RECOMMENDATIONS REPORT

Prepared by Nina Showell
Contact: showelln@uw.edu

A Policy Analysis of Civic Metadata Standards and Implications for the City of Seattle Open Data Program

15th August 2017

ABSTRACT

The purpose of this report is to assist the City of Seattle in achieving some of the goals expressed in the City of Seattle's 2017 Open Data Plan. Specifically, this report concentrates on the City's goals to improve the quality of the existing open data platform and to increase the discoverability of datasets for the public. To contribute to these efforts, this report considers the use of metadata standards from the perspective that metadata is a positive contributor to the quality, discoverability, and accessibility of datasets. This report analyzes the current use of metadata within the City of Seattle Open Data Program, considers how peer cities' open data programs employ metadata standards, and makes recommendations for how the City of Seattle can increase the quality of metadata provided. It recommends the City of Seattle follow the guidance of the federal government and adopt the Project Open Data Metadata Schema v1.1 as well as the data.json data format, which is natively supported by Seattle's current portal, Socrata. For data dictionaries and column-level metadata, it recommends the City of Seattle encourage departments to develop and implement standards as they see fit, ideally based on existing domain-specific metadata standards. Adopting the aforementioned metadata standards will improve the quality, usability, and discoverability of the datasets the City publishes, helping to achieve two of the goals of the 2017 Open Data Plan.

INTRODUCTION

Seattle is a progressive, mid-size city, and was one of the first in the country to publish data via an open data portal (Douglas, 2010). Seattle's efforts to promote civic data have been successful, but some of the internal practices for gathering, publishing, and presenting information have barely been updated since the Open Data Program began more than five years ago. This has contributed to the current state of the open data portal, which is somewhat convoluted and can be burdensome for users to understand. The availability of good metadata and other forms of descriptive information contributes positively to the discoverability and usability of datasets. As such, providing high-quality metadata should be considered an important part of the City of Seattle Open Data Program. This paper examines the application of civic metadata standards at peer cities, with a goal of providing recommendations to the City of Seattle. This information will ideally be used to strengthen the metadata that is a part of the City's Open Data Program.

One of the objectives expressed in the City's 2017 Open Data Plan is to improve the quality of the existing data platform (City of Seattle, 2017a). The City of Seattle has just launched a new dataset publishing process and efforts to clean up metadata and perform an audit of existing datasets will be completed in the second half of 2017. More specifically, the 2017 Open Data Plan lists five goals for the year: 1) Develop an improved publishing environment; 2) Improve the quality of the existing data platform; 3) Increase discoverability of data to the public; 4) Complete a privacy assessment; and 5) Promote general awareness of Open Data both internally and externally. While the Open Data Program is well-established, the scope of these efforts demonstrates that improving the Program is an ongoing project. The 2017 goals are large, challenging tasks, many of which may require a cultural shift for City employees.

Aspects of the broader Open Data Movement also shape the City's work and complement the City of Seattle's specific goals. From an academic perspective, privacy, metadata, data literacy, and data curation are examples of recent topics studied by open data researchers (Whittington, Calo, Simon, Woo, Young, & Schmiedeskamp 2016; Young & Yan, 2017; Gurstein, 2011). Peer cities are shaping the Smart Cities Movement, planning for the increased use of Internet of Things (IoT) devices, and are measuring progress via Global City Indicators (Zanella, Bui, Castellani, Vangelista, & Zorzi, 2014; Neirotti, De Marco, Cagliano, Mangano, & Scorrano, 2014). One hallmark of the Open Data Movement is its community-driven focus (Angarita, 2016). Additionally, it is critical to recognize that a central mission of city governments is to give voice to the people of the city. Seattle's Civic Technology Advocate, Candace Faber, works to foster connections with members of the community and strengthen the trust between the government and the people. Within the Open Data Program, the City is aiming to create a "virtuous cycle" of improvement (City of Seattle, 2017a). Publishing high-quality datasets will increase awareness about open data, which create valuable insights for the community, promoting further data-driven problem solving. This cycle will generate more demand and will strengthen the use of open data at the City of Seattle, ideally driven by both the community and by internal staff.

The goal of this particular project is to conduct an analysis of civic metadata standards and to make recommendations for the usage of standards within the Seattle Open Data Program. This work contributes to the project to improve the quality of the City's existing data platform.

Partly academic in nature, the work comprising this paper was made possible via a grant awarded to University of Washington professors Carole Palmer and Nicholas Weber. The grant, entitled "Open Data for Public Good: Data Literacy Education for Public Information Professionals," is provided by the Institute of Museum and Library Services (University of Washington, 2017). I want to thank the City of Seattle for their participation as a partner organization. Working with the City's Open Data Team has been a great experience. I would specifically like to thank the City of Seattle Open Data Program Manager, David Doyle, for his excellent mentorship. I am also grateful to Nic Weber at the University of Washington for his advisory role. Andrew Nicklin from GovEx provided helpful guidance for scoping the work and starting the project. I also want to thank the open data leaders and staff at the cities of Boston, Chicago, Los Angeles, New York, Philadelphia, and San Francisco, all of whom were willing to take the time to answer my endless questions. Finally, I am grateful to the Open Seattle organization, its organizers, and its participants, all whom enabled me to gain a greater understanding of how data is used in the community. I hope the work I present here will be of value to others.

BACKGROUND

According to the National Information Standards Organization, metadata is defined as "data about data," and is "the information we create, store, and share to describe things" (NISO, 2004, p.1). There are three types of metadata: Descriptive, Administrative, and Structural, and there are markup languages and data formats that can be used to express these elements in machine-readable ways (p. 6). Metadata informs our understanding of what information is, how it relates to other things, and how we can use it. Metadata is important because it forms the basis for the core functionality of the products, services, and objects we use every day. Looking for a movie to watch? Metadata elements include the title, director, actors, and date the movie was released, among other descriptors. Without this information, context about the movie would be sorely lacking. You would probably not be able to decide if you wanted to watch it, and you might not even know what the movie was about. A parallel situation exists for datasets. Without a title, date, or author, a dataset will likely be useless. If not useless, such a dataset would be untrustworthy or unreliable. You would probably not want to use it.

Luckily, most open data programs provide a good amount of metadata for their datasets. Cities work hard to publish these datasets, and relevant metadata for the datasets typically does exist in some format. The problem discussed in this paper arises when we look at the ways these metadata attributes are expressed. Specifically, inconsistent metadata makes the cross-comparison of datasets challenging, and metadata is often not expressed in machine-readable formats. Within all of the datasets in a single city's open data program, the information can be displayed in different ways. While fields such as title, date, and author are usually present, the formats of the information expressed in these fields is often not standardized. Even when metadata is standardized, it can be unclear which standards are used or if they are consistently enforced. These deficiencies negatively impact usability.

This paper discusses the existence and application of metadata standards within open data programs at cities in the United States. The work of seven peer cities, called the "G7" working group, is analyzed. These cities–Boston, Chicago, Los Angeles, New York City, Philadelphia, San Francisco, and Seattle–were the first in the country to champion the use of open data, starting a movement that spread across the United States in the early 2010s (Douglas, 2010). More than five years later, these cities have open data programs of similar maturity, although there are considerable differences between them, which I will discuss shortly.

Before delving into the work of the cities, it's useful to understand some additional background about how the open data movement began. In 2007, thirty people who were interested in open government met in Sebastopol, California (Trauberer, n.d). The group was diverse and included members ranging from Tim O'Reilly of O'Reilly Media, to Lawrence Lessig of Stanford, to the late Aaron Swartz, an open-source technology advocate. They created a document called "The 8 Principles of Open Government Data." The eight principles are as follows:

"Government data shall be considered open if it is made public in a way that complies with the principles below:

1. Complete
2. Primary
3. Timely
4. Accessible
5. Machine Processable
6. Non-discriminatory
7. Non-proprietary
8. License-free"

(Trauberer, n.d.).

These eight principles set the stage for a larger open data movement for governments. On President Obama's first day in office, he signed a memorandum about government transparency (Executive Office of the President, 2009a). Directing government agencies to be more transparent, this memorandum was soon followed by the Open Government Directive, which established deadlines for agencies to make the data under their purview be openly available (Executive Office of the President, 2009b). This work was then followed by two more directives, Executive Order 13642 "Making Open and Machine Readable the New Default for Government

Information" and Memorandum M-13-13 "Open Data Policy—Managing Information as an Asset." published in 2013 (Executive Office of the President, 2013a; Executive Office of the President, 2013b). The Obama administration's participation in efforts to increase government transparency was unprecedented.

At around the same time, cities and states also began to promote open data. Formed in 2009, the aforementioned G7 cities began to explore open data at the local level. Working as an informal consortium, their first effort was to work together to build an open API for 3-1-1 call center data. CTOs, CIOs, and representatives from the IT departments of the cities had many ideas for similar types of data and data services they thought they could share. Eventually, this work transformed into a discussion of open data portals and data formats. Bryan Sivak, CTO of the District of Columbia, suggested the following, "'The end goal is to create what we're calling the "civic stack" of software,' Sivak said. 'At the end of the day, we should have everything in this organization [the G7], so a city could literally come in and say, "We'll take the whole thing," and become a technology-enabled city.'" (Sivak in Douglas, 2010). This vision was large and optimistic, and the cities began to work together to try to achieve it.

As a member of the G7, Seattle began working with the other cities to translate these goals into reality. Former CTO Bill Schrier said, "We're sick and tired of reinventing the wheel. In most cases, if we had better information sharing, we wouldn't have to do that" (Schrier in Towns, 2010). In 2012, four of the cities–San Francisco, Chicago, Seattle, and New York–launched a shared portal in conjunction with the federal government (Knell, 2012). Hosted on cities.data.gov, this collaboration was the first centralized-yet-decentralized model for government data sharing. Each city hosted their own data, but the datasets were connected to each other via a federated system and could be accessed together on the federal government's site. The project was a success and paved the way for the "federated" data sharing model in which data from multiple cities can be integrated. But by the end of 2012, the G7 team was struggling. Due to leadership changes at three of the seven cities, the group was in flux (Douglas, 2012). The good news was that they had successfully created a national dialogue and had completed some important initial work. The bad news was that the group would soon dissolve.

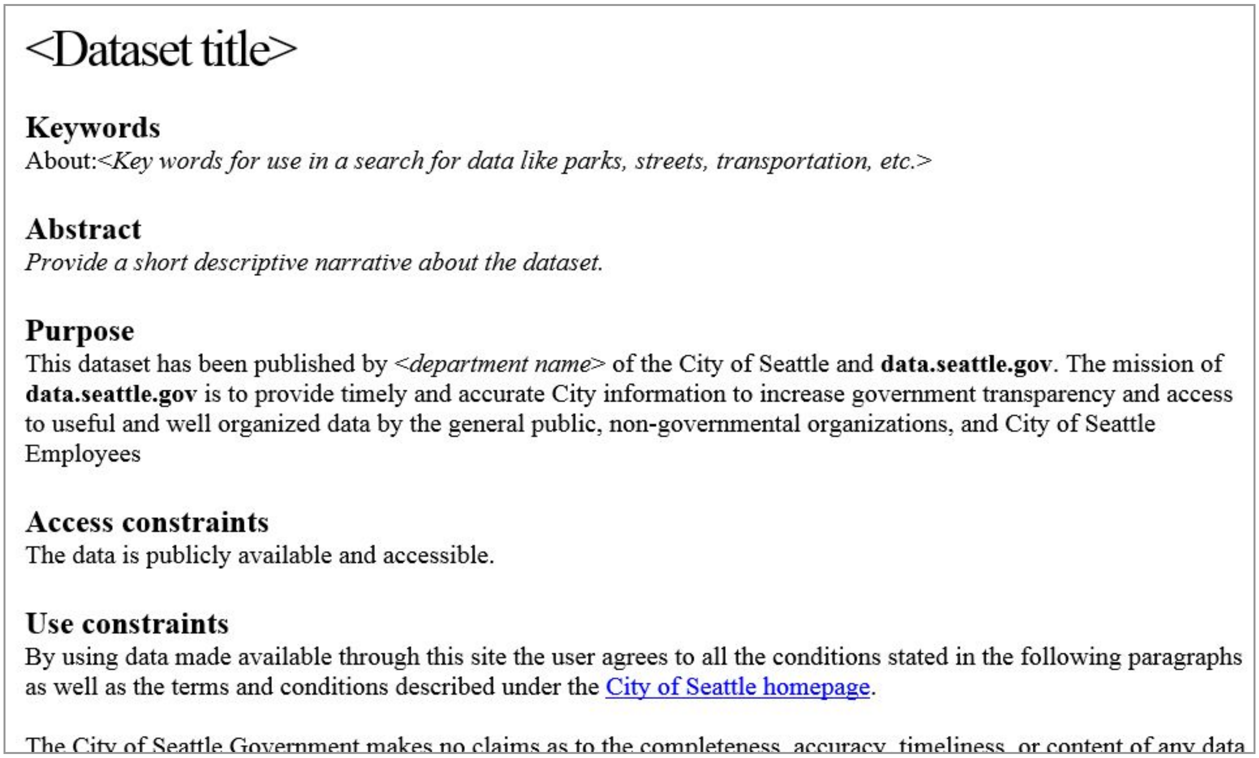
This paper examines the current work of the original G7 cities, concentrating on how these cities use metadata standards. The original focus of the G7 group was to work on data standards and shared services. While the original goals are no longer in service, the individual cities have retained a commitment to open data, and many have also retained the original commitment to standardization. Today, each city's open data program has a slightly different emphasis–for example, a few have chosen to focus more heavily on publishing geospatial datasets–but all are committed to open data work. Beginning with Seattle, I'll now discuss the current practices of each city.

The following analysis is my own creation, and while I have had a few short email exchanges with the staff of the open data programs at each city, it is possible that this paper contains some errors. If this is the case, I welcome any corrections. That said, I consider my perspective to be a positive one. My view of each city's work should be similar to that of a new resident or researcher who is unfamiliar with the city's existing practices. If a city is conducting important work that this paper has omitted, it is likely that the city's residents also have trouble discovering or understanding the same information. As such, this analysis may be helpful from a usability perspective.

ANALYSIS OF METADATA STANDARDS AT THE CITY OF SEATTLE

Although I have not spoken with any of the City of Seattle's previous Open Data Program Managers, it appears there were prior efforts to use metadata standards for the City's datasets. This initial work was conducted before the current Open Data Program Manager stepped into his role about a year ago, so I am relatively uninformed about the details of the City's previous work regarding metadata standards.

Until recently, the City of Seattle's intake and publishing process for new datasets was conducted manually, including the publication of associated metadata. The process was as follows: The owner of a dataset and that department's data advocate, called the Open Data Champion, would approach the Open Data Team about the dataset they wished to publish. The Champion and data owner would fill out the Open Dataset Submission Form, Metadata Template Document, and Open Data Risk Analysis Form. An example of the Metadata Template Document is shown below in Image 1. This information would go to the Open Data Team, and the dataset would be subject to a privacy review and quality review. The Open Data team would perform the ETL process and prepare to publish the dataset. Once published, datasets would be occasionally be reviewed for quality purposes and some datasets would receive either automated or manual updates. This was the process for the publication and upkeep of datasets. At some point in time, Seattle's use of the Metadata Template Document appears to have been discontinued. My impression is that the Open Dataset Submission form was updated to include several questions about metadata, leading to the discontinuation of use of the Metadata Template Document. I am unsure if the Open Data Risk Analysis Form was typically used.



<Dataset title>

Keywords

About:*<Key words for use in a search for data like parks, streets, transportation, etc.>*

Abstract

Provide a short descriptive narrative about the dataset.

Purpose

This dataset has been published by *<department name>* of the City of Seattle and **data.seattle.gov**. The mission of **data.seattle.gov** is to provide timely and accurate City information to increase government transparency and access to useful and well organized data by the general public, non-governmental organizations, and City of Seattle Employees

Access constraints

The data is publicly available and accessible.

Use constraints

By using data made available through this site the user agrees to all the conditions stated in the following paragraphs as well as the terms and conditions described under the City of Seattle homepage.

The City of Seattle Government makes no claims as to the completeness, accuracy, timeliness, or content of any data

Image 1: A Snapshot of the old City of Seattle Metadata Template Document.

In late 2016, David Doyle was hired as the City's new Open Data Program Manager. Hailing from a long career at Microsoft, he quickly recognized that the City of Seattle needed to move away from manual practices for dataset publication. While parts of the work–such as the privacy review–will likely always need to be handled manually, it was clear that other aspects of the publication process could be better implemented by using automated systems. The first task was to improve the intake process for publishing new datasets. This past spring, an intern from the Ada Developers Academy spent several months building an online ingress tool. Replacing the manual process, the new intake tool allows department representatives to identify datasets and submit them for review. It also allows other staff to conduct the privacy and quality reviews and to add internal notes about the dataset. While the City has just started using this new online tool, it has already proven itself worthy by saving a significant amount of employees' time.

Looking ahead toward the next three to five years, Doyle envisions a world in which more of the City's practices become automated. As IoT devices emerge as a common type of technology, the amount of data the City of Seattle holds will rise sharply. This data will be presented as a part of the Open Data Portal, and the portal needs to be able to handle presenting vast volumes of information. The number of datasets that automatically update will also increase, and it's likely the City will be publishing more datasets that update in nearly real time. Managing such a large amount of data will be challenging. The greater the number of datasets available, the harder the organization of the information becomes. Having good metadata will be key to making sure the Open Data Portal continues to be usable.

To get a better idea of the current state of metadata use on the City's Open Data Portal, I analyzed a corpus of 211 key datasets hosted data.settle.gov. The body of datasets was selected from the Asset Inventory Dataset, which is a dataset that contains information about of all datasets hosted in the portal (City of Seattle, 2017b). Seattle's portal contains a large amount information, including official datasets, external datasets, community-created filters, and forms, and geospatial maps. For this paper, the group of key datasets was selected based on the following criteria:

- Publication Stage = Published
  These datasets are published and are currently available online. This eliminates all test datasets
- Public = True
  These datasets are available to the public as opposed to being private to one user
- Provenance = Official
  These datasets were created by the City of Seattle, not by community members
- Domain = data.seattle.gov
  These datasets come from the City and not from an external source
- View Type = Dataset
  The information is represented as a dataset and is not a map, form, calendar, etc.
- Derived View = False
  These datasets are not subsets of other published datasets

Further information about the above terms and the Asset Inventory dataset is available from Socrata at https://support.socrata.com/hc/en-us/articles/218053527-Asset-Inventory-Overview.

The group of 211 datasets selected is not representative of all datasets published on data.seattle.gov, but it covers the major types of datasets departments are most likely to publish. I also believe it is a good representation of the types of datasets the City may want to focus on most closely when considering metadata standards. For instance, future IoT datasets are likely to fulfill the above criteria.

For these 211 datasets, I conducted some preliminary data profiling and analysis of the group. The current metadata practices are as follows:

- All (100%) have a title
- 116 datasets (54%) list a provider of the data

- 103 datasets (48%) display the email address for a person to contact

  Of these, 30 list the email address as "open.data@seattle.gov", meaning only 73 datasets (34%) list an individual point of contact.
- 176 datasets (83%) have a description
- 154 datasets (72%) include tags or keywords
- 115 datasets (54%) claim to receive periodic updates

  This count is comprised of datasets where frequency ≠ null, 0, never, or none
- 151 datasets (71%) include any type of license

As you can see, the above summary statistics demonstrate the lack of an application of metadata standards. Regardless of the particular metadata standard applied, virtually all accepted standards require the information described above to be provided. The incompleteness of certain fields is more problematic than others, but the above fields are generally considered the minimum amount of information a user needs in order to assess a dataset. Luckily, the City of Seattle is doing a good job providing the most critical piece of information: the title of each dataset. That said, if metadata standards were consistently applied, 100% of datasets would include all of the above information.

Let's now examine the metadata for several example datasets.

EXAMPLE DATASET: 911 DATA

This dataset, titled "Seattle Police Department 911 Incident Response," contains data about police responses to 9-1-1 calls. Updated multiple times per day, the dataset has 1.4 million rows, making it one of the larger datasets published on the City's portal (City of Seattle, 2017e). It was originally published in 2010, so it is also one of the oldest datasets available.

The Primer page for the dataset lists the following information:

Title	Seattle Police Department 911 Incident Response
Description	This dataset is all the Police responses to 9-1-1 calls within the city. Police response data shows all officers dispatched. To protect the security of a scene, the safety of officers and the public, and sensitive ongoing investigation, these events are added to the data.seattle.gov only after the incident is considered safe to close out. Data is refreshed on a 4 hour interval.
Updated	July 31st, 2017 [today's date]
Data Provided By	City of Seattle, Department of Information Technology, Seattle Police Department
Data Last Updated	July 31st, 2017 [today's date]
Metadata Last Updated	October 25, 2016
Date Created	October 8, 2010

Views	125k
Downloads	196k
Dataset Owner	Seattle IT
Data Owner	Department of Information Technology
Refresh Frequency	Hourly
Attachments	CadDataReleaserules.docx [Microsoft Word document]
Category	Public Safety
Tags	911, police, crime, incident response, census911incidents
License	http://creativecommons.org/publicdomain/zero/1.0/legalcode
Rows (count)	1.41 M
Columns (count)	19

Table 1: Primer page metadata for Seattle Police Department 911 Incident Response dataset.

Within the dataset, the column information is as follows:

Column Name	**Description**	**Type**
CAD CDW ID	CAD CDW ID	Plain Text
General Offense Number	General Offense Number	Plain Text
Event Clearance Code	Event Clearance Code	Plain Text
Event Clearance Description	Event Clearance Description	Plain Text
Event Clearance SubGroup	Event Clearance SubGroup	Plain Text
Event Clearance Group	Event Clearance Group	Plain Text
Event Clearance Date	Event Clearance Date	Date & Time
Hundred Block Location	Hundred Block Location	Plain Text
District/Sector	Sector	Plain Text
Zone/Beat	Beat	Plain Text
Census Tract	Census_Tract	Plain Text
Longitude	Longitude	Number
Latitude	Latitude	Number
Incident Location	[null]	Location

Initial Type Description	[null]	Plain Text
Initial Type Subgroup	[null]	Plain Text
Initial Type Group	[null]	Plain Text
At Scene Time	[null]	Date & Time

Table 2: Column-level metadata for Seattle Police Department 911 Incident Response dataset.

In addition, an attached Word document called “CadDataReleaserules.docx” includes a paragraph of descriptive information about how the data is recorded. For instance, it notes, “Narratives, Remarks, Text, Entities and Descriptions may contain personal, juvenile, and national security information and are not released” (City of Seattle, 2017e). The information included in this document is helpful for understanding how the data is released but it does not provide field-by-field guidance for the entire dataset, so it is not a data dictionary. Instead, it contains supporting summary information to help users understand the limitations of the data and how it was collected.

In terms of metadata, this dataset contains better metadata than many of the other datasets on data.seattle.gov. Developing high-quality metadata for datasets is a challenge for all City departments, especially given the current lack of guidance for how to do so. And for datasets like this one–a large, complicated dataset that happens to be about a popular subject matter–users often clamor for more information even when provided with a great level of detail. This is not to fault the Seattle Police Department or the users, but is simply to say that this work is challenging and complex, particularly for this type of dataset. In some sense, it appears that even when the City puts forward its best efforts, there is always more work that could be done.

When viewing the comments about the dataset, it is clear that users have questions. For example, one user says, “Any chance we can get a codebook to explain what the variables identify (such as Event clearance date, etc.)?” Another writes, “[to the above user] Did you get any reply from someone? I am curious to know the explanation of meaning of variables as well.” A third user says, “Is there any metadata available for the explanation of variables. If yes, please share the link” (City of Seattle, 2017e). From these comments, it is evident that users are interested in learning more about the dataset, and these three users are specifically requesting a data dictionary! Given that this dataset is one of the City of Seattle’s most popular, it would be a good candidate for additional metadata work. The popularity of this dataset is high enough that a comprehensive data dictionary should be included, and the comments above indicate there is demand.

Despite some of the negative things I’ve said about how the metadata is presented for this dataset, the metadata it includes it is better than most datasets published on the City’s portal.

From the above analysis of the group of 211 datasets, many are lacking information in not just one or two metadata fields, but in multiple fields. The Seattle Police Department 911 Incident Response dataset provides more information than average and is within the top 10% of datasets in terms of the metadata it provides.

Example Dataset: Road Weather Information Stations

Another example of a dataset that includes relatively strong metadata is a dataset called "Road Weather Information Stations," first published by the Seattle Department of Transportation in 2014. This dataset contains information from temperature sensors that take minute-by-minute recordings of road surface and ambient air temperatures (City of Seattle, 2017d). At present, the dataset contains 26.5 million rows and is updated every 15 minutes. Because this dataset is updated frequently and contains a vast volume of information, it is a good example of what datasets from IoT devices might look like in the future.

The metadata for this dataset is shown on the Primer page in Socrata and is similar to the metadata that is a part of the Seattle Police Department 911 Incident Response dataset. In addition, this dataset contains an attachment, called "RWIS Metadata.docx." The attachment is a completed copy of the Metadata Template Document referenced above. It contains free-form textual information about the following subtopics: Title, Keywords, Abstract, Purpose, Access Constraints, Use Constraints, Point of Contact, Credits, Distribution, Entity, Attributes, and Provided by [Department].

Within the attached metadata document, the "Attributes" field contains the following information:

Column	**Datatype**	**Description**
StationName	Text	Name of the station including location description
StationLocation	Location	Latitude and longitude of the station
DateTime	Date & Time	The date and time of the temperature readings
RecordId	Number	System unique identifier
RoadSurfaceTemperature	Number	Temperature of the road surface in degrees Fahrenheit
AirTemperature	Number	Ambient air temperature at the station in degrees Fahrenheit

Table 3: Attributes listed in the attached metadata file for the Road Weather Information Systems dataset (City of Seattle, 2017d).

The above table perfectly explains the information a user needs in order to understand the dataset. As a whole, the attached metadata form works as data dictionary. While it might be nice to present the information in a different format, such as in a non-proprietary file instead of a Microsoft Word document, the attached metadata document is complete and will be helpful for most users. This dataset provides a nice example for how metadata can be provided.

As you can see above, these examples illustrate that the use of metadata at the City of Seattle is good at times, but typically inconsistent. While the above cases are examples of datasets that provide relatively good-quality metadata, there are hundreds of datasets that do not provide this level of sophistication. For the body of 211 datasets analyzed, 176 datasets (83%) have a description, meaning 17% lack this information. It is not uncommon for the metadata to be provided, but to be out of date, inaccurate, or contradictory to other information provided.

Seattle's Open Data Policy is written so that the responsibility for managing both data and metadata lies with each City department, and more specifically, is the responsibility of the data owner (City of Seattle, 2016b). Each department's Open Data Champion works with the dataset owner(s) and with the Open Data Team and Open Data Program Manager, but the ultimate responsibility for providing complete metadata lies with the data owner. The Open Data Champion, Open Data Team, and Open Data Program Manager act in a supporting role, and the Program Manager has the responsibility of making sure the Policy as a whole is followed.

In addition, The City of Seattle Open Data Policy requires the City to use a platform-wide metadata schema. Specifically, it states in section 3, subsection D:

> In partnership with the Open Data Team and coordinated by their Open Data Champion, each City department and office shall…Make data and accompanying metadata open in machine-readable form. The metadata scheme shall allow data publishers to classify selected contextual fields or elements within their dataset as well as adhere to common Meta attributes identified platform-wide, thus empowering data consumers to build automated discovery mechanisms at a granular level. Each row of data shall utilize a unique identifier so that users can verify consistency over time. Using a common metadata taxonomy will allow the Open Data Program to increase discoverability and facilitate successful use of the data (City of Seattle 2016b).

As you can see above, the policy makes it clear that the City's Open Data Program is required to provide metadata and suggests that the use of some type of standard or common format is required. A discussion of the ways the City can comply with this portion of the policy is provided in a later section of this paper.

The benefits of providing high-quality metadata are clear. This document outlines the necessary steps to enable the City to improve its metadata offerings and increase the use of metadata standards for City-provided datasets. To guide this effort, I will now examine the current practices of the six other peer cities that were a part of the original G7 consortium. This information serves as a point of reference and should enable the City of Seattle to gain valuable insights about the work of peer cities.

ANALYSIS OF METADATA PRACTICES OF PEER CITIES

The City of Seattle's Open Data Program does not exist in isolation. Other cities and states are experiencing many of the same growing pains as their open data programs develop. Through researching the practices of cities, I've discovered that open data programs typically undergo a few years of rapid growth and development, followed by a levelling off where the amount of newly-published datasets decreases. This progression appears to be generally consistent across cities. For the purposes of this analysis, I am interested in specifically considering open data programs that are of a similar chronological maturity as Seattle's Open Data Program. While time may not be the best metric for comparison, the goal in choosing this measurement is to avoid focusing on newer open data programs that are still experiencing the initial upswing of program development. Instead, this paper focuses on open data programs that have passed the initial growth phase. As discussed previously, an appropriate selection of peer cities is those that were also part of the G7 group of open data pioneers. While not peers in regard to city size (most have larger populations than Seattle), the timing of these cities' adoption of open data is roughly in line. As such, these cities represent an appropriate group for analysis.

A summary of findings about the policies and practices of the cities is available below. For the sake of brevity, the bulk of the information about the work of peer cities is available in Appendices A, B, C, D, E, and F.

City	**Year Official Policy Enacted**	**Policy Explicitly Mentions Metadata?**	**Metadata Standard Currently Used**	**Portal Provider**
Boston	2014	No	Custom	CKAN
Chicago	2012	Yes	None	Socrata
Los Angeles	2013	No	Portal provider default	Socrata
New York	2012	Yes	Custom	Socrata
Philadelphia	2012	No	Custom	CKAN

San Francisco	2010	No	Custom	Socrata
Seattle	2016	Yes	None	Socrata

Table 4: Comparison of policies and practices of peer cities.

As you can see, the timeline for enacting an official open data policy has varied across cities. While the cities share the common origination of work in 2009 through their participation in the G7 working group, the exact timing of when they enacted open data policies differs. For instance, Seattle maintained an open data portal for years, but did not codify the program's existence until very recently.

In terms of how metadata is discussed as a part of policy, the work of the cities is diverse. Some policies are very short–fewer than 750 words total–but others are much longer. Short policies are often accompanied by a much more thorough implementation guide. In general, I found that the explicit discussion of metadata in a city's policy was not correlated with how the city chose to implement metadata standards. In most cases, mentions of metadata in the open data policies were relatively cursory and simply mentioned that metadata needed to be published.

The seven cities all have different ways of using metadata standards, but there are two general groupings: cities that use no metadata standards or that use the out-of-the-box standard from the portal provider, and cities that have developed a custom metadata standard, typically relying heavily on the established DCAT standard. The portal provider and metadata standard used did not appear to be correlated. You can read more about the exact efforts of the individual cities in Appendices A, B, C, D, E, and F.

The information presented in the above table is a concise way to illustrate the current policies and practices of the group of peer cities. But in reality, the application and practices are much more nuanced than this simple table suggests. The work of the peer cities is complex, has a rich history, and has developed within the context of many other civic efforts. Open data policies and practices are often restricted by issues of funding, personnel, and politics, and no city should be faulted for being subject to these common constraints. The work of managing and maintaining a successful open data program is challenging.

CONSIDERATION OF KEY AREAS AND POSSIBLE SOLUTIONS

This project was conducted with three major goals:

1) To explore the metadata standards currently available, including lack of standards, and to examine how cities apply these standards
2) To review the use of metadata within data.seattle.gov

3) To review the use of data dictionaries within data.seattle.gov.

In researching these three topics, it quickly became evident that the scope and consequences of metadata usage extend far beyond individual datasets. Metadata is highly correlated with increased discoverability and findability of datasets and contributes positively to usability. It is important to deliberately consider the long-term impact that the adoption of possible solutions might have on these related parts of the user experience. For example, how can the City of Seattle select a metadata standard that balances the needs of users with the needs of City departments and staff? How should Seattle weigh the desire for a metadata standard to be specifically relevant to the City with the potentially competing desire for metadata to be expressed in a nationally-recognized format? Does creating a customized standard toe the line of violating the principle that open data should be expressed in a non-proprietary format? (A customized standard would clearly not be developed in secret, but being able to understand customized work can be burdensome for users, and it is often thought that adherence to existing standards is preferable).

When I first set out to write this analysis, I anticipated that it would be relatively technical, examining the precise differences between metadata standards, and would consider field mappings, crosswalks, and the like. In contrast, I've found that it is often the cultural and organizational factors that drive how metadata standards can be best used. The expectations and involvement of constituents also drives decisions for how standards can be implemented. As a result of this insight, I've spent more time focusing on the high-level organizational components than on specific technical details. An overview of the main issues is below.

IMPACT OF DATA PORTAL PROVIDER

The portal provider a city uses can impact the metadata schema that is available for easy use. Both Socrata and CKAN, the leading portals, have certain requirements and default metadata schemas (CKAN, n.d. a; CKAN, n.d. b; Socrata, 2017a; Socrata, 2017b). The portals also have their own strengths and weaknesses. For example, the two have different workflows for how data is published.

Socrata's current required metadata schema is as follows (Socrata, 2017b):

- General Information
  - Dataset Title
  - Brief Description
  - Category
  - Tags / Keywords
  - Row Label (Describes what each row in the dataset represents)
- Licensing & Attribution

  - License Type
  - Data Provided By
  - Source Link
- Semantics & RDF
  - Row Class
  - Subject Column
- API Endpoint
  - Resource Name
  - Row Identifier
- Thumbnail Image
- Contact Information
  - Contact Email

Socrata employs the above fields as part of the default schema, but when uploading a dataset, only the Dataset Title must hold a value. The other fields can be blank. In terms of customization, administrators are able to add additional metadata fields. These metadata fields can be employed to enforce a controlled vocabulary via a drop-down menu of choices.

Seattle benefits from a close working relationship with Socrata and the company's headquarters are just a few minutes away from the civic center. One convenience of this is that the City is able to easily have face-to-face meetings with Socrata staff. Seattle has been able to offer lots of product feedback to Socrata and I anticipate that the two entities will continue to work together in the future. If needed, the City of Seattle should be able to work in conjunction with staff at Socrata to customize the metadata schema for the City's needs.

ORGANIZATIONAL STRUCTURE OF THE OPEN DATA PROGRAM

While the Open Data Policy is helping to create a culture of open data at the City, leadership and policy alone do not form the bulk of the cultural shift toward using data, which requires involvement at all levels.

In particular, the organizational structure of the Open Data Program at the City of Seattle has an impact on how changes to metadata can be adopted and enforced. At present, Seattle uses a decentralized model in which each City department identifies and contributes datasets. The current organizational structure may make it a little tricky to implement a metadata standard because it is likely that the open data team will need to communicate extensively with data owners. Achieving cross-department consistency can be difficult, which is a drawback of the current organizational model. That said, the current decentralized model is beneficial because the Open Data Champions are able to facilitate work between departments and the Open Data Team. The Champions play a large role in supporting the communication pipeline and promoting data use across the City. The decentralized model creates further data evangelism by involving a wide spectrum of staff members. Simply involving additional people in open data efforts can go a long way toward encouraging data use.

While implementing metadata standards is likely to be a time-consuming project, educating staff about metadata is also a key task. Luckily, the City of Seattle is well-equipped to do this via the involvement of the departmental Open Data Champions.

USABILITY & DISCOVERABILITY

In addition to the organization of the Open Data Program at the city level, the engagement of stakeholders and constituents can also impact metadata. Are residents able to find the data they're looking for? Can researchers understand what the columns in specific datasets represent? Who should people contact if they have questions about a dataset?

One option the City could consider would be allowing users to play a much larger role in driving the Open Data Program, including changes to metadata. The Open Seattle group, a local organization, is always looking to build projects using data the City provides. Could the City put in a more formal request for feedback about datasets? Probably. Leveraging the power of interested constituents would go a long way toward creating metadata improvements that are user-driven. For example, a user survey could provide guidance about how well the current metadata is working.

In addition to usability, quality metadata has a positive impact on discoverability. In fact, consistent metadata can be one of the best marketing tools available! Producing high-value datasets is key to the success of the Open Data program and offering quality metadata about these datasets is equally important. By improving the consistency of metadata, the discoverability of all datasets will increase, strengthening the work of the Open Data Program.

CATALOG-LEVEL & DATASET-LEVEL METADATA STANDARDS

The existence of a multitude of overlapping, competing, and irreconcilable standards and standards-governing bodies is not a new phenomenon. Such dilemmas plagued the creators of the internet from its earliest days and have haunted the traditional engineering disciplines for decades (Russell, 2014). The availability of multiple standards is not a new thing. In terms of selecting which standard to use, organizations generally have four choices: 1) Adopt the most popular, easiest, or otherwise “best” standard; 2) Adopt a standard that is less commonly used because it offers some specific desired features; 3) Create a customized standard or model designed specifically for the organization; 4) Not adopt any standard (including both deliberate and unintended forms of non-action).

For Seattle, in terms of adopting a metadata standard to use for all of the City's datasets, the four options are as follows:

1) Adopt the most commonly-used standard, which is Project Open Metadata Schema v1.1 (based heavily on DCAT) plus the data.json standard for providing machine-readable data
2) Adopt a different open data standard, such schema.org or the European standard, INSPIRE
3) Develop and adopt a customized standard (likely based on DCAT)
4) Not adopt any metadata standard

I will now examine these options one by one.

Option 1: Adopt the most common standard, which is Project Open Metadata Schema v1.1, plus the data.json standard for providing machine-readable data

Pros	Cons
● This standard is used by the federal government, so it is well-supported and recognized ● This standard was developed over time based on a wide consensus ● The standard is on its second iteration, so problems with the original version have been resolved ● Makes federation and harvesting of data and metadata easy for external users ● The standard has plenty of available technical documentation ● Machine-readable format is JSON, not XML (JSON is more efficient) ● Able to enforce data.json format programmatically ● Relatively easy to obtain buy-in from staff ● Socrata natively supports data.json ● Implementation of a standard would provide a more consistent user experience compared to the current practices	● May not be able easily implement and/or enforce this standard due to Socrata's existing metadata field requirements ● None of the six other G7 cities have chosen to fully implement this standard, indicating weaknesses ● The standard was developed with the federal government in mind, so it may not apply well to city data ● Federal commitment to the standard may be lagging; there have been no recent major updates ● Implementation would take time ● Enforcement of the standard might be challenging

Table 5: Assessment of adopting the Project Open Metadata Schema v1.1 and data.json standard.

Option 2: Adopt a different open data standard, such schema.org or the European standard, INSPIRE

Pros	Cons
● Seattle could find a standard that best fits	● May not be able easily implement and/or

the City’s needs ● Although I have conducted research about many of the available standards, more research would need to be done ● Implementation of a standard would provide a more consistent user experience compared to the current practices	enforce the standard due to Socrata’s existing metadata field requirements ● Alternate standards may not have strong support or be widely used ● Standards that are not widely used become obsolete quickly ● May be difficult to enforce programmatically ● Unsure if an ideal standard exists ● Developing a custom standard could be a better way to get exactly what the City wants ● Hard to obtain buy-in from staff ● Implementation would take time

Table 6: Assessment of adopting an alternative metadata standard.

Option 3: Develop and adopt a customized standard, likely based on one of the existing standards

Pros	Cons
● Metadata fields can perfectly match the language of the Open Data Policy, e.g. department leads = “Champions” ● Aside from policy language, the standard could better reflect the day-to-day workings of the City ● No redundant or unused fields that don’t apply to the City’s datasets ● Can easily fill in gaps that current standards that are missing, e.g. more fields for dataset update frequency etc. ● Easy to update and change the standard as needed ● City staff would likely be excited about the project ● Implementation of a standard would provide a more consistent user experience compared to the current practices	● May not be able easily implement and/or enforce the standard due to Socrata’s existing metadata field requirements ● Time-consuming to research and implement. Would require work from multiple staff members. ● Time-consuming to update and maintain ● May be difficult to enforce programmatically ● May require work for end users to interpret ● Potential grey area for violating the principle that open data formats be non-proprietary ● Could be too flexible, leading to less accountability. ● If only one or two custom fields are desired (low level of customization), it not be substantially better than using an existing standard.

Table 7: Assessment of adopting a customized metadata standard.

Option 4: Not adopt any metadata standard

Pros	Cons
● Requires no time	● Metadata-related user experience issues

● No buy-in from staff needed ● No cost ● Metadata can still be improved without the restrictions of using a standard ● Political uncertainty at the City may be a reason to not implement a new standard. New mayor to be elected in late 2017	in the current portal would continue to be unaddressed ● Does not contribute to the 2017 Open Data Plan goal of improving the quality of the current platform ● Does not contribute to the 2017 Open Data Plan goal of increasing the discoverability of data ● Metadata usage will likely only get worse over time. May become worse quickly with new datasets from IoT devices ● User complaints will increase as quality of metadata continues to decline ● Does not address current staff burden of fielding the same questions over and over again, which the adoption of a standard would partially address

Table 8: Assessment of not adopting a metadata standard.

A suggestion for which option to ultimately select is discussed in the Recommended Actions section below.

DOMAIN-SPECIFIC METADATA STANDARDS; COLUMN-LEVEL METADATA

In addition to catalog-level/dataset-level metadata standards, the City could also adopt metadata standards for column-level data. Typically domain-specific, column-level metadata is able to precisely explain the nuances of any given dataset. Following the model used above, there are four choices for the adoption of these standards:

- Adopt the most common domain-specific standards for column-level metadata

  Note: May not exist for all types of datasets or departments/service lines
- Adopt less-common standards for column-level metadata based on specific criteria

  Note: May not exist for all types of datasets or departments/service lines
- Have departments develop their own custom metadata standards
- Not use column-level metadata standards

Conducting a full analysis of this area is difficult because of how widely the standards vary. For example, the General Transit Feed Specification (GTFS) is a format for describing public transit schedules. Originally developed by Google, it is now an open-source standard that is used by transit agencies across the United States. By contrast, other domains feature a near total lack of standards. Because IoT devices are a relatively recent invention and can record a breadth of information, metadata standards for these devices do not yet exist. New York City has recently begun to formulate guidelines for managing the data produced by IoT devices and other cities are expected to follow suit (New York City, 2017). It is likely that standards for this data will be developed at a later point in time.

A list of some currently-available metadata standards is as follows (Govex Labs, 2017):

- Blue Button Toolkit, used for healthcare data
- Building & Land Development Specification (BLDS), used for building permits and construction data
- Content Standard for Digital Geospatial Metadata (CSDGM), created by the Federal Geographic Data Committee (FGDC), used for geospatial data
- DATA Act Information Model Schema v1.1, used for data about government spending
- General Transit Feed Specification (GTFS), used for transit data
- Green Button Data, used to describe energy usage
- Housefacts, used for data about residential buildings
- Open311, used for 3-1-1 call center data
- Open Contracting Data Standard (OCDS), used for data about government contracts
- Open Eligibility, used for health and human services data
- Open Trail System Specification (OpenTrails), used for data about public parks and trails
- National Information Exchange Model, used mainly for health and human services data

This list is not comprehensive; hundreds, if not thousands of these standards exist. Some are broad and can cover large amounts of the data that departments at the City of Seattle produce. For example, the Building & Land Development Specification (BLDS) could probably be used for most of the open datasets that the Department of Construction and Inspections publishes. Other departments would have significant difficulty finding metadata standards for the data they produce. For example, I am unsure if there are any standards that could be used to describe data from the Office of the City Clerk. Appropriate standards exist for many types of data but will not address the work of all departments.

DATA DICTIONARIES

Data dictionaries are another way to describe datasets. Typically somewhat free-form in nature, they can include information that is too challenging to express via specific field formats or controlled vocabularies. Data dictionaries often contain notes and comments, and many data dictionaries also provide examples. They essentially act as user guides for datasets.

Currently, the City of Seattle does not provide data dictionaries for all datasets, although some departments do occasionally publish them. In most cases, the information provided is not a full-fledged data dictionary, but is instead a partial data dictionary or guide to a portion of the information contained in the dataset. For instance, the attachment called

“CadDataReleaserules.docx” for the aforementioned Seattle Police Department 911 Incident Response dataset falls into this category (City of Seattle, 2017e).

When written well, data dictionaries are helpful to users because they express information in a plain-language, easy-to-understand format, typically presented as a simple two or three-column table. Through examining the open data portals of the G7 cites, I learned that data dictionaries are a top user request, especially for popular datasets and for datasets that cover complex topics. At the City of Seattle, many of the existing data dictionaries could be expanded to include more information. Data dictionaries could be created for some of the older datasets.

In practice, it would be relatively easy for the City of Seattle to recommend that departments provide data dictionaries for datasets. Data dictionaries can be hard to develop if they are created by people who are not familiar with the data, but if created by the data owner, the work can be accomplished in less than two hours per dataset. Achieving this for newly-published datasets would be easy. Retroactively creating data dictionaries for older datasets would be more challenging.

In addition to data dictionaries for individual datasets, some cities are working toward comprehensive data dictionaries the contain information about all datasets published. For example, the City could strive to express date/time formats in a standard format for every dataset, and this information could be included in a comprehensive data dictionary. This work would be a huge project but would produce a fruitful level of interoperability. It would also go a long way toward promoting cross-departmental collaboration. While likely too large of a project for the short term, developing a platform-wide data dictionary is something to consider for the future.

RECOMMENDED ACTIONS

RECOMMENDATION 1:
ADOPT THE PROJECT OPEN DATA METADATA SCHEMA v1.1 & data.json

As discussed in the section of this paper that covers the current use of metadata standards at the City of Seattle, the City’s Open Data Policy requires the City to use a “common metadata taxonomy,” which is essentially the same thing as a metadata standard (City of Seattle 2016b). Luckily, the City’s portal provider, Socrata, natively supports a well-recognized metadata standard. This standard is called data.json and is also referred to as “slash data.” Started by the federal government as a part of Project Open Data, data.json is a file format that allows the City to publish data in a machine-readable format (Project Open Data, 2017). JSON, shorthand for JavaScript Object Notation, is a lightweight standard for expressing data and is widely used by developers. At the very least, data.json allows Seattle to use a catalog-level metadata standard, enabling the discoverability and federation of datasets by external users. It is a powerful tool that encourages constituents to use the data the City provides. Data.json is already implemented at the platform level, so Seattle is already using this standard as a part of the Open Data Portal and is therefore currently in compliance with the City's Open Data Policy.

Data.json is usefully mainly for enabling data to be expressed in a structured, machine-readable format. This format allows open data programs at other cities to easily federate the data. By adopting this format, the City of Seattle will support the one of the most widely-recognized standard for data interoperability, which was one of the original commitments of the Open Data Movement. Data.json also aids developers who need a way to programmatically handle datasets.

Data.json is a useful tool, but it alone does not solve the usability and discoverability issues that Seattle's Open Data Program currently faces. While great for programmers, data.json is lacking in terms of how the information is presented to users who are less technically-inclined. Put simply, data.json is not a very accessible format. For this reason, I recommend that the City of Seattle adopt an additional metadata standard, called Project Open Data Metadata Schema v1.1. This standard can be used to express metadata information on the Primer pages in Socrata, which is where most users seek information. Even users who are technically-inclined will typically start their work by reading the less-technical information first, so the Primer page of a dataset should be seen as a key resource for all users. It is an important entry point to using a dataset and acts as a guide for how to interpret the data. Adopting the Project Open Data Metadata Schema v1.1 will enrich the quality, consistency, and accessibility of the information presented on the Primer page. These goals are expressed in the 2017 Open Data Plan and adopting the Project Open Data Metadata Schema v1.1 will directly resolve many of the current pain points that these goals strive to to address.

In addition to the reasons above, Seattle should adopt the Project Open Data Metadata Schema v1.1 because this is the schema that is used by the federal government, and because this schema supports data.json, which the City of Seattle is already using. Data.json is itself a part of the Project Open Data Metadata Schema v1.1, so the two formats can both easily be used. The Project Open Data Metadata Schema v1.1 is well-supported by government agencies, has extensive technical documentation, and is recognized within the community (Project Open Data, 2017). The standard has two variants: one for use by the federal government, and one for use by other parties. Seattle should adopt the non-federal variant.

Implementing this standard should be relatively straightforward, but the City of Seattle will need to work in conjunction with the City's portal provider, Socrata, in order to do so. Presently, due to some current limitations with Socrata's default metadata schema, the metadata fields used in the open data portal cannot be extensively modified. As Socrata's platform capabilities are updated on an ongoing basis, it is possible that the ability to change default metadata standards is something that will be available to their customers in the future. If so, the City of Seattle could then collaborate with Socrata to customize their metadata schema to match the Project Open Data specifications. Regardless of the exact practices for implementing such a standard, it is clear that the City would benefit from adopting some type of metadata standard, since additional consistency will increase the usability and discoverability of the datasets the City publishes, helping to achieve the goals presented in the 2017 Open Data Plan (City of Seattle, 2017a). At present, I do not recommend non-action on this matter. Usability and discoverability are goals which are important to the future of how data can and will be used. The City needs to hold itself to these high-level ideals even if implementing a solution is time-consuming (City of Seattle 2107a)."

While peer cities have created custom metadata standards, I am hesitant to recommend that the City of Seattle create its own metadata standard for several reasons. First, such work requires a significant amount of time and effort. While a custom standard might be a better fit, the time required to draft, adopt, implement, and enforce one would be substantial compared to what would be required for implementing a pre-defined solution. Additionally, I am somewhat hesitant to recommend using any type of customized standard based on philosophical grounds. One of the guiding principles of the open data movement is the idea that open data should not be presented in proprietary formats. While a custom standard would be public and not proprietary, the “ownership” of developing and maintaining the standard would mainly lie with the city. Without excellent documentation, customized standards are almost never as usable as standards that are widely adopted. Additionally, contributing to the proliferation of competing standards is not something to support, unless the City is willing to maintain the project over a long period of time. While similar in practice, an imperfect implementation of an established standard is potentially better than promoting the use of a custom metadata standard. For these reasons, the City of Seattle should adopt a pre-existing metadata standard and not create a new one.

In sum, the City of Seattle should adopt the Project Open Data Metadata Schema v1.1 and should work in conjunction with Socrata in order to implement it. Adopting the standard’s non-federal required fields is sufficient; there is no need for the City to use the federal government-focused or optional fields unless they are of value. Due to the pre-existing structure of fields in Socrata, adopting this standard may require help from staff at Socrata. If needed, Seattle can simply create new fields in the portal, but it would be nice to be able to rename or remap some of the existing fields, especially if this can be done programmatically. I encourage the City of Seattle to enquire with Socrata to see if such work might be possible.

RECOMMENDATION 2:
REQUIRE DEPARTMENTS TO INCLUDE COLUMN DESCRIPTIONS OR DATA DICTIONARIES FOR ALL DATASETS

Throughout the course of conducting this research, I discovered that data dictionaries are commonly requested by open data users. Importantly, data dictionaries enable users to research and answer questions on their own without needing to contact the City. This provides excellent service and decreases the burden on staff members, who are no longer tasked with responding to repetitive questions about datasets. If staff find that they are receiving many questions, they can easily respond by adding additional information to the data dictionary.

While data dictionaries take time to develop, writing the documentation is relatively straightforward when done by the dataset owner. Writing a data dictionary for a dataset should typically take less than two hours, even for very complex datasets. Data dictionaries for simple datasets can often be written in thirty minutes or less, assuming they are written by the data owner or by someone who is familiar with how the data was collected.

Aside from comprehensive text-based data dictionaries, short amounts of descriptive info can be added to the "Description" column in Socrata. For example, the previously-mentioned Road Weather Information Systems Dataset from the Seattle Department of Transportation lists a column called "AirTemperature" and a Description of "Ambient air temperature at the station in degrees Fahrenheit" (City of Seattle, 2017d). This short description is all that a user needs to know in order to understand the information expressed in the column. In this particular case, the information is provided in an attached document, but it could easily be listed in the Description column directly. Completing the description fields for dataset columns can be accomplished by data owners is less than an hour per dataset, even for large datasets with many columns.

Because of the simplicity of providing this information, the City of Seattle should require that departments provide column descriptions and/or data dictionaries for all datasets they publish. This information should simply be considered part of the dataset and datasets should not be published without it. Although writing this information takes a little time, it will not place an undue burden on departments. While this recommendation will require additional commitment from departments, asking departments do this work makes sense. Having departments and data owners complete these tasks enables the efficient use of domain-specific staff expertise.

Due to the differences in the types of data published, allowing departments to present the information as either column descriptions or data dictionaries (making the choice as they see fit) is appropriate. In addition, the format used for data dictionaries can also be at the discretion of the department or data owner. At a minimum, these data dictionaries should contain the same information that would be expressed by filling in the short Description field for every column. The existing Metadata Template document, available on the City's website, could easily be updated to serve this purpose, although the file format will need to be changed so that it it nonproprietary (City of Seattle, n.d.). In addition, a few departments have similar metadata templates that would also work; data owners do not need to be bound by the existing template.

Practically speaking, data dictionaries can be included as attachments in Socrata or as website links. To satisfy the requirement that these files be presented in non-proprietary formats, they should be available as .pdf, .txt, or .csv files. Hosting the documents on Github or directly on the City's website would also be appropriate.

As discussed previously, a global metadata dictionary covering the metadata for all datasets at the City of Seattle would be an interesting project. This work is being conducted by several peer cities, discussed in the appendices, and is something Seattle could strive toward. The City is not yet ready to tackle such a project but it would be a nice future extension to this work.

RECOMMENDATION 3: ENCOURAGE DEPARTMENTS TO PROACTIVELY USE METADATA; FOCUS ON ACCESSIBILITY

The City of Seattle Open Policy states in section 1, subsection B:

> When planning for new systems or data collection projects, or modifying existing systems or processes, City departments and offices shall consider which datasets and associated metadata should be published as Open Data (City of Seattle, 2016b).

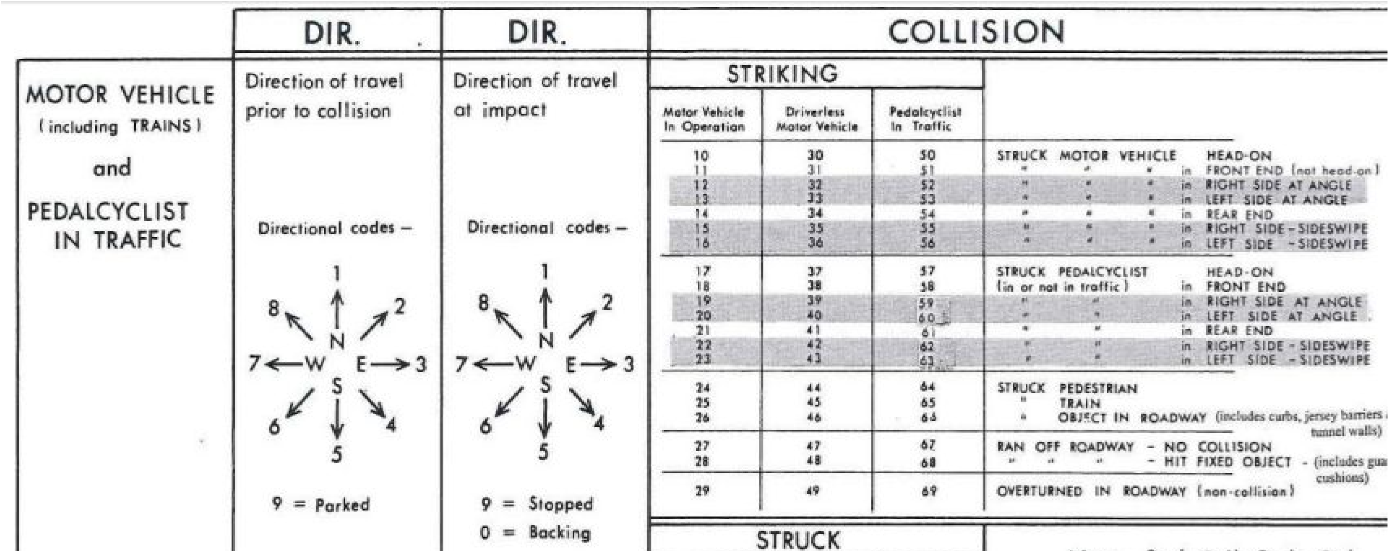
Based on the above, it follows that in addition to the data itself, staff should be considering metadata whenever they begin a new project that might be published as an open dataset. Encouraging departments to be proactive about developing and including metadata is a worthy goal, especially because it is often easiest to create metadata when a project is first started as opposed to tackling the work at a later date. Generally speaking, metadata should be considered as important at a dataset itself. The Open Data team should formally examine the metadata as part of the dataset publication process and should not hesitate to ask the Data Champion or data owner for clarification when needed.

In addition, the Open Data Policy is clear that departments need to recognize that the data they handle is public. One of the largest challenge for the City of Seattle is that of changing the culture to be more data-driven and publically-focused. Finding additional ways to demonstrate the value of Open Data work to departments would be helpful. One example is the recent publication of the residential Underground Storage Tank dataset by the Seattle Fire Department. The information published in this dataset was often the subject of public records requests, and proactively publishing it provides substantial value to the community and saves staff time. This is a great example of data that can be made open, and I encourage the Open Data Program Manager to find additional positive examples to share with the departments.

Returning to metadata, it is important for departments to understand that datasets are not automatically usable or accessible to residents. Metadata, plain language, and other design choices impact usability. Although most departments understand the value of open data, not all take the necessary steps to make sure their datasets are accessible. The Open Data Team should provide context for these efforts by educating department Open Data Champions about why usability and discoverability are important, and should stress that these concepts are just as vital as the data itself. Metadata is a part of accessibility and is something that departments need to proactively consider.

Through discussions with participants in the Open Seattle group, an organization which works on open-source projects for Seattle residents, I learned that data users were often unhappy with how datasets were presented. The sentiment that the publication of datasets seemed like an afterthought was common. In truth, this is not entirely incorrect, because many of the published datasets were originally created in analog formats that were not designed for publication. One interesting case is as follows:

Through the Open Seattle organization, I was introduced to Tim Ganter, creator of the Seattle Collisions web app (Ganter, 2017). He expressed frustration about being able to find the information he needed in order to create his app. Specifically, he was building an app to display data about vehicular collisions. While he had access to the dataset, he didn't know how to interpret it or how he could access the data via the API. Eventually, he was able to find the information he needed, but he had to dig through a lengthy .pdf file. I have a copy of this file, and upon reading it, I discovered that it contained the following guide for interpreting vehicular collisions:



MOTOR VEHICLE
(including TRAINS)
and
PEDALCYCLIST
IN TRAFFIC

DIR.
Direction of travel prior to collision
Directional codes –
1
2
3
4
5
6
7
8
N
E
S
W
9 = Parked

DIR.
Direction of travel at impact
Directional codes –
1
2
3
4
5
6
7
8
N
E
S
W
9 = Stopped
0 = Backing

COLLISION

STRIKING			
Motor Vehicle In Operation	Driverless Motor Vehicle	Pedalcyclist In Traffic	
10	30	50	STRUCK MOTOR VEHICLE HEAD-ON
11	31	51	" " " in FRONT END (not head-on)
12	32	52	" " " in RIGHT SIDE AT ANGLE
13	33	53	" " " in LEFT SIDE AT ANGLE
14	34	54	" " " in REAR END
15	35	55	" " " in RIGHT SIDE – SIDESWIPE
16	36	56	" " " in LEFT SIDE – SIDESWIPE
17	37	57	STRUCK PEDALCYCLIST HEAD-ON
18	38	58	(in or not in traffic) in FRONT END
19	39	59	" " in RIGHT SIDE AT ANGLE
20	40	60	" " in LEFT SIDE AT ANGLE
21	41	61	" " in REAR END
22	42	62	" " in RIGHT SIDE – SIDESWIPE
23	43	63	" " in LEFT SIDE – SIDESWIPE
24	44	64	STRUCK PEDESTRIAN
25	45	65	" TRAIN
26	46	66	" OBJECT IN ROADWAY (includes curbs, jersey barriers tunnel walls)
27	47	67	RAN OFF ROADWAY – NO COLLISION
28	48	68	" " " – HIT FIXED OBJECT – (includes gua cushions)
29	49	69	OVERTURNED IN ROADWAY (non-collision)

STRUCK

Image 2: Matrix for interpreting vehicular collision data. (City of Seattle, 2016c).

As you can see above, this information has a historic basis and was developed before the internet, so it's impossible for anyone to have known that the data would one day be published in bulk. Finding ways to make historic data be more understandable is something departments can accomplish by providing additional context via data dictionaries. Once he was able to find the supporting documentation (which contained more information that just the diagram above), Mr. Ganter was able to understand the dataset and successfully use it.

To address the other concerns of data users, there are several methods the City could employ. The first would be to create a forum or discussion tool for data users in order to solicit feedback and provide a mechanism for communication. While Socrata provides some of these features within the current portal, there are other ways the City could obtain feedback. Conducting a web-based survey of users would be another option.

In addition to soliciting feedback about existing datasets, the City of Seattle could work to conduct usability reviews before new datasets are published. For instance, the Open Data Team could ask Data Champions to complete peer reviews of each other's work, or could lead an "open data book club." Started in Canada, open data book clubs work to review datasets, not books. Typically, users view a dataset, offer feedback, and build projects or proof-of-concepts that use the data. This might be a good way to engage Open Data Champions so they can see the work of other departments.

Regardless of the exact methods used, the City of Seattle Open Data Program could do more to solicit and respond to feedback from both internal and external stakeholders. This work is an ongoing task and it would be ideal if some of it could be formalized as part of next year's Open Data Plan.

RECOMMENDATION 4:
CONDUCT AN AUDIT OF ALL OPEN DATASETS

A large portion of the work required to enable the City of Seattle to use a metadata standard will involve cleaning up the metadata of existing datasets. Unfortunately, applying the Project Open Data Metadata Schema v1.1. to older datasets will likely require manual cleanup because it appears that programmatically changing the metadata schema in Socrata is not possible.

The City of Seattle Open Data Program is planning to conduct an audit that covers several related topics, including archiving older datasets and assigning one common license to all datasets. Editing the metadata for all datasets can easily fit in with this work. The current state of the City of Seattle's Open Data Portal is more deficient than would be ideal. For instance, as discussed previously, 17% of datasets analyzed did not have descriptions. Despite the City's best efforts, metadata may not always be complete or accurate, but the City also needs to ensure that datasets do not lack basic information. Presently, the portal is in need of a full audit, and the City should take steps to ensure that the current situation is not replicated in the future.

I suggest formalizing the audit process so that every dataset is reviewed, at minimum, on a yearly basis. Developing automated tools to use to conduct the audit would be useful and publishing a "health status" dashboard to display the current status of the open data program would also be helpful. If it is too burdensome to audit all datasets at once, the audit could be conducted on a rolling basis throughout the course of the year. Ideally, metadata fields for all datasets can be analyzed, cleaned, and re-formulated to match the standards discussed above. Providing data dictionaries should also be part of the audit, although this may be better managed as a later stage of the audit project.

In addition to the parts of the audit that will be completed by the Open Data Team, there is also an opportunity for the City of Seattle to engage stakeholders. The Open Data Team should leverage resources within the City to perform some of the work. For example, as mentioned above, department Champions could conduct peer reviews. In addition, participants in Open Seattle might be interested in evaluating a handful of datasets or might have comments on datasets they have previously used. Finally, I encourage the City to continue to collect feedback from data users. This feedback should inform the audit and provide guidance for how the City can improve and prioritize the work. For example, it might be useful to start the audit project by making improvements to the top 20 most-downloaded datasets, or to begin with the datasets that receive the most user feedback. Either of these methods would be appropriate for prioritizing the tasks. Working to develop an ongoing and participatory relationship with users should be a longer-term goal for the Open Data Program. Ideally, an ongoing level of audit work and additional communication with stakeholders should be able to prevent the need for such far-reaching audits in the future.

CONCLUSION

In summary, the City of Seattle Open Data Program is currently grappling with issues of data quality, standardization, and usability. Like peer cities, Seattle's early efforts to promote Open Data have been successful, but the City should examine and revise some of the current practices in order to solidify the program's continued success for the future. These efforts are expressed in the 2017 Open Data Plan, which lists five goals (City of Seattle 2017a). The first goal of developing an improved publishing environment has already been accomplished. This report focuses on the next two goals, which are to improve the quality of the existing data platform and to increase discoverability of data to the public, and touches on the final goal, which is to promote the general awareness of Open Data both internally and externally.

Metadata is a major contributor to platform quality and discoverability. As such, options for improving the City's use of metadata are important. This report considers the available options, placing the choices within the context of the City's current implementation, the practices of peer cities, and the desires of external constituents.

To aid the City in accomplishing the goals set forth in the 2017 Open Data Plan, this report makes four recommendations. The first recommendation is to adopt the Project Open Data Metadata Schema v1.1 and the data.json format for machine-readable data. Adopting these metadata standards will facilitate use and discoverability by increasing the quality and consistency of metadata provided. Second, this report recommends requiring departments to include column descriptions or data dictionaries for all datasets. This information will provide further context that may be difficult to express as a part of the standard metadata schema and encourages data owners to express additional information in a comfortable, easy-to-read format. Third, this report recommends that the City encourage departments to proactively use metadata and to consider issues related to accessibility. The goal of this recommendation is to provide a more explicit focus on how data can be used. Finally, this report recommends the City conduct an audit of the existing datasets available on the platform. Cleaning up the existing platform is necessary in order to successfully implement the rest of the recommendations, and it would be a good idea to formalize a process for conducting periodic audits. Taken together, these four recommendations should help the City of Seattle improve the quality of the platform.

Following the aforementioned recommendations will enable the City of Seattle to successfully achieve several of the 2017 goals. Because the goals in the 2017 Open Data Plan are ambitious and visionary, is is likely that additional goals following similar themes will be relevant for years to come. Additionally, the ideas expressed in the 2017 Open Data Plan are significant enough such that any solution will require an extended commitment for improvement. Implementing the recommendations listed in this report will take time and thorough planning, but will set the City on a good path toward achieving the ambitious targets set forth in the 2017 Open Data Plan.

I fully believe that implementing these recommendations will help achieve the current goals of the Open Data Program. I wish the City the best of luck in these efforts.

REFERENCES

Angarita, J. (2016). Unlocking the Potential of Open Data through Community Engagement. *Ash Center at Harvard Kennedy School: Data-Smart City Solutions*. Retrieved from http://datasmart.ash.harvard.edu/news/article/unlocking-the-potential-of-open-data-through-community-engagement-941

Socrata. (2017). *Knowledge Base: Asset Inventory Overview*. Retrieved from https://support.socrata.com/hc/en-us/articles/218053527-Asset-Inventory-Overview

CKAN. (n.d. a). *Metadata*. Retrieved from https://ckan.org/portfolio/metadata/

CKAN. (n.d. b). *Customizing dataset and resource metadata fields using IDatasetForm*. Retrieved from http://docs.ckan.org/en/latest/extensions/adding-custom-fields.html?highlight=metadata

City of Boston. (2014). *Open and Protected Data Policy*. Retrieved from https://data.cityofboston.gov/City-Services/Open-and-Protected-Data-Policy/2rjs-rb6r

City of Boston Analytics Team. (2017). Open Data to Open Knowledge Data Platform is Going Beta. *Latest City of Boston News*. Retrieved from https://www.boston.gov/news/open-data-open-knowledge-data-platform-going-beta

City of Chicago. (2012). *Open Data Executive Order No. 2012-2*. Retrieved from https://www.cityofchicago.org/city/en/narr/foia/open_data_executiveorder.html

City of Chicago. (2017a). *City of Chicago Github Repository*. Retrieved from https://github.com/Chicago

City of Chicago. (2017b). *City of Chicago Open Source Projects*. Retrieved from http://dev.cityofchicago.org/projects/

City of Los Angeles. (2013). *Executive Directive No. 3*. Retrieved from https://www.lamayor.org/sites/g/files/wph446/f/page/file/Executive-Directive-3-Open-Data.pdf?1426620075

City of Los Angeles. (2014). *LA Open Data Policy and Playbook*. Retrieved from http://datala.github.io/od-policy/

City of Los Angeles. (2015). *Wastewater Sewer Database And Data Dictionary April 14 2015*. Retrieved from https://data.lacity.org/dataset/Wastewater-Sewer-Database-And-Data-Dictionary-Apri/yres-8rzw

City of Los Angeles. (2017). *LAPD Arrests 2016* [dataset]. Retrieved from https://data.lacity.org/A-Safe-City/LAPD-Arrests-2016/m58u-93eu

City of New York. (2016). *Open Data Policy and Technical Standards Manual*. Retrieved from https://www1.nyc.gov/assets/doitt/downloads/pdf/nyc_open_data_tsm.pdf

City of New York. (2017). *Guidelines for the Internet of Things: Data Management*. Retrieved from https://iot.cityofnewyork.us/data-management/

City of Philadelphia. (2012). *Executive Order 1-12*. Retrieved from http://www.phila.gov/data/executive-order/

City of San Francisco. (n.d.). *San Francisco: Final Metadata Standard*. Retrieved from https://docs.google.com/spreadsheets/d/1VoM6CPf21s7qNyg5fWRai4lSckn9e9PG18eiHniJ8W8/edit#gid=1785109203

City of San Francisco. (2013). *Open Data Policy*. Retrieved from http://sfbos.org/ftp/uploadedfiles/bdsupvrs/committees/materials/gao_032813_121017.pdf

City of Seattle. (n.d.). *Open Data How To's*. Retrieved from http://www.seattle.gov/tech/open-data-how-tos

City of Seattle. (2016a). *Open Data Playbook v1.0*. Retrieved from http://www.seattle.gov/Documents/Departments/SeattleGovPortals/CityServices/OpenDataPlaybook_Published_2016.08.pdf

City of Seattle. (2016b). *Open Data Policy*. Retrieved from http://www.seattle.gov/Documents/Departments/SeattleGovPortals/CityServices/OpenDataPolicyV1.pdf

City of Seattle. (2016c). *SDOT Collisions*. Retrieved from https://data.seattle.gov/api/assets/B82F757B-0846-4B9E-A242-A943046DF9FD?download=true

City of Seattle. (2017a). *2017 Open Data Plan*. Retrieved from http://www.seattle.gov/Documents/Departments/SeattleIT/City%20of%20Seattle%202017%20Open%20Data%20Plan.pdf

City of Seattle. (2017b). *Asset Inventory* [data file], Retrieved from https://data.seattle.gov/dataset/data-seattle-gov-Asset-Inventory/dcqc-5hci

City of Seattle. (2017c). *Open Data Program 2016 Annual Report*. Retrieved from http://www.seattle.gov/Documents/Departments/SeattleIT/Open%20Data%20Program%202016%20Annual%20Report.pdf

City of Seattle. (2017d). *Road Weather Information Systems* [data file]. Retrieved from https://data.seattle.gov/Transportation/Road-Weather-Information-Stations/egc4-d24i

City of Seattle. (2017e). *Seattle Police Department 911 Incident Response* [data file]. Retrieved from https://data.seattle.gov/Public-Safety/Seattle-Police-Department-911-Incident-Response/3k2p-39jp

Douglas, M. (2010). G7: CIOs From Seven Big-Cities Work Together to Develop Open-Source IT Solutions. *Government Technology*. Retrieved from http://www.govtech.com/e-government/G7-Big-City-CIOs-Work-to-Develop-Open-Source-IT-Solutions.html

Douglas, M. (2012).Gang of 7 Big-City CIOs Forges Ahead Despite Turnover. *Government Technology*. Retrieved from http://www.govtech.com/pcio/Gang-of-7-Big-City-CIOs-Forges-Ahead-Despite-Turnover.html

Executive Office of the President [Obama]. (2009, January 21). *Memorandum on Transparency and Open Government*. Washington, DC: U.S. Government Printing Office. DCPD Number: DCPD200900010. Retrieved from https://www.archives.gov/files/cui/documents/2009-WH-memo-on-transparency-and-open-government.pdf

Executive Office of the President [Obama]. (2009, December 8). *M-10-06: Open Government Directive*. Washington, DC: U.S. Government Printing Office. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf

Executive Office of the President [Obama]. (2013, May 9). *Executive Order 13642: Making Open and Machine Readable the New Default for Government Information*. Washington, DC: U.S. Government Printing Office. Retrieved from https://www.gpo.gov/fdsys/pkg/CFR-2014-title3-vol1/pdf/CFR-2014-title3-vol1-eo13642.pdf

Executive Office of the President [Obama]. (2013, May 9). *M-13-13: Open Data Policy - Managing Information as an Asset*. Washington, DC: U.S. Government Printing Office. Retrieved from https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2013/m-13-13.pdf

Ganter, Tim. (2017). *Seattle Collisions*. Retrieved from http://seattlecollisions.timganter.io/collisions

Govex Labs. (2017). *Civic Data Standards*. Retrieved from http://labs.centerforgov.org/open-data/civic-data-standards/

Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday*, 16(2).

Hillenbrand, K. (2017). Case Study | Boston's Citywide Analytics Team. *Harvard Kennedy School: Data-Smart City Solutions*. Retrieved from http://datasmart.ash.harvard.edu/news/article/case-study-bostons-citywide-analytics-team-1043

Kanowitz, S. (2017). Boston puts open-data quality first. *GCN*. Retrieved from https://gcn.com/articles/2017/03/20/boston-open-data.aspx

Knell, N. (2012). 4 Big Cities Launch Shared Data Platform. *Government Technology*. Retrieved from http://www.govtech.com/e-government/4-Big-Cities-Launch-Shared-Data-Platform.html

National Information Standards Organization (U.S.). (2004). *Understanding Metadata*. Bethesda, MD: NISO Press.

Neirotti, P., De Marco, A., Cagliano, A. C., Mangano, G., & Scorrano, F. (2014). Current trends in Smart City initiatives: Some stylised facts. *Cities*, 38, 25-36.

Project Open Data. (2017). *Project Open Data Metadata Schema v1.1*. Retrieved from https://project-open-data.cio.gov/v1.1/schema/

Russell, A. L. (2014). *Open Standards and the Digital Age: History, Ideology, and Networks*. New York, NY: Cambridge University Press.

Socrata. (2017a). *Best Practices For Metadata Management*. Retrieved from https://support.socrata.com/hc/en-us/articles/115008609707-Best-Practices-for-Metadata-Management

Socrata. (2017b). *Metadata Schema*. Retrieved from https://support.socrata.com/hc/en-us/articles/115008368427

Towns, S. (2010). 'Gang of 7' Big-City CIOs Partner to Share Information and Apps (Opinion). *Government Technology*. Retrieved from http://www.govtech.com/pcio/Gang-of-7-Big-City-CIOs-Share-Information-and-Apps-Opinion.html

Trauberer, J. (n.d.). *The Annotated 8 Principles of Open Government Data*. Retrieved from https://opengovdata.org/

University of Washington. (2017). *University of Washington: About Open Data Literacy*. Retrieved from https://odl.ischool.uw.edu/about

Weisberg, Peri. (n.d.). *San Francisco Metadata: Background and survey of options for San FrancIsco metadata standards that help users find data and allow for tracking and summarizing*. Retrieved from https://docs.google.com/document/d/1dz_-yWePLvfNPX8KZRn2SiP1SCpFcojZeO-8U8SPXgE/edit#heading=h.vcagqsmll6iq

Whittington, J., Calo, R., Simon, M., Woo, J., Young, M., & Schmiedeskamp, P. (2016). Push, Pull, and Spill: A Transdisciplinary Case Study in Municipal Open Government. *Berkeley Technology Law Journal*, 30(2), 1967.

Young, M., & Yan, A. (2017, January). Civic Hackers' User Experiences and Expectations of Seattle's Open Municipal Data Program. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2014). Internet of things for smart cities. *IEEE Internet of Things journal*, 1(1), 22-32.

APPENDIX A

BOSTON

Boston’s open data policy, entitled “City of Boston Open and Protected Data Policy,” was adopted in 2014 following an executive order signed by Mayor Martin Walsh (City of Boston, 2014). The City’s policy does not explicitly discuss metadata, but the policy says that further guidelines are provided in a separate document, called “The City of Boston Technical Standards Handbook.” According to City staff, this exact document is not currently available, but they have other guidelines for the use of metadata. Regardless of the exact presentation of the information, it is clear that the City of Boston’s open data policy is less stringent than Seattle's policy in terms of metadata standards. Boston is doing work related to metadata, but metadata requirements are not explicitly written into the City’s Open Data policy.

Recently, Boston chose to relaunch their open data portal, introducing a new portal called Analyze Boston (City of Boston Analytics Team, 2017). The new portal was developed as a part of the City’s Open Data to Open Knowledge Initiative, funded by grant from the Knight Foundation. One of the explicit purposes of the grant was to create a centralized data catalog. Working with representatives from each City department (akin to Seattle’s departmental Open Data Champions), staff identified datasets for publication. Boston’s previous open data portal was administered via Socrata. The new site, Analyze Boston, uses CKAN and is maintained by Opengov. At present, both sites are available concurrently, although the old site is no longer being updated. The new portal is more organized, features consistent licenses, and is open source. Because this is a relatively new project, the City is continuing to identify and add new datasets and make other improvements to the way the information is presented.

Although the City of Boston’s Open Data Policy does not explicitly mention metadata, the open data team has been working on some major efforts related to metadata and usability, and their work is generally well thought-out. For instance, when launching the new Analyze Boston portal, the number of datasets available was purposely decreased. The reason behind this was that staff had been able to clean up and consolidate older datasets. For example, instead of having separate datasets, each representing a different year, it was possible to combine some related datasets into one comprehensive dataset covering all years (Kanowitz, 2017). Other outdated datasets were retired and the City is currently working to publish new datasets. Actions like these go a long way toward making sure datasets are easy to discover and use. These actions also show that staff are methodically auditing the available datasets. This level of attention to detail is helpful for data users, and this is a project that the City is continuing to work on.

Before contacting anyone at the City, I first conducted a short exploratory analysis of metadata fields in the new portal. It appeared to me that the following fields were used:

Field	**Required / included for all datasets?**

Title	Yes
Description	Yes
Publisher	Yes
Contact Point	Yes
Contact Point Email	Yes
License	Yes
Temporal Notes	No?
Tags	No
Attachments (Excel Docs Etc.)	No
Classification	No
Contact Point Phone Number	No
Landing Page	No
Location	No
Modified	No
Open	No
Released	No
Source	No
Temporal From	No
Theme	No
Type	No

Table 9. Dataset-level metadata requirements at the City of Boston.

The above list was simply based on my observations and is not an official list of the City's requirements. Instead, it represents a publically-accessible view, that is, an impression based entirely on publicly-available information. This can be an informative way to examine the current practices and to see a snapshot of what the metadata for datasets looks like from the perspective of end-users. Because Boston's portal is new, it is likely that the current implementation will continue to change. One interesting thing to note is the use of the "Temporal notes" field. The most common presentation of metadata I found was for datasets to include a Title, Description, Publisher, Contact Point, Contact Point Email, License, and Temporal Notes. Approximately 50% of datasets are shown with this format. Temporal Notes is a freeform text field where information about the exact date range, update frequency, or other temporal information can be recorded, and the City makes good use of this format.

After conducting my own examination of the metadata fields used, I contacted the City of Boston to learn about their internal processes and procedures for metadata use. They sent me an extensive amount of documentation about the development of the new Analyze Boston portal. I also learned that they had put a lot of consideration into metadata standardization. The City of Boston decided to develop their own metadata standard, which is similar to DCAT and to the Project Open Data Metadata Schema, but which includes some adjustments that are customized for the City. While the specifics of this information are the City's info to share, not mine, I feel confident in saying that Boston is doing thoughtful work regarding both metadata and documentation.

Despite the lack of an official policy regarding metadata, Boston is making a superb effort to make datasets usable and discoverable. Metadata is a part of data quality, and it is evident that Boston is attuned not only to issues of quality and usability, but also to the education and inclusion of all residents in these efforts. While the City could do a little more to further the use of metadata–such as providing plain-language data dictionaries for all datasets–their efforts are already leaps and bounds above what many other cities provide.

Finally, I want to touch on the structure of the City's open data program. On an organizational level, much of the City of Boston's open data work is conducted via the Citywide Analytics Team, a relatively new group that works on data issues across the city (Hillenbrand, 2017). The citywide focus allows for strong data evangelism at all levels of city government, with frontline work led by the data coordinator of each department. The existence of a citywide analytics team appears to be a strong model for promoting data use, and I wish the Analytics Team and other staff the best as they continue to work on the Analyze Boston portal.

APPENDIX B

CHICAGO

Signed by Mayor Rahm Emanuel in 2012, the Open Data Executive Order directs agencies and departments to make their data available online via an open data portal (City of Chicago, 2012). The policy does not contain strong language about metadata. Specifically, the executive order makes one reference to metadata: "Each city agency shall...make available online...all datasets and associated metadata under such agency's control." Despite this single reference to metadata, the policy is somewhat similar to the City of Seattle's policy in that it places responsibility for datasets and metadata with the individual agencies and departments.

Datasets are provided in the City's portal, https://data.cityofchicago.org/, which is administered by Socrata. Following the same format as when analyzing the metadata from for Boston, I conducted a short assessment of the fields used in the portal. The fields are as follows:

Field	**Required / included for all datasets?**

Updated (date)	Automatically generated & filled by Socrata
Data Last Updated (date)	Automatically generated & filled by Socrata
Metadata Last Updated (date)	Automatically generated & filled by Socrata
Date Created	Automatically generated & filled by Socrata
Views	Automatically generated & filled by Socrata
Downloads	Automatically generated & filled by Socrata
Data Provided by	Yes? Default field in Socrata
Dataset Owner	Yes? Default field in Socrata
Contact Dataset Owner	Yes? Default field in Socrata
Dataset Title	Yes; required by Socrata
Category	Yes? Default field in Socrata
Tags	Yes? Default field in Socrata
License	Yes? Default field in Socrata
Source Link	Yes? Default field in Socrata
Data Owner	No, but included for most datasets
Frequency	No, but included for most datasets
Time Period	No, but included for most datasets
Last Updated Date via Automated Load	No

Table 10. Dataset-level metadata requirements at the City of Chicago.

Again, the above information is simply my assessment of the City's use of metadata based on the available datasets. It is not an official specification provided by the City. In considering this information, it appears that the City of Chicago fully meets the requirements for metadata as mentioned in the City's policy because all datasets include a basic amount of metadata information (City of Chicago, 2012). Even though this information is simply the default metadata that Socrata uses, it meets the requirements of the policy. It is less clear if Chicago's work fulfills the spirit of the policy, which makes reference to the far-reaching goals of using open data to "empower Chicago's residents by providing them with information necessary to participate in government in a meaningful manner, to assist in identifying possible solutions to pressing governmental problems, and to promote innovative strategies for social progress and economic growth." These ideas expressed in the policy are ambitious, so it appears that Chicago will always have more to strive toward. Like other cities mentioned in this report, Chicago could do further work; providing data dictionaries is an example of a future project the City could undertake, but this work would be significantly above and beyond what the policy requires.

One noteworthy project the City of Chicago has been working on is an effort to engage with developers, researchers, and other advanced data users. For example, Chicago previously completed the technical work for implementing custom data dictionaries. Although not used by the City today, the code and documentation for the project are available on Github (City of Chicago, 2017a). Other projects include OpenGrid, a mapping tool, and RSocrata, a way to connect data from Socrata to R, which is an open-source software package used for statistical analysis. In addition to Github, Chicago maintains a page on the City's website that showcases open source projects built using the City's data (City of Chicago, 2017b). Although not explicitly related to metadata efforts, many of these projects increase the discoverability and usability of the datasets Chicago provides. Furthering stakeholder engagement is clearly important to the City and this is one of the biggest strengths of Chicago's open data program. Promoting these types of projects allows the City to receive feedback from experienced stakeholders and data users. This work also contributes positively to the aspirational goals mentioned above.

APPENDIX C

LOS ANGELES

Los Angeles adopted an open data policy in 2013. Formalized in Executive Directive No. 3, the policy establishes the Open Data Portal and instructs departments to publish data (City of Los Angeles, 2013). The City's policy does not mention metadata. The executive order does say, "Whenever possible, data should be made available in machine-readable format(s)." which indicates that some level of metadata is required, although it does not explicitly use the word "metadata." To implement the policy, The City of Los Angeles developed a strategic policy and implementation playbook (City of Los Angeles, 2014). This playbook contains guidance for how Los Angeles can achieve departmental buy-in and begin the process of opening up the data their government holds. The playbook discusses metadata as a contributor to data quality and is notable in that it explicitly considers the use of controlled vocabularies in metadata development.

Los Angeles maintains two data portals: https://data.lacity.org/ and http://geohub.lacity.org/, used for tabular data and geospatial data. These portals are administered by Socrata and ESRI. The City uses the out-of-the-box metadata fields that are provided for each portal, with some minor tweaks. Los Angeles does not currently implement a standardized metadata format, but does require that a few fields be completed, namely the dataset title, description, owner, and tags. The customization of the portals' standard formats is minimal; for instance, the "owner" field in Socrata has been renamed to "department" to simplify the language, but the practical use of the field is similar regardless of the name.

Like Seattle, Los Angeles follows a distributed organizational model with data stewards from each department contributing to the City's overall open data program. This model is generally viewed positively because it allows departments to use their domain expertise to present the data in an appropriate manner.

In terms of data dictionaries, Los Angeles does not require departments to use them, but encourages departments to make use of the description fields in Socrata. For example, the LAPD Arrests 2016 dataset lists a column within the dataset named "Time" and the Description field reads "In 24 hour military time" (City of Los Angeles, 2017). While this example is rudimentary, it demonstrates that simple plain-language data dictionaries can be easily implemented. On the opposite end of the spectrum, the open data portal includes a few substantially-developed data dictionaries, such as the 44-page guide to the City's sewer system and wastewater network (City of Los Angeles, 2015). When I spoke to staff at the City, they mentioned that they are interested in increasing the use of data dictionaries and are currently exploring options for how to do so. Los Angeles has demonstrated that encouraging departments to fill in the column-level metadata description fields in Socrata is not a substantial burden, suggesting that Seattle could easily adopt a similar practice.

APPENDIX D

NEW YORK CITY

In New York City, Mayor Michael Bloomberg signed Local Law 11 of 2012, requiring city data to be published in a single web portal. This law directed the IT department to develop specific guidance for doing so. Within 6 months, staff prepared the NYC Open Data Policy & Technical Standards Manual which provides the official implementation guidelines. Nearly thirty pages long, this document is a comprehensive resource for how the new law can be implemented. It provides substantial guidance for the specific implementation of metadata standards, and includes entire paragraphs explaining how agencies need to use metadata.

A few portions of the document are below (City of New York, 2016):

Section 3.4.3: Metadata

"Every Agency must create, publish, and maintain on the NYC Open Data Portal accurate metadata for each public data set as set forth in the City Standards for metadata in this Document."

Section 3.4.4: Maintenance

"Every Agency must ensure that each public data set and associated metadata is kept current to the extent that the agency regularly maintains or operationally updates the public data set."

Section 4.1.2.1: Data Dictionaries

“As mandated by Local Law 107 of 2015, all datasets on the Open Data Portal must be accompanied by a plain language data dictionary, with the goal of making the data more understandable.
Outlined below are the minimum standards that must be adhered to:

- Agency name, dataset name, dataset description, and update frequency must all be Provided.
- Each column name should be listed and defined.
- Where applicable and reasonable, terms, acronyms, codes, and units of measure should be defined.
- To the extent practical, a range of possible values should be included. History of modifications to dataset format should be documented.

Agencies may choose to provide additional information deemed relevant, including but not limited to, method of collection, relationship with or between other datasets, system of record, field lengths, etc. Data dictionaries can be provided in a file format of an agencies choosing, but must include the above minimum requirements.”

In addition, Open Data Policy & Technical Standards Manual lays out requirements for agencies to provide data to the larger catalog of datasets, to identify update frequencies for datasets, and to obtain feedback from users (City of New York, 2016). Based on the above policy, the creation and maintenance of metadata is the responsibility of each individual agency. Note that “agency” is the same as a department here in Seattle; there are about 50 in New York City. Each agency is required to meet the minimum metadata standards defined above: agency name, dataset name, dataset description, and update frequency. Compared to other cities, New York City’s policy manual is exceptionally thorough. The requirement to provide a baseline level of information while letting individual agencies determine what additional information to include strikes a nice balance between maintaining catalog-level consistency without minimizing the expertise of individual agencies.

I struggled to conduct a useful analysis of New York City’s implementation of standards. Because of the number of agencies and the differences between them, checking for compliance with standards was too large of a task to undertake. Presently, there are more than 11,000 datasets published on the portal, which is run by Socrata. At a glance, all datasets have titles, which is due to a basic requirement of Socrata’s platform. Almost all datasets appear to have agency names. Descriptions are sometimes missing, and while update frequency is typically listed, some frequencies are inaccurate. Many datasets appear to be missing the plain language data dictionaries that are required in the City’s policy, although the City does have a fantastic standard format for providing this information. That said, New York City does a relatively job of providing basic metadata, especially given the size of the portal and scope of data available, as well as the fact that the portal was established before that most other cities, so it features some older datasets. The strong policy provides substantial guidance for agencies, and even if adherence to the policy is not perfect, it serves as a good guide for how agencies should think about, use, and promote data. The balance between the minimum city-wide requirements and allowing individual agency oversight of additional information is a nice way to implement metadata standards. The work New York City has done regarding data dictionaries is also stellar, and this will continue to be an area of focus over the next year.

APPENDIX E

PHILADELPHIA

Philadelphia adopted an open data policy in 2012, following an executive order by Mayor Michael Nutter (City of Philadelphia, 2012). Like many of the policies of peer cities, the executive order makes no mention of metadata. As such, there are no formal requirements for Philadelphia's open data program use metadata standards.

Philadelphia's open data is available via a CKAN portal and the City also provides a metadata-focused website. In 2015, Philadelphia created a comprehensive metadata catalog, nicknamed Benny. The City developed their own metadata standard based in part on older GIS metadata standards, and it contains six required fields. When I inquired with City staff about Philadelphia's metadata requirements, they expressed the perspective that aside from some basic information (such as the dataset title, description, and date), the contact information for the dataset owner is most important. The thinking is that while other metadata fields provide context, serious data consumers will often have additional questions about the dataset. By connecting data users with data owners, users are able to receive better service and can ask any questions they may have. From a practical perspective, if data owners are receiving multiple similar questions, this feedback should encourage them to publish additional metadata or contextual information for the dataset. In some ways, this model is similar to New York's policy, and to what I have recommended for the City of Seattle: require a few basic fields, and leave the rest of the information at the discretion of the individual departments. User needs inform how information is presented, which is a good way to prioritize the work.

One of the city of Philadelphia's particular strengths is their commitment to working with partner organizations and members of the community. The City's open datasets are available at http://www.phila.gov/data/, but the City also works with http://www. opendataphilly.org, a group which focuses on regional data. On the city's CKAN platform, user engagement is excellent. There are some ways the city could improve—such as providing data dictionaries—but Philadelphia has a strong metadata footing, especially considering that the City's policy does not require one.

APPENDIX F

SAN FRANCISCO

San Francisco was one of the first cities to adopt an open data policy, adopting a policy in 2009 and codifying it in 2010 under the direction of mayor Gavin Newsom. It was then revised in 2013. The policy creates a formal position for a Chief Data Officer (CDO), whose job includes the responsibility of determining the technical standards the City will use for open data. While the original version of the policy contained several references to metadata, the revised version does not. Instead, the amended policy speaks more broadly, stating that the City shall, “Use open, non-proprietary standards when practicable” (City of San Francisco, 2013). The responsibility for providing open data is shared by the CDO and City departments. Departments are required to open their data, but the ultimate responsibility for establishing rules and standards lies with the CDO.

In terms of metadata, the City of San Francisco decided to develop their own custom metadata standard, and conducted a substantial amount of research as a part of the project. The finalized standard includes fields which have been customized to reflect the operational language the City uses. For example, it includes fields such as “Frequency - data change,” and “Frequency - publishing” (San Francisco, n.d.). San Francisco has chosen to use these terms because they accurately describe the City’s operations. San Francisco’s final metadata standard consists of fifteen required fields and eleven conditional/optional fields, as described below:

Field Name	Consensus	Condition	Discussion/Justification	Definition
Title	Required		Title helps discover and select datasets as well as differentiate between similar datasets.	Human-readable name of the asset. Should be in plain English and include sufficient detail to facilitate search and discovery. Avoid acronyms.
Description	Required		Description helps discover and select datasets as well as differentiate between similar datasets.	What the dataset describes. Provide a longer description of the data that can be readily understood by non-technical users.

Category	Required		Category provides a distinct navigation method and groups similar datasets together regardless of source	The category of the data set identified by the list of possible values. If a data set can fall into multiple categories, select the one which is most significant. This list will be subject to change on an ongoing basis.
Department	Required		Department name is needed for navigation and to ensure a single responsible department.	The department that collects and manages the data as the canonical source.
Data dictionary	Required		Data dictionaries are essential to understanding how the data can be used, whether it is to understand fields, differences in fields, and assessing whether or not the data is appropriate for the intended use. However, we want to strike a balance between making it easy to publish and providing enough information for data users. The existing means of adding data dictionaries are either burdensome or clumsy. The implementation of this requirement must be flexible and provide sufficient guidance.	Data dictionary should explain the fields within the dataset in terms of their definition, type, size, and any other pertinent information that describes the dataset
Last updated	Required		Last updated gives an indication of the recency of the data, which helps users determine if it is appropriate to use	Most recent date and time when the dataset was changed, updated or modified.
Frequency - data change	Required		Together with the publishing frequency, this gives us an indication of our timeliness; It also gives data users insight into the rate of data change for planning and use.	Frequency with which dataset changes.

Frequency - publishing	Required		Together with the data change frequency, this gives us an indication of our timeliness; Understanding the frequency of publication is valuable in terms of planning and use of the dataset.	Frequency with which dataset is published.
Unique Identifier	Required		A unique dataset identifier is required for dataset management.	A unique identifier for the dataset or API as maintained within an Agency catalog or database.
Permalink / Identifier	Required		A permalink helps provide continuity for accessing the dataset.	Persistent link to the dataset
URL	Required		A URL provides a more user friendly link.	More descriptive link to the dataset
Public access level	Required - Private		While most data on the platform will be public, public access level gives us a means to track protected or sensitive data and provide a means for internal users to discover and access non-public data.	The degree to which this dataset could be made publicly-available, regardless of whether it has been made available.
Public access level comment	Conditional - Private	Required if not public	If the data is not public, we should provide an explanation and a means for people to access it if eligible.	An explanation for the selected “accessLevel” including instructions for how to access a restricted file, if applicable, or explanation for why a “non-public” or “restricted public” data asset is not “public,” if applicable. Also note options for making the data public, where appropriate, including obfuscation, aggregation, or anonymization.
License/Rights	Required		A license reduces legal uncertainty for data consumers or users	The license with which the dataset or API is published.

Data Steward name	Required - Private		We want to include internal contacts for each dataset to support the data coordinators and to answer dataset questions. However, we do not want this public as it limits our ability to track data questions and response times. In addition, this role contrasts with the data coordinator role, which will be the initial public contact point for department datasets.	Data Steward's name. Who manages the data and is responsible for making changes to the data? Who understands what the dataset includes and can answer questions about it?
Data Steward email	Required - Private		See previous field 15	Data Steward's email address.
Row Count	Conditional	Required if automatically provided by platform.	Row count is a useful indicator of dataset size. However it is too burdensome to manually generate and update this.	
Endpoint	Conditional	Required if the dataset has an API or is an API	An API endpoint facilitates programmatic access to the data.	Endpoint of web service to access dataset.
Geographic unit	Conditional	Required if data includes a geographic column	Geographic unit indicates the geographic level at which the dataset is collected; also helps us track the need to aggregate or summarize data	At what geographic unit is the data collected? For example, if the data is collected by address, it would be Street Address.
Temporal coverage	Conditional	Required if a) temporal data and b) the platform automates it	Temporal coverage provides an easy way to determine the value of a dataset; however, maintaining this manually is too burdensome and prone to error.	The range of temporal applicability of a dataset (i.e., a start and end date of applicability for the data).
Download URL	Conditional	Required if not natively hosted on Socrata or if future platforms do not provide a download mechanism.	A download URL provides access to the data for the purpose of open data.	URL providing direct access to the downloadable distribution of a dataset.

Format	Conditional	Required if not natively hosted on Socrata or if future platforms do not provide multiple download mechanisms.	Knowing the format for non-hosted datasets helps users determine if they can use it, including software needs.	The file format, physical medium, or dimensions of the resource. Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types [MIME].
Tags	Optional		Tags provide a means to include technical language, secondary categories, and acronyms. While we want to encourage the use of tags, we don't believe it should be required as implementing a rigorous approach is too burdensome.	Tags (or keywords) help users discover your dataset; please include terms that would be used by technical and non-technical users.
Program Link	Optional		Program links can provide more information on the origin of the dataset. Not all datasets will have this information.	The URL to the program area web pages
Data notes	Optional		Data notes provide an opportunity to include information not captured in other fields. Not all datasets will have this information.	Are there any concerns about overall data reliability? Are there any changes in data collection or methods that the user should be aware of? Are there any constraints with data accuracy? What levels of confidence with this dataset could the user reasonably assume?
Related documents	Optional		This provides an opportunity to include forms or other types of documents that help to understand the data. Not all datasets will have this information.	Related documents such as technical information about a dataset, developer documentation, etc.

Table 11: Fields in San Francisco's metadata standard (City of San Francisco, n.d.).

As is evidenced above, San Francisco has put a tremendous amount of effort into developing a customized standard. They conducted a comprehensive analysis of open data metadata standards, the application of said standards at the national and state/local levels, and the relevance of available standards to the City of San Francisco (Weisberg, n.d.). I highly recommend reading this report, since it offers one of the best analyses of metadata standards for government data that I have been able to find.

In conversation with staff, I learned that they chose to develop a custom standard because existing standards did not contain all of the elements they felt they needed, especially for describing the frequency at which data is updated. Some schemas (e.g. DCAT) have extensible fields, but additional fields can become burdensome if too many are added. In developing their own standard, San Francisco was able to limit the fields to only those they felt were necessary.

In sum, San Francisco has put in a lot of effort to analyse possible metadata options and implement a solution that works best for the city. While such a project is a large undertaking, the documentation and reports that were produced as a part of the project are an extremely useful resource, and the finalized metadata standard seems to be working well.