Year Established	-0.09066	0.02146	**0.001423 ****
Current Population	-0.000001166	0.0000007041	0.125784
Government Type: Council-Manager	0.04674	0.09678	0.638620
Vote Margin	-0.1871	0.2969	0.541553
Host			
ArcGIS	0.007306	0.1367	0.958325
ckan	-0.5934	0.1632	**0.003922 ***
Socrata	-0.1849	0.1372	0.204912
Total Datasets	-0.000022440	0.000004079	0.561835
Current Voting Age Population	0.000001850	0.000001103	0.121806
Gini Index	6.340	1.281	**0.000435 *****
Percent College Graduates	0.007779	0.004754	0.130059
Percent Unemployed	0.03546	0.05517	0.533577
CDO: Yes	-0.3204	0.1017	**0.009222 ***
Policy			
Internal	-0.09320	0.1674	**0.588757 ***
Legislative	0.1458	0.1276	0.277687
Executive	0.1912	0.1319	0.175107
Percent White	0.002018	0.003347	0.558819

Model 3: Complete; Linked Data

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	-0.1292	0.1571	0.4282
Year Established	0.008341	0.02749	0.7672
Current Population	-0.000001945	0.0000009021	0.0541
Government Type: Council-Manager	0.4080	0.1240	**0.0072 ****
Vote Margin	0.6785	0.3804	0.1021
Host			
ArcGIS	-0.4825	0.1758	**0.0187 ***
ckan	-0.1932	0.2091	0.3755
Socrata	-0.4271	0.2091	**0.0334 ***
Total Datasets	-0.00001360	0.00005226	0.7995
Current Voting Age Population	0.000003038	0.000001414	0.03547
Gini Index	1.442	1.641	0.3983
Percent College Graduates	-0.008617	0.006091	0.1849
Percent Unemployed	0.02123	0.07068	0.7695
CDO: Yes	0.02436	0.1303	0.8551
Policy			
Internal	0.2145	0.2144	0.3386
Legislative	0.3226	0.1635	0.0741
Executive	0.1275	0.1690	0.4665
Percent White	-0.0005452	0.004288	0.9011

Model 4: Complete; Total Score

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	0.08228	0.05645	0.17289
Year Established	-0.02058	0.009879	0.06137
Current Population	-0.0000007779	0.0000003242	**0.03527 ***
Government Type: Council-Manager	0.1137	0.04456	**0.02695 ***
Vote Margin	0.1228	0.1367	0.38816
Host			
ArcGIS	-0.01188	0.06293	0.08567
ckan	-0.01966	0.07517	**0.02399 ***
Socrata	0.1530	0.06317	**0.03391 ***
Total Datasets	-0.000009500	0.00001878	0.62296
Current Voting Age Population	0.000001222	0.0000005080	**0.03491 ***
Gini Index	1.945	0.5897	**0.00709 ****
Percent College Graduates	-0.0002094	0.002189	0.92550
Percent Unemployed	0.01417	0.02540	0.58809
CDO: Yes	-0.07402	0.04682	0.14221
Policy			
Internal	-0.03033	0.07706	0.70138
Legislative	0.1171	0.05877	0.07170
Executive	0.07966	0.06073	0.21633
Percent White	0.0003682	0.001541	0.81558

Model 5: Time; Period

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	0.09009	0.02592	**0.005182 ****
Year Established	-0.006716	0.004536	0.166754
Current Population	0.00000008453	0.0000001488	0.581497
Government Type: Council-Manager	-0.006059	0.06277	0.290320
Vote Margin	0.06973	0.06277	0.290320
Host			
ArcGIS	0.02404	0.02889	0.422970
ckan	-0.02707	0.03451	0.449362
Socrata	0.05034	0.02900	0.110500
Total Datasets	-0.000005310	0.000008623	0.550560
Current Voting Age Population	-0.0000001243	0.0000002332	0.604716
Gini Index	-0.6064	0.2707	**0.046716 ***
Percent College Graduates	0.001609	0.001005	0.137763
Percent Unemployed	0.05332	0.01166	**0.000802 *****
CDO: Yes	-0.07486	0.02150	**0.005128 ****
Policy			
Internal	0.09146	0.03538	**0.025355 ***
Legislative	0.07782	0.02698	**0.014855 ***
Executive	-0.06030	0.02788	0.053461
Percent White	-0.0009959	0.0007076	0.186942

Model 6: Time; Update Frequency

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	-0.3827	0.2991	0.227
Year Established	0.04217	0.05234	0.438
Current Population	0.000001692	0.000001718	0.346
Government Type: Council-Manager	0.1465	0.2361	0.547
Vote Margin	0.5823	0.7244	0.439
Host			
ArcGIS	-0.2118	0.3334	0.538
ckan	0.09230	0.3982	0.821
Socrata	0.1032	0.3347	0.764
Total Datasets	0.00002283	0.00009951	0.823
Current Voting Age Population	-0.000002895	0.000002692	0.338
Gini Index	2.2247	3.124	0.487
Percent College Graduates	0.006829	0.01160	0.568
Percent Unemployed	0.1831	0.1346	0.201
CDO: Yes	0.2269	0.2481	0.380
Policy			
Internal	-0.2626	0.4083	0.533
Legislative	0.1444	0.3114	0.652
Executive	-0.04094	0.3217	0.901
Percent White	0.002026	0.008166	0.809

Model 7: Time; Last Updated

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	-0.1371	0.1624	0.4167
Year Established	0.03897	0.02843	0.1978
Current Population	-0.000001149	0.0000009328	0.2436
Government Type: Council-Manager	0.1436	0.1282	0.2865
Vote Margin	-0.01084	0.3934	0.9785
Host			
ArcGIS	0.3388	0.1811	0.0881
ckan	0.3885	0.2163	0.0999
Socrata	0.3585	0.1818	0.0743
Total Datasets	-0.00003039	0.00005404	0.5851
Current Voting Age Population	0.000001803	0.000001462	0.2431
Gini Index	1.226	1.697	0.4850
Percent College Graduates	-0.007870	0.006299	0.2374
Percent Unemployed	-0.1625	0.07310	**0.0481** *
CDO: Yes	0.1511	0.1347	0.2858
Policy			
Internal	-0.08367	0.2217	0.7131
Legislative	0.06161	0.1691	0.7225
Executive	0..01090	0.1747	0.9514
Percent White	0.003392	0.004435	0.4605

Model 8: Time; Total

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	-0.1672	0.09200	0.0965
Year Established	0.02654	0.01610	0.1275
Current Population	0.0000003573	0.0000005284	0.5129
Government Type: Council-Manager	0.09989	0.07263	0.1964
Vote Margin	0.2506	0.2228	0.2847
Host			
ArcGIS	0.02415	0.1026	0.8182
ckan	0.1453	0.1225	0.2604
Socrata	0.1639	0.1030	0.1396
Total Datasets	-0.000001579	0.00003061	0.9598
Current Voting Age Population	-0.0000005742	0.0000008280	0.5024
Gini Index	1.085	0.9611	0.2830
Percent College Graduates	0.0008532	0.003586	0.8154
Percent Unemployed	0.04046	0.04140	0.3494
CDO: Yes	0.1137	0.07631	0.1645
Policy			
Internal	-0.1027	0.1256	0.4308
Legislative	0.09957	0.09578	0.3208
Executive	0.004983	0.09898	0.9608
Percent White	0.001529	0.002512	0.5551

Model 9: Machine Readable

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	0.09912	0.1195	0.4246
Year Established	-0.01688	0.02092	0.4368
Current Population	0.0000002815	0.0000006865	0.6896
Government Type: Council-Manager	-0.05607	0.09436	0.5644
Vote Margin	0.01888	0.2895	0.9492
Host			
ArcGIS	-0.2044	0.1332	0.1533
ckan	-0.2114	0.1592	0.2110
Socrata	0.2606	0.1338	0.0773
Total Datasets	0.000008882	0.00003977	0.8274
Current Voting Age Population	-0.0000004279	0.000001076	0.6984
Gini Index	-0.2423	1.249	0.8497
Percent College Graduates	0.002647	0.004635	0.5794
Percent Unemployed	0.04057	0.05379	0.4666
CDO: Yes	-0.1343	0.09914	0.2028
Policy			
Internal	0.05519	0.1632	0.7416
Legislative	0.02757	0.1244	0.8287
Executive	0.03149	0.1286	0.8111
Percent White	-0.001138	0.003263	0.7340

Model 10: Non-Proprietary

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	0.01212	0.01855	0.5270
Year Established	0.0002690	0.003247	0.9355
Current Population	-0.00000003869	0.0000001065	0.7234
Government Type: Council-Manager	-0.008800	0.01464	0.5601
Vote Margin	-0.1004	0.04493	**0.0471** *
Host			
ArcGIS	-0.006611	-0.02068	0.7552
ckan	0.009666	0.02470	0.7030
Socrata	-0.0007229	0.02076	0.9728
Total Datasets	0.000001990	0.000006172	0.7532
Current Voting Age Population	0.00000006793	0.0000001669	0.7532
Gini Index	0.01106	0.1938	0.9555
Percent College Graduates	0.001546	0.0007194	0.0548
Percent Unemployed	-0.008527	0.008348	0.3290
CDO: Yes	-0.02195	0.01539	0.1814
Policy			
Internal	-0.03386	0.02532	0.2082
Legislative	-0.02609	0.01931	0.2038
Executive	-0.005363	0.01996	0.7931
Percent White	-0.0005705	0.0005065	0.2839

Model 11: License-Free

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	-0.03730	0.1165	0.75494
Year Established	0.03763	0.02040	0.09211
Current Population	-0.0000002904	0.0000006693	0.67271
Government Type: Council-Manager	0.06262	0.009200	0.51017
Vote Margin	0.6984	0.2823	**0.03088 ***
Host			
ArcGIS	-0.03176	0.1299	0.81139
ckan	0.1512	0.1551	0.35074
Socrata	0.1962	0.1304	0.16074
Total Datasets	-0.00006793	0.00003877	0.10756
Current Voting Age Population	0.0000004889	0.000001049	0.65021
Gini Index	-0.5236	1.217	0.67546
Percent College Graduates	0.0002384	0.004519	0.95888
Percent Unemployed	0.06981	0.05245	0.2`0``
CDO: Yes	-0.09302	0.09666	0.35658
Policy			
Internal	-0.1594	0.1591	0.33804
Legislative	0.1045	0.1213	0.40731
Executive	-0.01111	0.1254	0.93098
Percent White	0.009983	0.003182	**0.00945 ** **

Model 12: Permanence

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	0.2080	0.2176	0.3598
Year Established	0.04883	0.03808	0.2261
Current Population	-0.000002602	0.000001250	0.0615
Government Type: Council-Manager	0.4886	0.1718	**0.0160 ***
Vote Margin	0.1998	0.5271	0.7118
Host			
ArcGIS	-0.1744	0.2426	0.4873
ckan	-0.08384	0.2898	0.7777
Socrata	0.1544	0.2435	0.5390
Total Datasets	-0.0001055	0.00007240	0.1730
Current Voting Age Population	0.000004116	0.000001958	0.0594
Gini Index	1.133	2.273	0.6281
Percent College Graduates	-0.004585	0.008439	0.5977
Percent Unemployed	0.04072	0.09793	0.6856
CDO: Yes	0.05673	0.1805	0.7592
Policy			
Internal	-0.2831	0.2971	0.3611
Legislative	0.4881	0.2265	0.0542
Executive	0.4104	0.2341	0.2341
Percent White	0.005258	0.005941	0.3951

Model 13: Complete; Downloadable, Complete; Machine-Readable, Primary, Accessible, Non-Discriminatory, Usage Cost

Variable	Estimate	Standard Error	*p*-Value
Source: Forbes	5.879×10^{-16}	3.911×10^{-16}	0.1609
Year Established	-1.073×10^{-16}	6.6844×10^{-17}	0.1454
Current Population	6.817×10^{-23}	2.246×10^{-21}	0.9763
Government Type: Council-Manager	2.091×10^{-16}	3.087×10^{-16}	0.5122
Vote Margin	5.612×10^{-16}	9.472×10^{-16}	0.5655
Host			
ArcGIS	-2.773×10^{-16}	4.359×10^{-16}	0.5377
ckan	-6.411×10^{-17}	5.207×10^{-16}	0.4364
Socrata	-3.535×10^{-16}	4.377×10^{-16}	0.2440
Total Datasets	-2.295×10^{-20}	1.301×10^{-19}	0.8632
Current Voting Age Population	-1.174×10^{-22}	3.519×10^{-21}	0.9740
Gini Index	1.090×10^{-15}	4.085×10^{-15}	0.7945
Percent College Graduates	-3.110×10^{-18}	1.517×10^{-17}	0.8413
Percent Unemployed	1.477×10^{-17}	1.760×10^{-16}	0.9346
CDO: Yes	-7.948×10^{-17}	3.244×10^{-16}	0.8110
Policy			
Internal	1.152×10^{-15}	5.339×10^{-16}	0.0539
Legislative	2.063×10^{-16}	4.071×10^{-16}	0.6223
Executive	1.266×10^{-16}	4.207×10^{-16}	0.7692
Percent White	-5.027×10^{-18}	1.068×10^{-17}	0.6470