2019

A rubric driven evaluation of open data portals and their data in transportation

Archana Venkatachalapathy
Iowa State University

A rubric driven evaluation of open data portals and their data in transportation

by

Archana Venkatachalapathy

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Civil Engineering (Intelligent Infrastructure Engineering)

Program of Study Committee:
Anuj Sharma, Major Professor
Christopher Day
Soumik Sarkar

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2019

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

ACKNOWLEDGMENTS

I would like to thank my major advisor Dr. Anuj Sharma for his constant guidance and support. His valuable inputs have been relevant in shaping the structure and content of this thesis. He consistently allowed this thesis to be my own work, steering me in the right direction along the way. I would also like to thank my program of study committee members for their helpful suggestions on the topic of study. A special mention to Skylar Knickerbocker (Research Engineer, Institute for Transportation) for his valuable inputs and suggestions throughout this project. He has been of tremendous help specially in designing the tableau visualization tool for this project.

I express my deepest gratitude to my family and friends for their unfailing support and continuous encouragement.

ABSTRACT

In recent years, the open data movement is gaining momentum in the transportation industry with multiple State's Department of Transportation (DOT) launching their own repository of datasets. The quality of data, ease of usage and availability of metadata varies from source to source. There is an imminent need to assess the quality of open data portals to provide agencies a yardstick to measure their performance and draw inspirations from higher ranking portals. We propose a data portal evaluation rubric (DPER) which can serve this purpose. DPER is designed to capture the essence of the National Open Data Policy. The DPER was used to evaluate 43 data portals at the state (39) and national level (4) which provide transportation datasets. DPER evaluates the quality of the portal, the openness of data, and the relevance of its content to the transportation sector. The portal of the State of New York scores the highest due to its user-friendly interface with interactive visualization tools, relevant data content, detailed data information and useful API references for application developers.

CHAPTER 1. INTRODUCTION

"Data is a precious thing and will last longer than the systems themselves", says Tim Berners-Lee, father of the World Wide Web which celebrated its 30th anniversary recently (*1*). In current times data has grown from scarcity to abundance. With many hailing data as the new oil, it is becoming a highly influential agent in decision making (*2*). Data driven research is strongly establishing itself in every sector of development and is responsible for numerous innovations. This surge of data is expanding several industries such as data science, cloud computing, big data analytics and management. However, the access to this data remains restricted which raises the question of whether the data is being used to its full potential. In this digital age, the internet has grown into the most powerful resource tool which can be used to exploit the complete potential of this data, if access is provided.

This chapter begins by introducing the concept of open data and its principles. To understand the roots of the open data movement we discuss the key events in history leading to the initiative for open government data. Shifting into the field of transportation we have identified instances which highlight the benefits of open publication of data. Another dimension of open data is big open data which is also discussed in this chapter. Finally we conclude with the objective of the thesis focusing on the evaluation of open data portals and their data in transportation.

Definition of Open Data

Accessibility is the core principle of open data. According to the Open Data Handbook, open data "... is data that can be freely used, re-used and redistributed by anyone – subject only, at most, to the requirements to attribute and share alike." The openness of data is defined by certain key features such as Availability, Redistribution and Universal

Participation. Data must be provided free or at a low cost and conveniently modifiable formats i.e. machine-readable formats for easy handling. The use and distribution of data should be unrestricted to allow flexibility and interoperability with other data sources to make effective use and derive benefits. Universal Participation is a highly important for open data, which calls for no discrimination among user groups in terms of access provided. (*3*)

Data can be an ambiguous term so naturally, the question arises regarding the type of data considered for open publication. This openness is attributed to a category of data which is non-personal and do not create concerns for national security. The ultimate goal of providing data as a free source is to empower citizens to innovate and develop. To understand the concept of open data clearly, we take a look at the history of the Open Data movement.

Open Data Movement

The theory of Open Data has played a larger role before impacting the Transportation Industry. Dating back to the early 1940s, Robert King Merton voiced his support for opening scientific data and research results for the common good. He believed that researchers must relinquish their intellectual property rights and contribute towards the common goal of accelerating the growth of knowledge. The term "Open Data" first appeared in a document released by the National Research Council (Committee on Geophysical and Environmental Data) which called for the exchange of open scientific data between nations to collectively devote efforts in understanding the global environment (*4*).

Moving into the 21st century, openness had entered the software industry. The open source software movement advocated open collaboration of programmers. It supported the exchange of programming codes for software development. This movement led to several innovations in the Internet, an important one being Wikipedia (*5*). The pioneers of the same

movement later met in December 2007 at Sebastopol, California to discuss the concept of open public data and provide a mandate to be adopted by the government. This meeting consisted of many key attendees such as Tim O'Reilly (defined the idea of open source and Web 2.0) and Lawrence Lessig (founder of Creative Commons License) who were well-known faces in the open source software movement. This marked an important moment in the history of open data movement, as it led to the creation of eight principles which define open data as we know it today. (*6*)

Open Public data concentrated on bringing transparency and accountability to the government. On May 9, 2013, an executive order was signed by President Barack Obama, "Making Open and Machine Readable the New Default for Government Information" (*7*). This was followed by the introduction of an Open Data Policy whose subject was "Managing Information as an Asset". This was a key juncture for the open data movement in the United States, an important step towards an open government. By this memorandum, open data is defined as "publicly available data structured in a way that enables the data to be fully discoverable and usable by end users".(*8*)

Following these events, the U.S. Government launched its very own open data portal which holds over 280,000 datasets today. This open data portal hosts datasets across 14 different topics ranging from health to finance and agriculture to education. Inspired by this U.S. Government initiative and policy, many state governments have followed stride and launched their own open data portals. Currently, there are 995 open data portals in the U.S as shown in Figure 1.1. As per the report from the U.S government portal, 48 of these portals function at the state level, another 48 at city/county level (*9*).

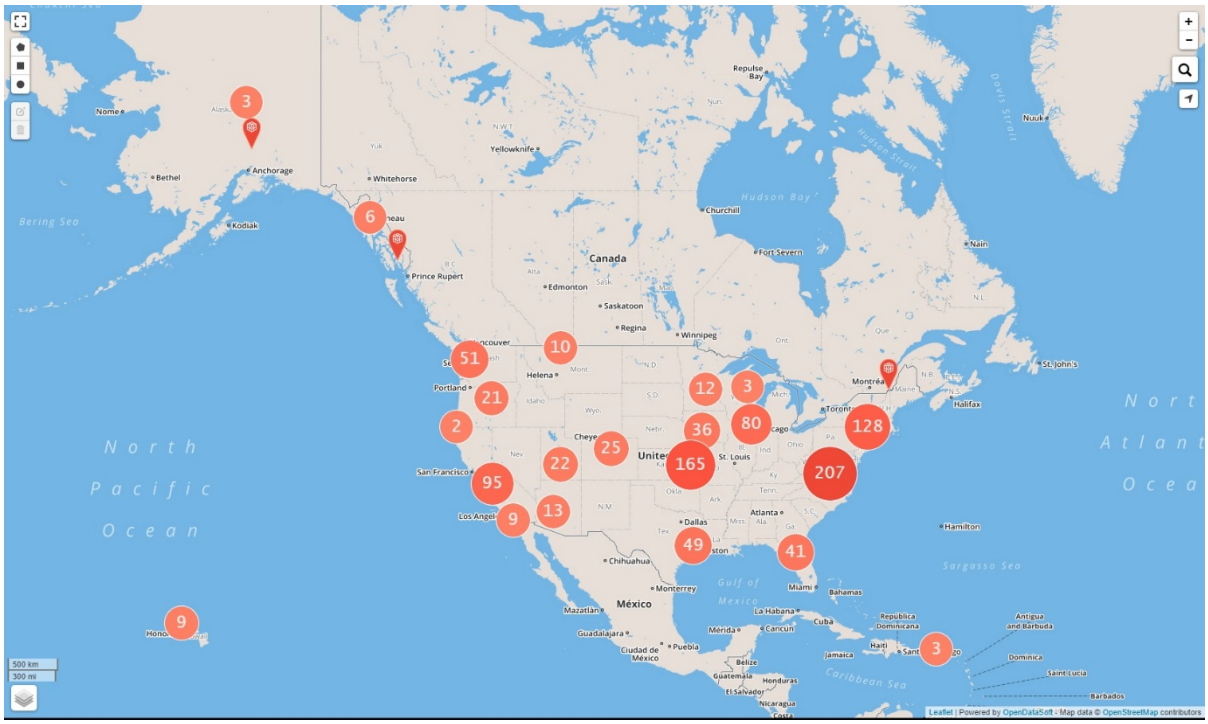

Figure 1.1 Open Data Portals in the United States (10).

Open government data is inclusive of data from different diverse fields such as Agriculture, Health, Finance, Environment, Transportation etc. The government's initiative towards open data was driven with the motive of building trust with citizens through transparency and accountability. Open data is not confined to just promoting a responsible government. Another important aspect of Open data is collaboration which means providing the resource for everyone to work together and contribute towards creating solutions without any constraints. In this prospect, open data for different fields must be studied separately and in detail. Hence, in the next section we focus on open data in the field of transportation.

Open Data in Transportation

Over the years data has grown in volume in the transportation sector. This increase in volume is credited to the use of new technologies such as traffic detectors, tracking mobile and vehicle devices and many infrastructures, environmental and meteorological devices.

Transportation data collected includes a wide range of topics such as traffic volume, crash data, data from different modes of transport such as railways, waterways and airways, intelligent transport systems, pavement conditions and public transit. Open Data movement in the transportation field began with public transit data. (*11*)

North America has seen a boom in open publication of transit data with many agencies reaping the benefits of their efforts. The Tri-County Metropolitan Transportation System of Oregon (TriMet). Massachusetts Bay Transportation Authority (MBTA), Chicago Transit Authority (CTA), Washington Metropolitan Area Transit Authority (WMATA), San Francisco Bay Area Rapid Transit (BART) and New York's Metropolitan Transit Authority (MTA) are the major transit agencies to have published data openly. TriMet was the pioneer of this open transit data publication. All these transit agencies have published data and several Application Programming Interfaces (API) openly along with additional guide and documentation to aid application developers. Most of the applications are aimed at improving trip planning for transit users and have been highly beneficial. These agencies have also contributed to open data publication in their respective city or state portals.

In a survey conducted for the Transit Cooperative Research Program (TCRP) reports (*12*), revealed that 66% of the responding agencies acknowledged that opening their data has improved their agency's perception on transparency and openness. 78% of these agencies agreed that open data initiative has helped to increase awareness among the public regarding transit services in the city. Similarly, public users of transit services were satisfied with the high-quality applications and data now prevalent through open data publication. Hence, open transit data has contributed towards efficient and comfortable travel, improved administration and encouragement towards using public modes of transportation with solid information (*13*).

Currently, there are many State's Department of Transportation (DOT) contributing significantly to providing open transportation data. After the first memorandum on Transparency and Open Government in 2009, many states embraced the initiative of open government and came forward with data to be published openly (Figure 1.2). With the increase in data in the transportation sector, many State's DOTs have launched their own data portals to publish highly relevant open transportation data.

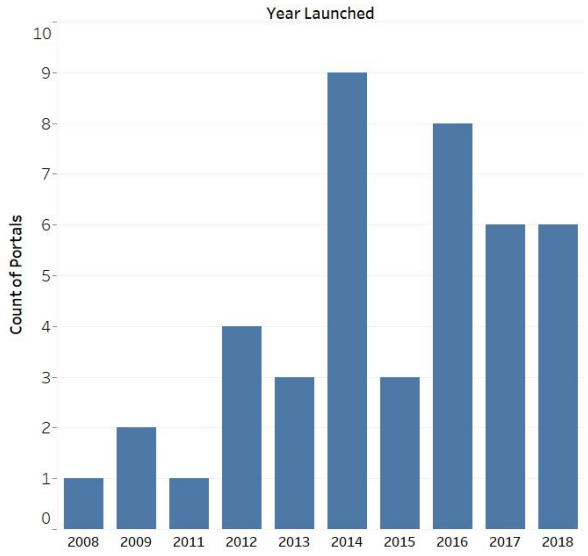

Figure 1.2 Open Data Portals launched each year.

Improvements in data collection techniques have increased the amount of data collected i.e. big data. Handling and analyzing big data can involve complex computations. These computations can be time-consuming owing to the size of input data. However, larger the data, more is the information available to develop better solutions. The size of data collected and processed is constantly increasing. Since, the economic costs involved in collecting, managing and storing these large size data are high, big data is rarely provided as open source. With the government's strong favor for open data, the next step should be

towards Big Open data. Adopting the same for the transportation field can lead to great success. (*14*)

Regional Integrated Transportation Information System (RITIS) data portal developed by The Center for Advanced Transportation Technology Laboratory (CATT Lab) at the University of Maryland is an example of big data services in transportation studied for the thesis. This is a leading platform which gathers and analyzes large streams of road or traffic data. This data portal is not typically open source, but with adequate permissions, access can be acquired without any fee. The data portal provides real-time feeds of incidents occurring on different roadways of the U.S. as obtained from different public and private sector firms. They also run an analytics platform which is highly beneficial to transportation officials, first responders, planners and researchers (*15*).

The quality of data, ease of usage and availability of metadata varies from source to source. To capitalize this open data publication it is necessary to identify and understand the cause of this variability. Across different portals it is important to analyze the key features such as categorization of data, visualization of data, its spatial and temporal characteristics, API guide and tools, completeness and accuracy of data. Minimizing the variance and standardizing the format and design of open data publication can thus be beneficial.

Thesis Objective

With little guidance, many DOTs and agencies are left with questions about what an open data portal is, what transportation data can they offer, how user-friendly are these web portals and is this openness leading towards targeted innovations and discoveries. There is an imminent need to assess the quality of open data portals, to provide agencies a yardstick to measure their performance and draw inspirations from higher ranking portals. To achieve this, we aim to evaluate the quality of open transportation data portals and data content.

Quality assessment is a subjective task which can be best handled by a scoring rubric. A scoring rubric is a clearly defined scheme which can provide an easy performance assessment and meticulous description of expectations for better performance. Hence, we propose a data portal evaluation rubric (DPER) which can serve this purpose. DPER is designed to capture the essence of the National Open Data Policy. It is then used to evaluate 43 data portals at the state and national level which provide transportation datasets. The DPER constitutes of three levels of evaluation. The overall score calculated for every portal is based on three categories, Portal Usability, Data Information and Content Relevance to Transportation. Each category of the rubric is described by several parameters which best highlight its essence. Each of these parameters are described using features observed in portals and their data during the course of the study. The weights for the rubric have been designed based on a feedback survey circulated among several open data publishers.

In the process of this evaluation, each portal was observed in detail across different factors such as ease of usage, accessibility of data, data formats, license information etc. All this information denotes the variability that exists among different data portals although they all aim at achieving the same goal. To highlight these differences and assimilate this information in one place, we also created a visualization tool using Tableau software. The tableau visualization creates a repository of open data portals providing significant transportation datasets. It highlights the features provided in these portals and their ranking as per the DPER. It serves as an effective tool to compare the features available in these different portals.

CHAPTER 2. LITERATURE REVIEW

In the previous chapter we reviewed the principles of open data, key events of the open data movement, its impact on the government and transportation industry. We also discussed the need for evaluation and the thesis objective.

In this chapter, we reflect upon the literature essential in designing the different elements of the rubric. The National Open Data Policy created by the United States Government lays down the definition of open data. We analyze the open data initiative at different transit agencies to identify crucial factors which enabled their success. There are several evaluation strategies employed in literature for assessment of open government data whose pros and cons are discussed. Lastly, we discuss the aptness of a rubric to this problem and steps in designing a flexible and valid rubric.

National Open Data Policy

The National Open Data Policy (*16*) enumerates the standard to be upheld for openness and also highlights the importance of information and open government. It defines Open Data and its principal qualities which are Public, Accessible, Described, Reusable, Complete, Timely and Managed Post Release.

The U.S. Government encourages agencies to publish open data adhering to the Open Data Policy. The policy aims at creating open data with machine-readable formats, compliant to standards, open-licensed, and common and extensible metadata. The information published should be flexible and interoperable for use in an interface. Standard practices should be followed such as maintaining a repository and managing feedback post-release of data. The privacy and confidentiality of citizens must not be harmed due to the nature of the data released.

Huijboom and Van Den Broek (*17*) highlight the propaganda behind open government data publication which is to strengthen citizen engagement, encourage innovative business and enhance law enforcement. This research, aimed at analyzing the strategy and activities adopted by five different countries namely the United States, United Kingdom, Spain, Denmark and Australia towards open data. The results of the analysis explain the key elements responsible for the progress and failure of Open Data. Some important progress factors identified in the U.S. include Strategies, Regional Initiatives, Citizen Initiatives and Emerging Technologies. Following the above factors, the U.S. government has actively organized Hackathons and developed web portals and applications to improve citizen participation. The Open Data Policy has been a progressive strategy in shaping the benefits and impacts of Open Data.

Open Transportation Data

The research background for Open Data in transportation is currently limited. Till date, the prime focus has been on Open Government Data. As discussed previously in the Introduction chapter, transit agencies were the first to publish data openly. Rojas (*18*) reviewed open data initiatives at five transit agencies across the United States. The strategies, initiatives and consequence of publication of open transit data by these agencies was compared in depth. Their transition from a closed to an open system was not easy. Preparing the data in machine readable formats was an arduous task which was aided by the intelligent technology available. Crossing all barriers these agencies provided open access of data and Application Programming Interface (API) to the people. In return, developers took to generating many mobile and web applications for transit users. This section discusses in detail the open data initiative across the major transit agencies which have contributed several open datasets to their city or state portal counterparts.

In 2005, The Tri-County Metropolitan Transportation District of Oregon (TriMet) was the first agency to publish their data openly. It began collaborating on open data with Transit Surfer and Google Transit apps. The results of this partnership were highly beneficial. Today TriMet website hosts 49 mobile applications such as PDX Transit Map, Rail Bandit, Roadify etc covering different modes of transport. It holds the highest number of software applications developed second in place to New York City's Metropolitan Transit Authority (MTA). (*19*) (*20*)

The arrival of several navigation apps made it easier for the public to acquire driving direction but still had difficulty in accessing information regarding transit schedules and routes which directly affected the public transit ridership. With the release of transit schedule data in open and non-proprietary formats (CSV), it created an opportunity for developers to use the data to create effective transit applications for public use. An increase in transit applications made available instant and effective information which encouraged the public to use public transit modes. This event was crucial in the creation of the General Transit Feed Specification for standardized publication of open public transit data. (*21*)

Massachusetts Bay Transit Authority (MBTA) was an early adopter to publishing open data taking cues from the TriMet's open data success. MBTA had invested efforts for providing transit users with for trip planning and arrival information systems. However, they were reluctant in creating an open system. The change came in 2009 brought about by two of their employees who were strongly driven by the open data movement at TriMet. MBTA released data of five bus lines of the 200 they had running. In response to the data release the saw many developer come forward with exciting applications in predicting bus arrivals. This boosted their commitment and in 2010 data from all MBTA system was published openly.

The authorities realized the benefits of this action and kept forth to encourage developers and involved them in meetings, hackathons and conferences. MBTA set forth a great example for other transit agencies to start their open data movement. Expanding their efforts, OpenMBTA website was created which hosts several open source tools developed for transit trip planning (*22*).

MTA is the largest public transit authority in the United States and one of the largest in the world (*23*). In 2010, MTA's chairman Jay Walder initiated the open data movement with an open data policy (*24*). The policy led to the creation of a web portal which published MTA data with open access to all. The web portal consists of schedule, route and fare details published as open, accessible and updated data. This initiative improved the MTA's ratings significantly. It led to the development of nearly 80 applications which are equally effective to different groups of people. MTA has revised its website adopting a standard approach to offer an array of real-time data which are not widely available. (*25*) . MTA was one of the forerunners in contributing data for open publication to the State of New York open data portal launched in 2013 (*26*).

The Chicago Transit Authority (CTA) is the second largest public transportation system in the United States (*27*). CTA's move towards open data publication came after a citizen built API led to better and efficient mobile applications for transit users to track arrival timings. In 2009, CTA released its official Bus Tracker API and provided enormous support to developers in adopting them to develop applications. By 2010, a developer's center page was launched with detail documentation on published API's and a guide for using them (*28*). As of today, CTA hosts JSON versions of three different APIs (CTA Train Tracker, CTA Bus Tracker, Customer Alerts API) which have been utilized in created 22

mobile and web applications to aid planning of transit users (*29*). Like MTA, CTA has offered several datasets for open publication in Chicago City Data Portal (*30*).

BART emerged as a National Leader by openly publishing real-time transit data feeds. Located in downtown Dan Francisco, across the bay, consists of 48 stations along six routes of rapid transit lines (*31*). BART had the advantage of owning all the Automatic Vehicle Location (AVL) data which it published openly to aid application developers (*32*). Almost immediately transit planner applications were developed. BART invested more efforts and encouraged competition among developers to utilize their open data to provide the best service to customers. BART was ardent in its open data efforts as it published data in open Google's General Transit Feed Specifications (GTFS) formats. It also helped create the real-time feed standard extension to the GTFS. Currently, BART offers three official applications with several features such as trip planning, real-time departures, contact to police services and airport connectivity (*33*). In 2018, BART took the initiative further and introduced an open data policy and a web portal. This web portal hosts over 50 datasets across 12 categories such as Economy, Environment, Finance, Ridership, Safety, Performance, Workforce etc (*34*).

Washington Metropolitan Area Transit Authority (WMATA) faced two major issues in its efforts to open data. The agency's willingness towards open data was low and complete, accurate data was not available for publication. Finally, the push from the citizens demanding transparency and access to what is rightfully their data made the WMATA fix the challenges with data and publish them openly. In 2010, WMATA released public API for its Metro Rail data and real time positions of all metro buses. However, WMATA did not reap many benefits from this efforts has it failed to encourage the most important stakeholder in

this initiative in this movement, the application developers. In the Washington DC region, local governments have collaborated with developers and set up real-time transportation screens for businesses to use and relay timely updates on transit facilities (*35*).

Transit agencies were the first in the transportation industry to dive into the open data movement. This also because the transit agencies have direct impact in the public's daily life. These case studies are perfect examples of the benefits to be reaped from open data publication. An important aspect to these open data stories is that their efforts were not restricted to just publishing data. The Open Data Movement does not confine only to accessibility. Accessibility is rather the first step. The larger focus lies in creating an environment and providing the right resources towards effective use of data. This is evident both in the success of MBTA which encouraged developers greatly and the loss of WMATA which failed to do so.

Today, as many DOTs and other state agencies have started to expand their open data base, these case studies offer key points to keep in mind. Adapting a similar model as the transit agencies many DOTs host websites to publish data and API for developers. These efforts will be fruitful only when the resources provided are relevant. Publishing complete, accurate open data is highly important. Maintaining contacts with developers and customers to understand their needs in updating the open data system is essential. Hence, these features were added as criteria for evaluation of data in the DPER.

Prevalent Methods of Evaluation

There exists another dimension of research work that focuses on the evaluation of open datasets. Some standard methods include Tim Berners-Lee 5 Star Open Data, the Open Data Barometer (ODB) and the Global Open Data Index (GODI). The 5 star method provides a simple five scale evaluation which can be used to rate open data described in Table 2.1(*36*).

Table 2.1 5 Star Open Data

Stars	**Scale**
*****	Link your data to other data to provide context
****	Use URIs to denote items within data
***	Make data available in open non-proprietary format
**	Make structured data available
*	Make data available online under an open license

The ODB is a methodology developed by the World Wide Web Foundation which analyzes open government data and scores them on a scale of 100. It analyzed the data based on its readiness for initiatives, implementation of active programs and the impact created. This assessment system is dependent on three surveys. There is a government self-assessment survey highlighting their efforts for the progression of the open data movement. A peer-reviewed expert survey is conducted to understand the current scenario open data in specific countries which is then peer-reviewed and scored. There is also a dataset assessment where the presence of 15 categories of data is identified in each country. Further for each category, there are 10 detailed questions to assess their quality. There is also a secondary data survey conducted to evaluate the readiness of open government. Each of the variables are normalized as per weights described in Table 2.2 and an overall score out of 100 is calculated. (*37*)

Table 2.2 Open Data Barometer

Readiness (35%)			
Government Policies (1/4)	Government Action (1/4)	Entrepreneurs & Business (1/4)	Citizens & Civil Society (1/4)
Implementation (35%)			
Accountability data cluster (1/4)	Social Policy dataset cluster (1/3)	Innovation dataset cluster (1/3)	
Impacts (30%)			
Political (1/3)	Social (1/3)	Economic (1/3)	

GODI is an independent assessment which is calculated by conducting a survey to understand the feedback of users and owners. There are four key assumptions in the analysis carried out. The open data evaluated must be defined according to the 'Open Definition' of Open Knowledge International. The government must be in the cardinal role of publishing data. The GODI is indicative of the open data publication at the national level. GODI is not country specific in order to address the open data efforts of government bodies at sub-national levels. The survey results are analyzed to understand the shortcomings and to highlight the progress compared to other parties in the world. Based on the overall score, places are classified as open data (100%), public data (up to 80%), access-controlled data (up to 85% but limitations to user) and data gaps (0%). (*38*)

Susha et al. (*39*) draws a comparison by evaluating the popular methodologies for benchmarking open data. The methods evaluated include Open Readiness Assessment by World Bank, ODB, GODI, PSI scorecard by European PSI platform and Open Data Economy by Capgemini Consulting. The methods were chosen based on similarity found in their scope and accessibility. However, they had significant differences in the features of the

open data that they evaluated. In conclusion, the existent benchmarks are specific to a purpose and need to be validated to improve the quality of evaluation.

The existent methods of evaluation suggest that a rubric evaluation is an effective way to assess open data. Each of the described methods have a similar framework although they differ in features studied and the scoring design. Hence, a rubric is considered a competent method to evaluate the quality of open data. A rubric essentially allows the evaluator to arrive at subjective criteria to be met by the assessed work. It converts this subjective criterion to an objective benchmark by using a precise scoring guideline. It is used when there is a need to identify the extent to which a benchmark is met. A scoring rubric is highly apt for evaluation when the work is graded to meet a certain benchmark and provide an assessment for its improvement. (*40*)

Designing a Rubric

Perlman (*41*) describes steps for performance assessment using a scoring rubric. This paper defines two types of scoring rubrics, analytical and holistic. It clearly elucidates the steps in choosing the type of rubric, labels and scoring scale. It highlights the importance of choosing the appropriate length of a scoring scale. The scoring scale should delineate the differences in the feature. It should not be extremely short or long such that the differences are neglected or inadequately described. Based on the importance of each feature evaluated, the labels can be equally or unequally weighted. Rubrics can be selected, modified or created newly pertaining to the objective. This paper poses significant questions to be answered in designing an effective rubric for evaluation.

An effective rubric needs to be valid and reliable. Moskal (*42*) defines validity and reliability and their role in designing an effective scoring rubric. Validity defines the aptness of the results of the evaluation to the predetermined objectives. A valid scoring rubric is one

where the labels are clearly defined based on the objectives of the evaluation. Reliability is defined as maintaining uniformity in the scoring criteria such that results are not influenced by different evaluators. It is necessary to achieve this consistency for obtaining stable results which relate to the objective of the evaluation.

A notable research by Thorsby et al. (*43*) focused on both the portal and data contents for 37 city level Open Data Portals using five indices. A checklist evaluation was employed to calculate two indices namely an Open Government Data Portal Index and Dataset Content Index. The others indices include an Overall Index, Number of Datasets and Number of Datasets per a population of 100,000. Based on their evaluation, the team hypothesized six different factors which may impact the evaluation indices. The population of a city had the largest impact while a regional consortium had a limited impact. The other factors identified which showed no impact on the evaluation indices were level of education, type of government (open or closed), the degree of innovation and the age of the portal.

Data Portal Evaluation Rubric

Influenced by the existent research, a data portal evaluation rubric (DPER) has been designed using the features listed by National Open Data Policy to evaluate the quality of open transportation data. The focus of the DPER lies on three different aspects of an open data portal. Similar to the OBD the DPER can be used to calculate an overall score for the portal out of 100. This score is a reflection of the performance of the portal in terms of Portal Usability, Data Information and Relevance of Content to Transportation, DPER is a weighted rubric where the different rubrics are assigned weights based on the thesis objective.

Each of the categories are further broken down into several parameters. Parameters are abstract attributes representing a category. Based on review of literatures and observation of current portals, parameters such as Data Formats, Application Developer Tools, and Legal

Provisions etc were chosen. To create a valid and reliable rubric each of these parameters are broken down to simple features such as search bar, graphical representation of data, licensing, update frequency of data etc are either present or absent in a portal. The final category of Content Relevance consists of different topics (transportation) similar to GODI, OBD or the Thorsby et al. paper. The scale of content relevance is set such that more the topics covered more the score obtained by the portal.

After the selection of parameters and features for the rubric, it was essential to assign reliable weights to the rubric. To this aspect, we adopted the analytical hierarchy process (AHP) approach and involved the input from transportation professionals in identifying the importance of the criteria for evaluation among one another. We conducted a survey using the AHP format where in participants were asked to compare the importance of different parameters towards the category of scoring. Based on the responses received, we calculated the weights for the parameters and the subscoring categories. Hence, the input from key stakeholders in open data also influenced the design of the rubric.

DPER is a hierarchical rubric with three levels, category, parameters and features. The scores calculated at each level are normalized to ensure flexibility and extensibility of the portal. For example, if the DPER was to be expanded for a particular parameter, new features can be easily added as the cumulative of features is normalized to a parameter score. Also, the DPER is simple to design and contains many parameters which are important to open data portals in general. Hence, it can be easily adapted for open data portals of different fields. Considering both the portal as well as data contents, the rubric focuses on evaluating the user interface to bring about the desired impact of open data publication in transportation.

CHAPTER 3. DATA DESCRIPTION

In this chapter we discuss the different open data portals studied. Open Data Portals are websites which serve as a platform for open publication of data. We focused on transportation data and the involvement of state agencies in publishing these data. A total of 43 portals were identified, 39 of these portals are maintained by the respective State Department of Transportation (DOT) or other state agencies. US Government Open Data Portal, US DOT Open Data Portal, Bureau of Transportation Statistics (BTS) Open Data Portal and Regional Integrated Transportation Information System (RITIS) Data Portal are the four national portals included in the study (Figure 3.1). These portals vary significantly in terms of the design of user interface, features offered, topics of transportation covered and publishing agency.

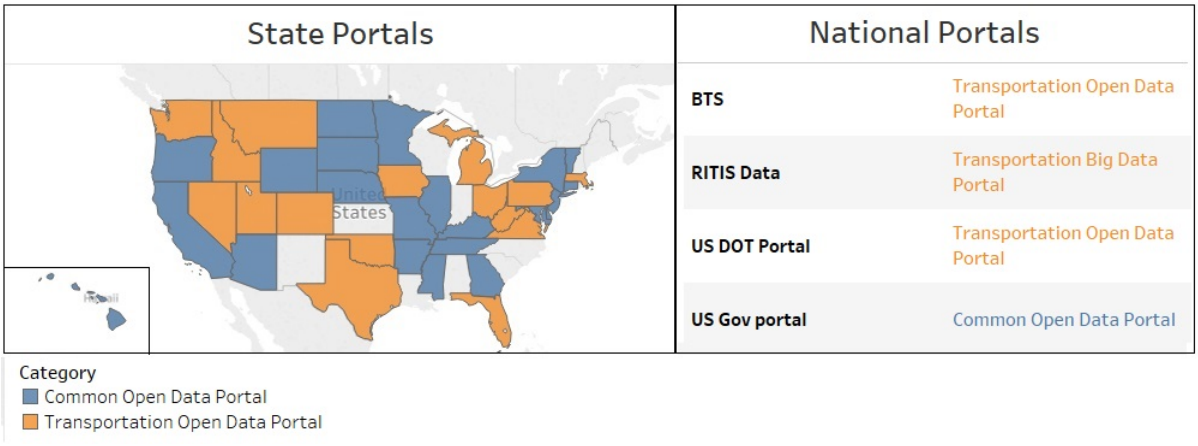

Figure 3.1 State and National Portals Studied

Common Open Data Portals are launched by state and national governments as a platform for publishing large number of datasets. These portals offer all kinds of data such as Agriculture, Education, Energy and Environment, Finance, Public Safety, Transportation and Human Services and 24 such portals are included in the study. In contrast, there are 19 open

data portals which are launched by the Department of Transportation agency at the state and national level. These portals cover datasets pertaining to transportation only. (Figure 3.2)

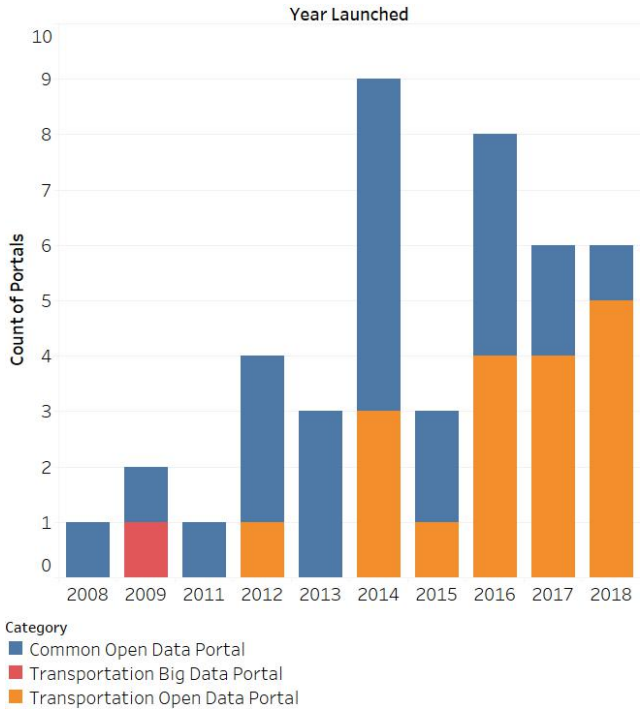

Figure 3.2 Count of Open Data Portals by year of launch

The design variability is due to the different developers who aid these agencies in developing the online data portals. Although there are many different developers, majority of these portals are designed by Socrata, ESRI, CKAN or DKAN. These developers are typically software companies who assist agencies in establishing open or public data services. ESRI has developed the most number of portals owing to its prior contribution with geospatial data publication for these agencies. There are a few portals who have been designed by the data publishing agency itself which are categorized as others. (Figure 3.3)

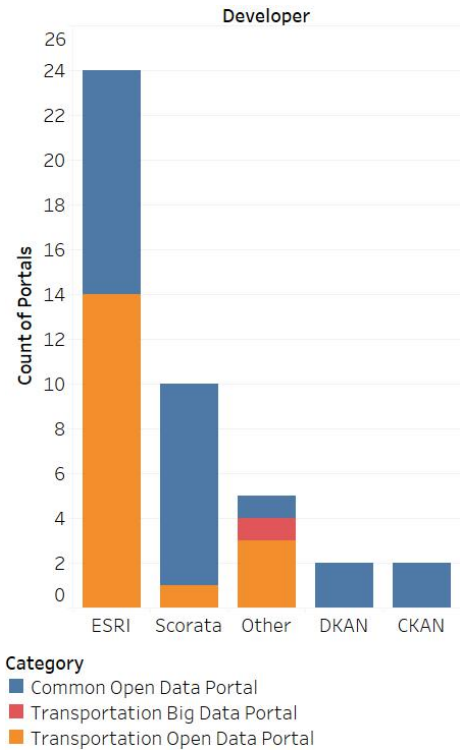

Figure 3.3 Count of Open Data Portals by developers

Transportation in itself a huge field covering several topics. The number of transportation datasets offered in each of these portals varies significantly (Figure 3.4). Roadways, Traffic Violations data and Transit data are the most widely available types of transportation data. In contrast, weather updates on roadways, data on parking facilities and freight data are scarcely available (Figure 3.5). There is a need to standardize the data content in every portal, as datasets though widely present, lack uniformity. For example, Iowa DOT provides many datasets from the Road Asset Management System describing different features such as medians, curb lines, shoulders etc. In contrast, the Arkansas GIS repository provides only route information. Authorities should focus on standardizing the categories of data available to the user.

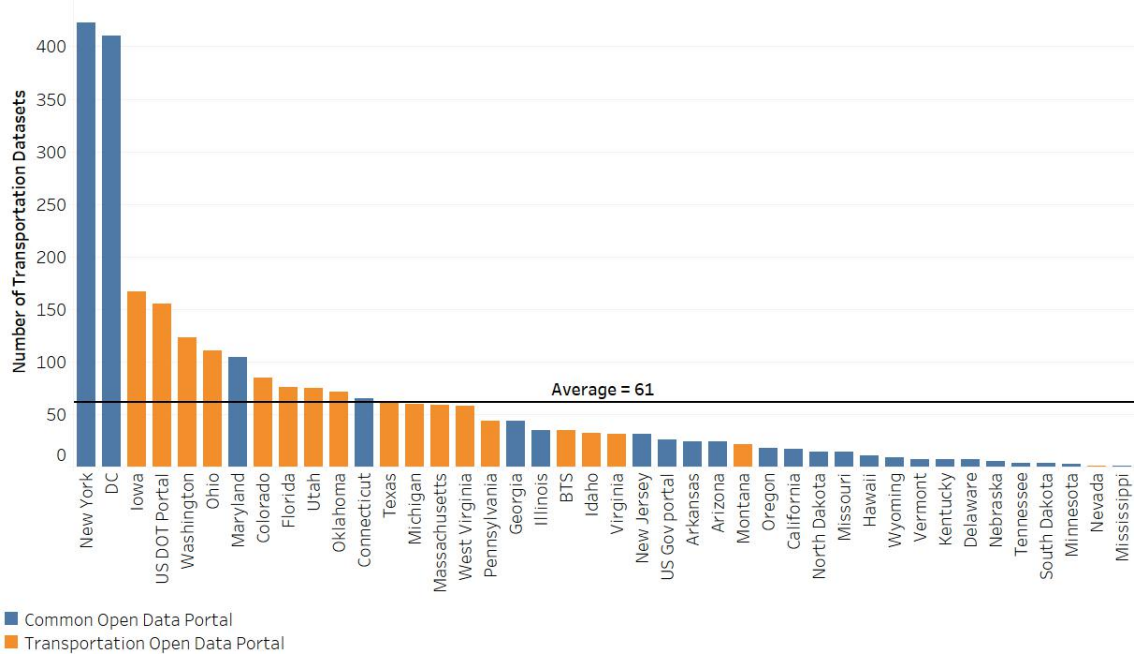

Figure 3.4 Number of Transportation datasets in each portal studied

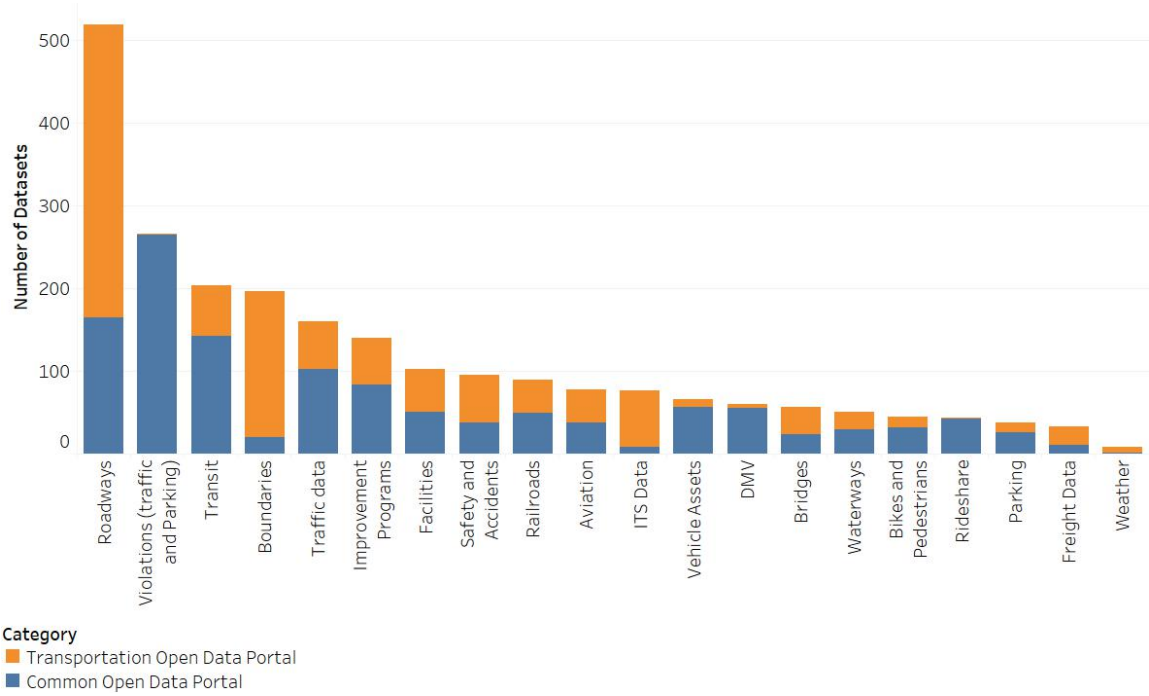

Figure 3.5 Total Number of datasets across all portals in each transportation topic

Most of the open data portals available today offer small size datasets as observed in the portals studied. New York State portal offers a few datasets of large size ranging from 1GB to 12GB. Similarly, the US DOT Portal offers large datasets pertaining to the recent trials conducted in connected vehicle environments test bed around the U.S. RITIS is a big data portal which collects data from various sources, streams them real-time and uses the same data feed in developing several awareness tools for application on roadways. Its design and structure are slightly different from the other portals studied as the system is well built for ingesting more than 6 billion real-time streaming records per day and analyzing them in real-time to help planning agencies with crucial decision making (*44*). (Figure 3.6)

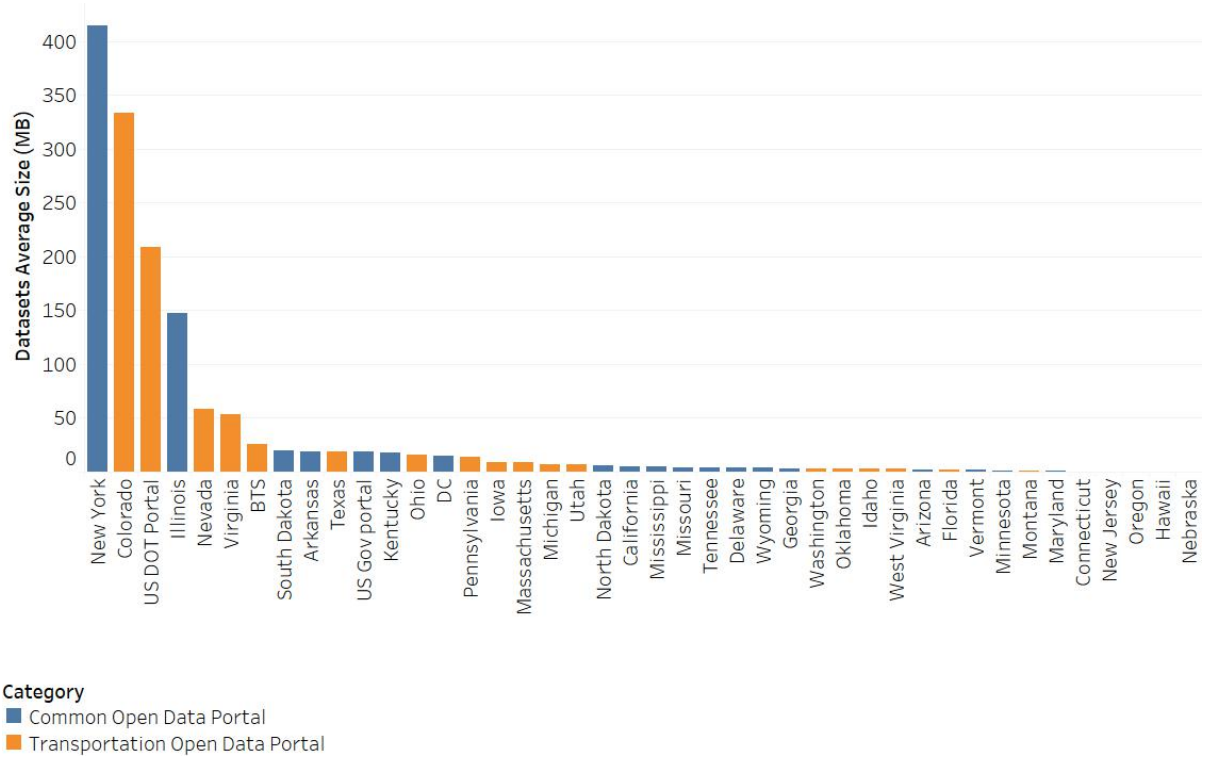

Figure 3.6 Average Data Size in each portal studied

CHAPTER 4. METHODOLOGY

In the previous chapter we discussed the different portals studied. There are mainly two types of open data portals with respect to the content they offer i.e. common open data portals and transportation open data portals. These portals serve at the national or state level and differ significantly in design due to different developer agency. In the chapter we focus on the DPER which is the methodology to evaluate the portals studied. This chapter will discuss in detail the categories, parameters and features described in the DPER. It will also highlight the scoring schema adopted.

With the identified open data portals, the DPER was designed to evaluate the quality of open transportation data. The objective of the DPER was to evaluate the usability of the data portal, the openness and details of the data available and the relevance of this content to the transportation community. The rubric was also designed to serve as a guideline for any agency which would be developing an open data portal.

The scoring rubric is designed to calculate three subscores, one each for the portal, data content and its relevance to transportation respectively. The final overall score is calculated by the summation of the weighted subscores designed based on survey feedback from data publishers. The three divisions of evaluation were chosen as they are inclusive of the medium and the composition of open data which can highly impact the open data movement. As a flexible rubric, evaluator can adopt different weighing patterns to assess data in a manner closer to their objective. For each subscore, the rubric is designed to assess its various characteristics.

Evaluation of Relevance of Data Content

The first step of evaluation is to calculate the transportation relevance subscore. After reviewing the categories of datasets across portals, a list of twenty topics were shortlisted. The shortlisted topics depict a category of data which is relevant to the transportation community. Each topic includes an array of subjects whose availability varies with each portal. If data present in a portal relates to any of these topics, then it can be stated that the portal offers data relevant to the transportation community. For every topic covered the portal is awarded one point. The points obtained for a given topic is normalized to a scale out of five.

The normalization exists to provide flexibility to the rubric. Each portal covers different aspects of the same topic described in the rubric. In future, we can study the level of importance of these different subtopics, in which case this rubric can be modified to include the same as a scaling feature. The presence of different sub-topics can be scored and normalized to a scale out of five for final subscore. With twenty topics each scored out of five, their cumulative results are calculated to a score out of hundred. The topics shortlisted and their respective datasets found under each of them is described below.

Aviation

Popular aviation datasets relate to airports and runway information. Airports information include arrivals, departures, airline capacity and facilities. Airfare information is also provided in some portals. Zone data with flying restriction may also be available.

Bikes and Pedestrians

Bikers and pedestrians are important users of the roadway. Information on bike stations, availability of bikes, bike routes and traffic counts are published openly. Pedestrian data include locations of pedestrian signals and traffic counts.

Boundaries

This data consists of city, county, district or other demarcation in the state. Boundary demarcation as recorded by regional agencies such as Metropolitan Planning Organization (MPO) are also published.

Bridges

There is usually an inventory of bridge data consisting of bridge locations. Apart from locations, current conditions and required maintenance of bridges is also published as open data.

Department of Motor Vehicle (DMV)

Data related to the DMV include license registered, DMV office locations and various facilities licensed by the DMV.

Facilities

Locations of rest areas and gas stations are published openly. Utilities provided at the rest areas such as restrooms, Wi-Fi, food purchase, telephone, and pet exercise are also provided. Roadside signs, lamp posts and signal locations are provided.

Freight

Data related to freight volume, corridor, and type of freight facilities such as warehouse, fuel plants or grain processing facilities are available.

Improvement Programs

Data on improvement programs highlight the future plans or the current plan underway for improving pavement, structures, bridges or roadways.

Intelligent Transport Systems (ITS) Data

This includes data from various traffic devices such as CCTV, dynamic message signs and highway advisory radio. Trajectory data from US DOT projects are also published

openly. Results from recent tests on connected vehicle environment test beds from different cities in the U.S. are published openly by the U.S. DOT.

Parking

Open data on parking spaces are provided as park and ride lots, parking fares, parking meters.

Public Transit

There are various transit authorities publishing open data in the portals. Common transit information published include schedules, routes, stops, fares and ridership details.

Railroads

Railroads information consist majorly of routes of rail lines, locations of railroad crossings, and dataset available relating to the ridership on trains which provides a passenger count of rail line travelers.

Rideshare

Rideshare is a popular mode of transport, developing as an alternative to public transit. These datasets cover trip information published year wise. These datasets provide non-personal information such as drop off locations, pick up locations, trip distances and fare rates.

Roadways

Roadway data includes information about all classes of roads such as highways, major and minor roads, local roads and ramps. Road geometry data consists of median width, shoulder width, shoulder curb and the number of lanes. Route information for different roads is available. Surface conditions of the road measured in terms of roughness index and friction index provide data on pavement condition. Information on scenic roads and trails will also be

covered under this topic. These subtopics consist of the popular roadway data published openly.

Safety and Crash Data

Most portals provide archived crash data from previous years. Crash data specific to vehicle type, travel mode or location may also be published. Safety structures such as barriers, guardrails, rumble strips are also reported as open data.

Traffic Characteristics

Open data on various traffic characteristics such as traffic volume, delay and signal timings are published. Additionally, locations of traffic signals, signal cabinets, traffic poles and traffic count stations are also published.

Vehicle-related information

This topic includes data on vehicle miles traveled, travel times. Origin-destination counts, vehicle assets such as vehicle type, sustainability measurements, freight vehicle miles traveled, parking location of trucks, routes for trucks and electric vehicle charging stations.

Violations (Parking and Traffic)

This topic covers tickets issued for parking and traffic violations. This data is archived year wise and available with Incident type, ticket type, incident location, date and time information. These datasets do not contain any person data of the violators, instead provides account of violations recorded.

Waterways

There are very few datasets pertaining to waterways available concentrating on water taxis, ferry routes, and ports.

Weather

Weather updates from roadways are published as open data. This is popular among states with severe winter conditions where many applications have been developed using them.

Evaluation of Data Portal

The platform for providing this data is important as it greatly improves the usage of data. This is evident in the efforts made by different developers in creating a data portal which aims at providing a convenient experience for its users. Hence, the next step of the rubric focuses on the usability and functionality for end-users. Each of the parameter listed have a varying scale which will be normalized to a scale out of five for uniformity. Once all parameters have been evaluated, the total score is normalized to a score out of hundred with survey based weights for each parameter. The parameters evaluated are discussed below.

Ease of Usage

This category aims to evaluate the convenience of using the data portal. Search bars are an important tool which enables a user-friendly data portal by saving time and helping the user easily find data. Hence, the presence of a search bar earns the portal one point. Also, if the data portal can be easily discovered by a user, the usage increases. Therefore, the next feature is the number of clicks (navigation steps) taken to reach the data portal from a Google search. To design a rubric for this feature the number of steps for reaching a portal was recorded. From the results obtained, two steps was decided as the threshold value to calculate the score. If the number of steps was less than two, the portal received two points. If the number of steps was equal to two, the portal received one point. If the number of steps is greater than two, the portal received zero points. (Table 4.1)

Table 4.1 Scoring Design for Clicks to reach Portal

Number of Clicks to reach Portal	Score (Points Awarded)
> 2	0
= 2	1
< 2	2

The categorization of datasets allows the user to narrow their search pool. If the transportation data were categorized into groups such as roadways, traffic data, crash data, etc. the portal received one point. If the portal provided video or document tutorials for the user to understanding navigation through the portals and accessing the datasets, the portal received one point.

Accessibility

This refers to the different forms by which data can be accessed. Data can be previewed at the website or downloaded in various formats. Some portals also provide tools to filter datasets with some including an additional feature to download filtered datasets which will include only the required data. For preview, download and filter download features present, the portal is awarded one point each. If links to external websites are provided, the portal receives one point. These external links usually provide further information on data source or collection technique.

Interactive Visualization

Data visualization is an important aspect of analyzing data. This can be achieved through visualization tools such as maps, bar charts, pie charts or line charts. Representation of data using maps clearly demarcates the jurisdiction of data. If portals offer interactive geospatial maps (geohash) for visualization it is rewarded one point. It receives another point

if graphical representations of data is possible (Figure 4.1). Interactive tabular representations of data is given an additional point.

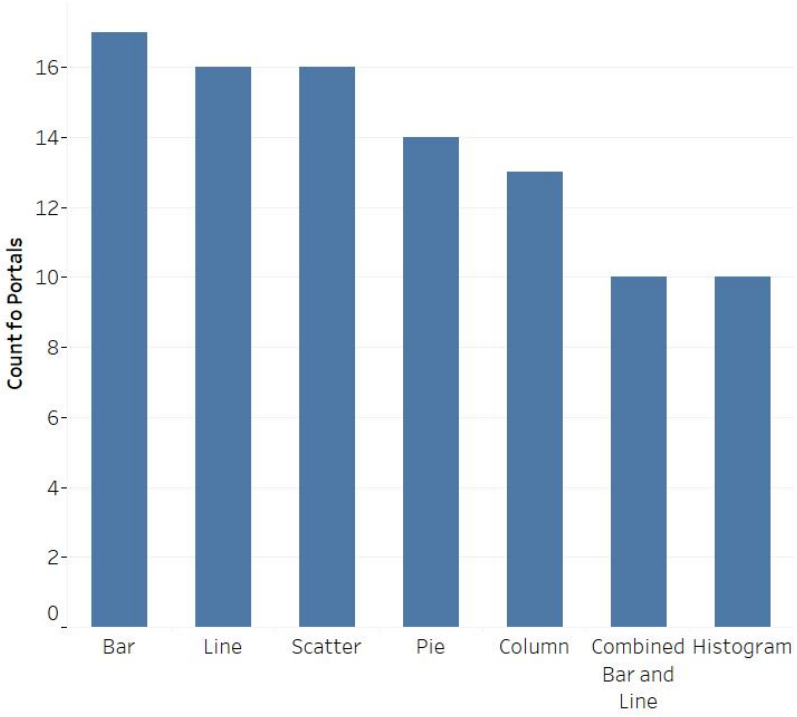

Figure 4.1 Count of Portals with specific Graphical Representations

Statistical Tools

The option to filter the data helps the user focus on specific data attributes that suits their interest. The ability to quickly provide descriptive statistics such as mean, mode and median would be an asset as they are the quickest way to analyze data. For the presence of each of the above-listed tools, the portal is awarded one point each.

Application Developers Tool

The biggest advantage of the open data portals is for developers, researchers and others who use this data to foster innovation. An important tool for this is a well-documented portal which has Application Program Interface (API) guide as well as an API Query tool which provides a platform to raise queries, filter and aggregate data. The portal is

awarded one point each if the above tools are present. To understand the impact of application developing tools, the number of applications developed using this data has been measured. A scoring design was created using the recorded numbers with a five-scale evaluation (Table 4.2). Frequency Distribution Statistics was employed to design the class intervals for the scoring design. As a result, if the number of applications developed are high, the points rewarded also reflect the same. Another area of use of open data is research, due to the lack of clarity in identifying this use, this parameter was not added.

Table 4.2 Scoring Design for Number of Applications Developed

Range of Number of Applications Developed	Score (1-5)
1-5	1
6-10	2
10-15	3
15-20	4
21-25	5

The Number of Transportation Datasets

The use of this parameter is to highlight the need for more transportation datasets which would drive the publishers to provide more transportation data. To design the scoring scale, the total count of transportation datasets in each portal was recorded. This value ranged from numbers as small as 1 to numbers as large as 423. Frequency Distribution Statistics was employed to design the class intervals for the scoring design. (Table 4.3)

Table 4.3 *Scoring Design for Number of Transportation Datasets*

Range of Number of Transportation Datasets	Score (1-5)
1-85	1
86-170	2
171-255	3
256-340	4
341-425	5

Feedback

Active participation is encouraged and providing the ability to give feedback or comment sections will allow users to critic the datasets or suggest additional categories of datasets. If a comment section or an email for contact is provided, the portal is awarded one point.

Evaluation of Data Content

The last category of the rubric focuses on the data provided by the portals. The Open Data Policy is a concrete and comprehensive document which clearly highlights the prerequisites of open data. These parameters are designed referring to the National Open Data Policy. Each topic described below is normalized to a scale of five to maintain uniformity. The total score is then normalized to a scale out of hundred based on weights from feedback survey. These parameters are discussed below.

Data formats

Data must be published publicly without any restrictions and should be available in accessible and non-proprietary formats such as CSV, KML, SHP, XML, PDF, XLS etc

Figure 4.2). Across portals 24 unique data formats were identified (APPENDIX A.). Since, this is a data specific parameter, points are awarded to the portal based on the overall accessible data formats. The total number of data formats found accessible in the portal datasets is recorded. This value is converted to a score using a scoring design based of Frequency Distribution Statistics mentioned in Table 4.4.

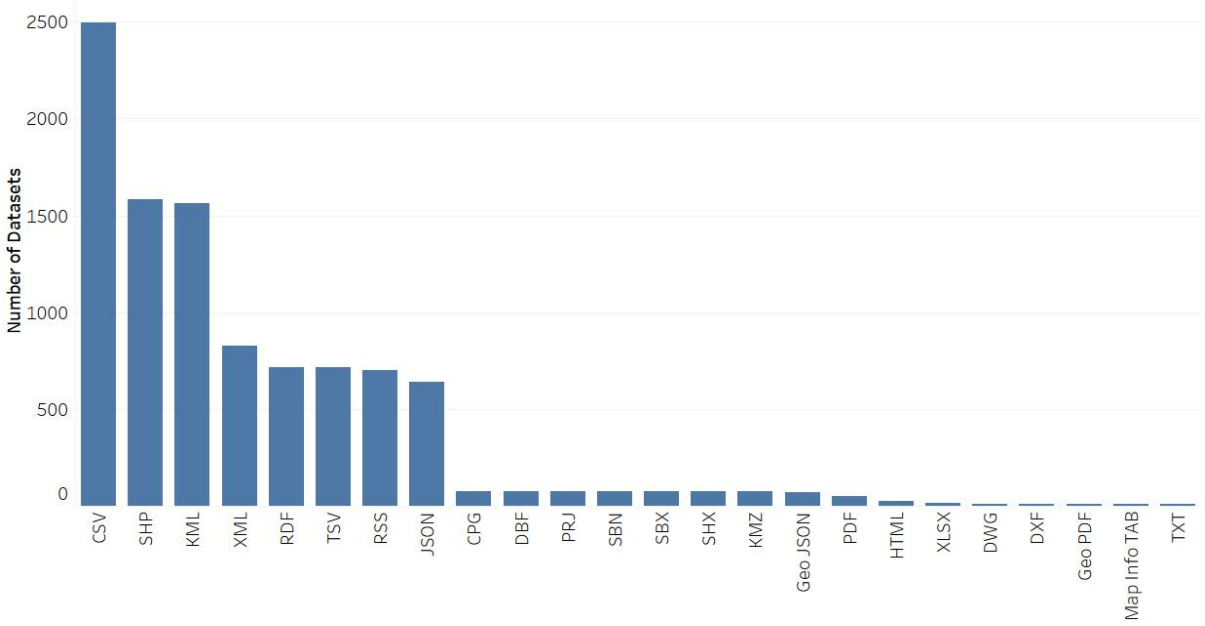

Figure 4.2 Total Number of datasets in each data format across all portals

Table 4.4 Scoring Design for Data Formats

Range of Number of data formats	Score (1-5)
1-3	1
4-6	2
7-9	3
10-12	4
13-15	5

Data Description

Data should be described clearly in terms of the attributes, metadata and details of data owner or publisher. Data dictionary clearly defines the attributes and the values it can assume. This specific information must be provided uniformly for every dataset in the portal to provide a clear understanding to the user. Metadata is a document that describes the data. A metadata covers various topics such as content information, spatial information, reference information, theme or keywords and standard information. There are many standards for metadata of geospatial information that are issued by Federal Geographic Data Committee (FGDC) or International Standards Organization (ISO). If the metadata is standard compliant it becomes interoperable and flexible to use (Figure 4.3). A point is awarded for the availability of a metadata document and another point if it is compliant to either of the standards described in detail in APPENDIX B. . Contact information of the owner of the data is helpful for users to connect with the source to gain additional insight and results in the portal gaining a point.

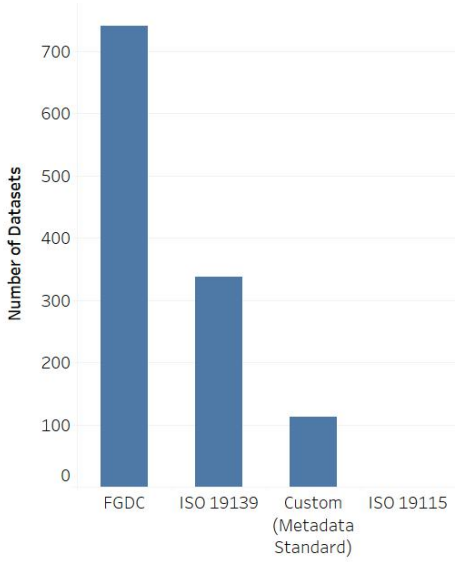

Figure 4.3 Total Number of datasets with specific standard metadata across all portals

A common drawback of these portals studied is that the information provided is not uniform among all datasets. Hence, each dataset of the portal is first evaluated for the above mentioned features. A cumulative average score is taken as the score for the portal. This way the score on a scale of one represents the proportion of datasets with above features.

Data Characteristics

Information about the data in terms of update frequency, temporal and spatial characteristics are essential for cleaning and analyzing the data (Table 4.5). Update frequency informs the user about how frequently the data is updated. Temporal Coverage is defined as the time period in which the data was collected or is applicable. Temporal resolution provides the smallest time interval in which the data was collected. Spatial Coverage is defined as the geographical area covered in the dataset, this can be at city, county or state level. Spatial Resolution is the smallest geographic unit used for data collected. Scores for each of these features are calculated similar to data description features. Thus, each feature is scored on a scale of one.

Table 4.5 Example of Data Characteristics provided across portals

Data Characteristics	**ESRI**	**Socrata**	**CKAN**	**DKAN**	**RITIS**
Update Frequency	Updated Annually	As needed	Irregular	Annually	2hrs 30 min ago
Spatial Coverage	Entire state of Michigan	Statewide	National	North Dakota	State and road functional class
Spatial Resolution	-	Thruway exit	1-arc second	POLYGON ((-102.151748791 48.998722201, …, -104.048726205 48.99981441))	Road selection tool – segments,
Temporal Coverage	2012	2015	1995 through 2014	Wednesday, December 26 2018 -06:00	Chosen date-time range
Temporal Resolution	24-hour intervals	1 hour interval	-	-	Seconds, minutes

Data Performance

Provision of good quality data is highly essential for its effective use. To evaluate this we designed features, Views or Downloads of dataset, Accuracy report of data and Completeness of data. Socrata designed portals provide real-time information of views and

downloads for each dataset. If this information is available for all the dataset, the portal is awarded one point. An accuracy report is a data quality information which provides a general assessment of the quality of the dataset. This piece of information is crucial in creating reliable results or products from the dataset. Missing data pose problems during analysis, hence, we looked at whether the publishing agencies provide complete datasets or account for the missing data. If all datasets were complete with no missing data points, the portal was awarded one point.

Legal Provisions

Although the data is provided openly, it is imperative to comply with certain terms and conditions. These are referred to in the terms of a license. Creative Commons is a popular license used by open data publishers which protects the copyrights of the owner of the data. It allows for openness in distribution and reuse, on the condition that the owners are attributed for in such acts. There are different types of Creative Commons with defined conditions described in detail in APPENDIX C. . Many portals design their own custom license providing clear terms of data validity and reliability (Figure 4.4). If the dataset is licensed with the above traits, it is awarded one point. If no license is specified it is awarded zero points. A cumulative average of the points awarded to all the datasets is calculated as the score for portal. The portal is awarded one point if it is adhering to a custom data policy or the National Open data policy.

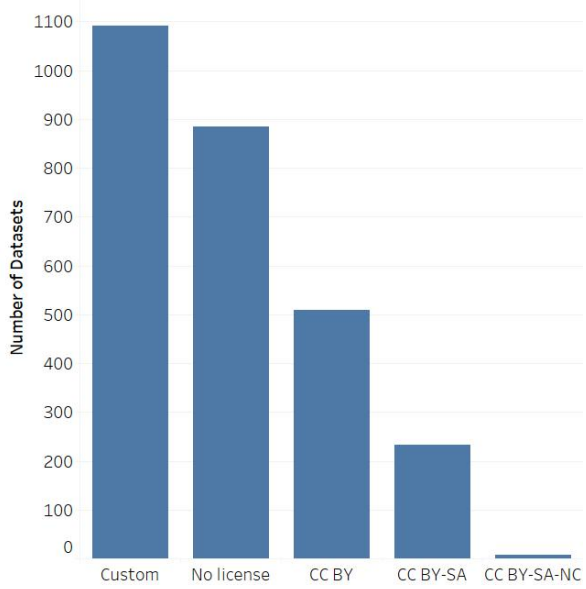

Figure 4.4 Total Number of Datasets with specific license type across all portals

Designing the Rubric Weights

Data Portal Evaluation Rubric was designed to evaluate the portal usability, data information and relevance of content to transportation. The different criteria for evaluation under each of these categories were designed based on the National Open Data Policy, existent methods of open data evaluation and impactful case studies of open transportation data publication. The next step in designing the rubric was assigning weights to the different criteria of evaluation. To assign these weights with strong reasoning we sought to conduct a survey among popular user groups in the open transportation data community and obtain their feedback on essential elements for open data portals and their data.

Analytical Hierarchy Process

Analytical Hierarchy Process (AHP) is a technique popularly used for multi-criteria decision problems. It provides a structured approach to fit the criteria into a hierarchical structure and assign weights to them. AHP describes the problem in terms of goal, criteria and alternatives. Each of these factors occupy a different level in the hierarchy which is

referred to as the node. Alternatives are compared in pairs in terms of their importance based on a certain criteria. Similarly, criteria are compared in pairs based on their importance towards achieving the ultimate goal. These decisions are completed by people who are believed to have sound knowledge and expertise in the field under study (*45*).

Constructing the hierarchy for our rubric design, there are four levels – Features, Parameters, Category and Final Weighted Score. There are three categories of evaluation which represent the criteria of assessment for open data portals and their data. Each category is described by certain parameters which are concepts or constructs highlighting the key aspect of evaluation. Features are the tools which describe the parameter whose presence in the portal or data is identified. Figure 4.5 shows the hierarchy in DPER and provides a count of the number of parameters and features described under each category.

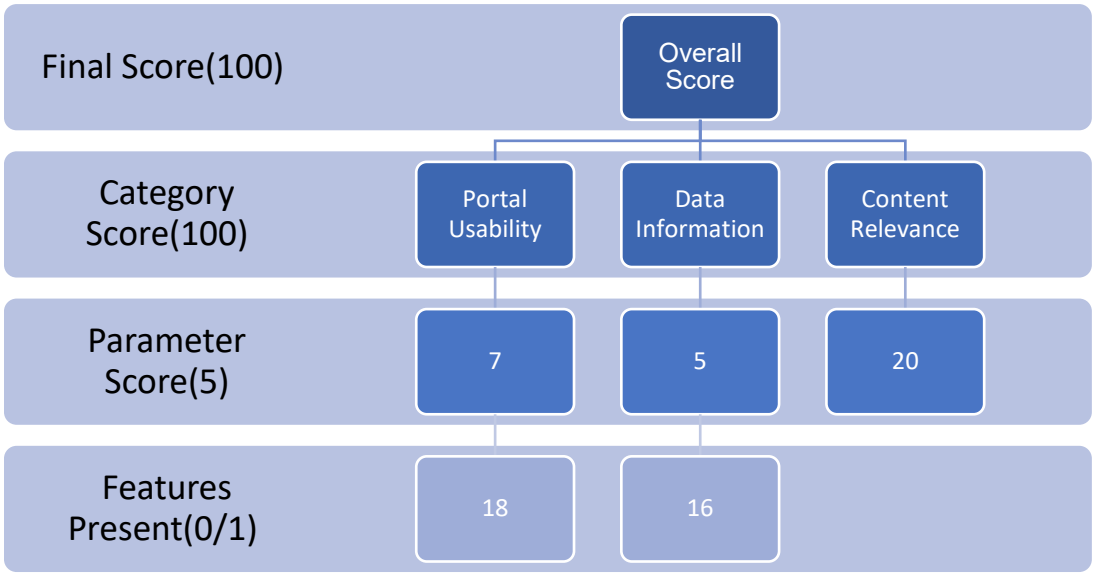

Figure 4.5 Hierarchy structure of DPER

AHP uses a subjective scale for comparisons of the different criteria which is similar to a standard Likert scale based on the user's subjective opinions. These subjective opinions can be converted to a numerical scale and used to calculate weights or priorities to the criteria

based on their importance towards achieving the ultimate goal (Figure 4.6). Consistency in an important factor of AHP which is allowed for with a small tolerance. Among three criteria's A, B and C, if the user chooses A over B (A>B) and B over C (B>C) and by consistency A should be chosen over C (A>C). This is translated in numerical value as the consistency ratio which must be less than or equal to 0.1.

The Fundamental Scale for Pairwise Comparisons		
Intensity of Importance	**Definition**	**Explanation**
1	Equal importance	Two elements contribute equally to the objective
3	Moderate importance	Experience and judgment moderately favor one element over another
5	Strong importance	Experience and judgment strongly favor one element over another
7	Very strong importance	One element is favored very strongly over another; its dominance is demonstrated in practice
9	Extreme importance	The evidence favoring one element over another is of the highest possible order of affirmation
Intensities of 2, 4, 6, and 8 can be used to express intermediate values. Intensities of 1.1, 1.2, 1.3, etc. can be used for elements that are very close in importance.		

Figure 4.6 Standard AHP scale for pairwise comparisons (45)

Survey and Weights for criteria

The survey was sent to transportation officials from Federal Highway Administration (FHWA), American Association of State Highway and Transportation Officials (AASTHO) and state DOTs. The survey was created in the AHP format where pairwise comparisons were created for each set of criteria to be evaluated. The survey focused on four topics. There were 3 questions comparing the categories of the rubric asking the participant which they find is more important and how more important based on the standard AHP scale. The next 21 questions followed a similar pattern and compared the 7 parameters under Portal Usability. The next 10 questions compared the 5 parameters under Data Information. The

final question listed the 20 topics under Content Relevance and asked the user to choose which topics of data they prefer. We received 17 responses which have been compiled using the BPMSG AHP software to calculate consistent priorities (*46*). The results from the survey are summarized below.

Table 4.6 Priorities for Categories of Rubric (Consistency Ratio: 0.09)

Category	Weights
Data Information	0.4178 (41.78%)
Content Relevance	0.3624 (36.24%)
Portal Usability	0.2197 (21.97%)

Table 4.7 Priorities for Parameters of Portal Usability (Consistency Ratio: 0.01)

Parameter	Weights
Accessibility	0.2433
Ease of Usage	0.2067
Number of Transportation Datasets	0.1521
Interactive Visualization	0.1310
Application Developer Tools	0.0951
Analytical Tools	0.0933
Feedback tools	0.0785

Table 4.8 Priorities for Parameters of Data Information (Consistency Ratio: 0.02)

Category	Weights
Data Performance	0.3348
Data Description	0.2410
Data Characteristics	0.1827
Data Formats	0.1353
Legal Provisions	0.1063

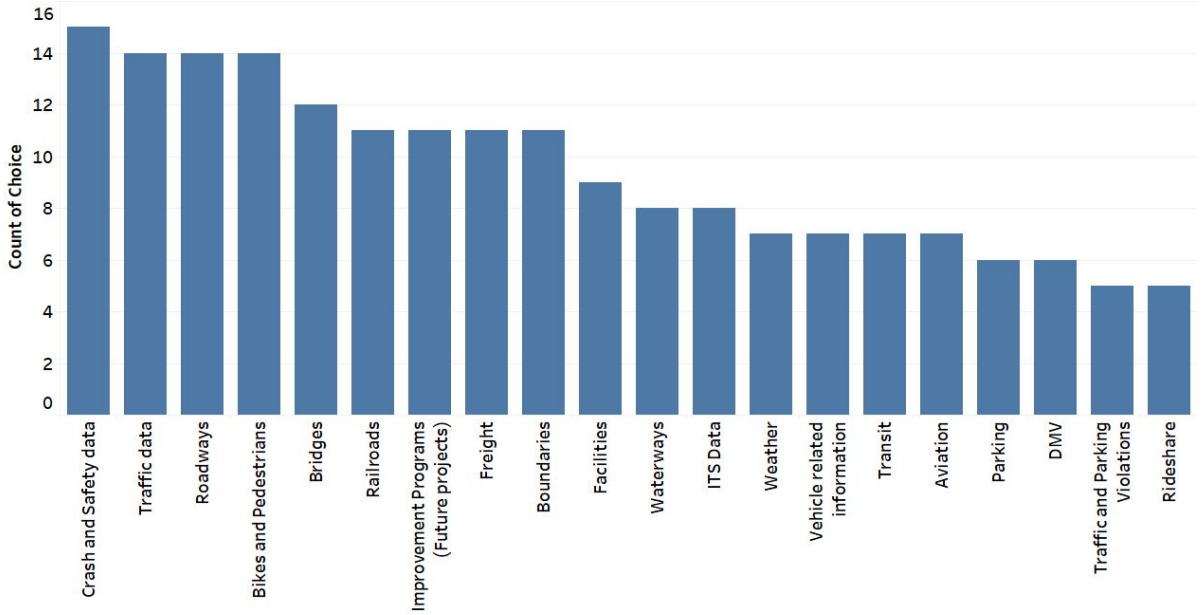

Figure 4.7 Results from Preferred choice of transportation data

The weighted overall score is the summation of the three described topics areas subscores: relevance of data to transportation (36.24 %), evaluation of data portal (41.78%) and the evaluation of data content (21.97%). Using this overall score, the portals studied are ranked and analyzed. A summary of the rubric design is shown in Table 4.9 .

Table 4.9 Summary of Data Portal Evaluation Rubric (DPER)

Parameters (Weights)		Features (Scale)	
Relevance (36.24 %)			
Aviation (0/1)	Facilities (0/1)	Railroads (0/1)	Transit data (0/1)
Bikes and Pedestrians (0/1)	Freight Data (0/1)	Rideshare (0/1)	Vehicle related information (0/1)
Boundaries (0/1)	Improvement Programs (0/1)	Roadways (0/1)	Violations – Traffic and Parking (0/1)
Bridges (0/1)	ITS Data (0/1)	Safety and Crash Data (0/1)	Waterways (0/1)
DMV (0/1)	Parking (0/1)	Traffic Data (0/1)	Weather (0/1)
Portal (21.97 %)			
Ease of Usage (20.67%)	Search Bar (0/1)	Clicks To Reach Portal (0-2)	Categorization (0/1)	Tutorials (0/1)
Accessibility (24.33%)	Preview Of Data (0/1)	Download Data (0/1)	Download Filtered Data (0/1)	links to other info (0/1)
Interactive Visualization (13.1%)	Geospatial Maps (0/1)	Graphical representation (0/1)	Tabular Representation (0/1)
Analytical Tools (9.33%)	Filter and Sorting tools (0/1)	Statistical Tools (0/1)	

Table 4.9 (continued)

Application Developer Tools (9.51%)	API Guide (0/1)		API Query Tool (0/1)		Number Of Applications Developed (0-5)
Number of Transportation datasets (15.21%)	1-85 (1)	86-170 (2)	171-255 (3)	256-340 (4)	341-423 (5)
Feedback Tool (7.85%)			Comment Section / Contact Email (0/1)		
Data Content (41.78 %)					
Number of Data Formats (13.53%)	1-3 (1)	4-6 (2)	7-9 (3)	10-12 (4)	13-15 (5)
Data Description (24.1%)	Descriptive Text (0-1)	Data Dictionary (0-1)	Metadata (0-1)	Standard Compliance For Metadata (0-1)	Contact Info of data owner (0-1)
Data characteristics (18.27%)	Update Frequency (0-1)	Temporal Coverage (0-1)	Temporal Resolution (0-1)	Spatial Coverage (0-1)	Spatial Resolution (0-1)
Data Performance (33.48%)	Views/ Downloads (0/1)		Accuracy Report (0-1)	Complete datasets (0/1)	
Legal Provisions (10.63%)	License (0-1)		Data Policy (0/1)		

CHAPTER 5. RESULTS AND DISCUSSION

In this chapter we discuss the evaluation of different portals as conducted using the DPER. Based on the overall scores we ranked the portals. The features largely present or absent in these portals are highlighted. Individual portals which performed the best and worst are compared. RITIS data portal scores are also discussed in detail to understand its performance compared to other portals. With different developer, there is significant difference in the design and data information published which is highlighted through a comparison of their scores.

Each of the 43 data portals was then analyzed based on the DPER scoring. Each portal and its datasets was evaluated and scored out of 100. The states were then ranked based on their final scores in Table 5.1 (Figure 5.1). The top 5 ranked portals were in New York, Maryland, District of Columbia, RITIS and US DOT Portal respectively. The bottom five ranked portals were Wyoming, Tennessee, Mississippi, Nevada and Minnesota state portals respectively.

Table 5.1 Ranking of Open Data Portals by DPER Scoring

Portal	**Overall Score (100)**	**Portal Usability Score (21.97%)**	**Data Information Score (41.77%)**	**Content Relevance Score (36.24%)**
New York	72.29	19.15	18.71	34.43
Maryland	61.75	17.46	15.30	28.99
DC	58.83	17.89	11.95	28.99
RITIS Data	55.18	20.18	24.13	10.87
US DOT Portal	53.63	16.55	15.34	21.74

Table 5.1 (continued)

Connecticut	52.61	15.88	14.99	21.74
Iowa	50.18	16.18	6.82	27.18
Washington	46.83	15.29	9.80	21.74
Delaware	46.18	14.54	17.14	14.50
Florida	44.77	14.92	9.92	19.93
US Gov portal	42.69	10.56	17.64	14.50
Missouri	41.78	16.79	12.31	12.68
Pennsylvania	41.71	14.92	5.05	21.74
Ohio	40.94	9.37	6.20	25.37
Texas	40.12	14.62	9.19	16.31
Michigan	38.91	14.40	10.01	14.50
Oregon	38.27	15.45	11.95	10.87
Massachusetts	36.85	13.28	3.64	19.93
Vermont	36.50	16.79	14.28	5.44
Oklahoma	36.09	13.58	8.01	14.50
Georgia	35.35	12.22	6.82	16.31
New Jersey	34.12	14.54	14.14	5.44
Bureau of Transportation Statistics	33.96	12.52	5.14	16.31
Utah	33.88	15.83	3.56	14.50
Hawaii	33.56	15.15	12.97	5.44
Montana	33.30	12.89	4.09	16.31
Virginia	33.29	14.91	3.89	14.50
Illinois	32.49	12.91	3.27	16.31

Table 5.1 (continued)

West Virginia	32.31	14.18	5.45	12.68
California	31.92	14.24	10.43	7.25
North Dakota	31.27	11.18	11.03	9.06
Arizona	30.46	13.28	4.50	12.68
Idaho	29.93	12.97	2.47	14.50
Arkansas	28.83	7.66	8.48	12.68
South Dakota	25.14	13.28	6.42	5.44
Colorado	24.83	7.72	8.05	9.06
Kentucky	24.01	11.61	5.16	7.25
Nebraska	23.76	8.70	7.81	7.25
Wyoming	22.61	14.01	4.98	3.62
Tennessee	22.15	11.99	6.54	3.62
Mississippi	19.77	13.28	4.67	1.81
Nevada	18.54	13.58	3.14	1.81
Minnesota	13.05	8.09	3.14	1.81

The color highlights in Table 5.1 indicate the top performances of the portals for each category. The highest scoring portals are colored green, followed by yellow and blue coding. The weight of the subscores plays an important role in assessing the performance of the portals. A difference in weights allotted would cause slight changes in the ranking but ultimately the evaluation is dependent on the parameters chosen which allows for easy adaptability of the rubric to assessing different objectives.

Four of the top five portals are developed by Socrata. These portals performed well in terms of relevant content, map visualization, data charts, feedback platforms, views per data, data dictionary and accessible data formats. New York State portal offers the largest number of transportation datasets covering nineteen of the twenty topics listed in the rubric. Socrata provides an easy user interface with interactive visualizations and API Query tools. There is a clear data dictionary that accompanies 82% of the datasets. Datasets are published with information on Update Frequency (95%), Temporal (61%) and Spatial characteristics (55%). State of New York has launched its own open data project with clearly drafted directives and guidelines.

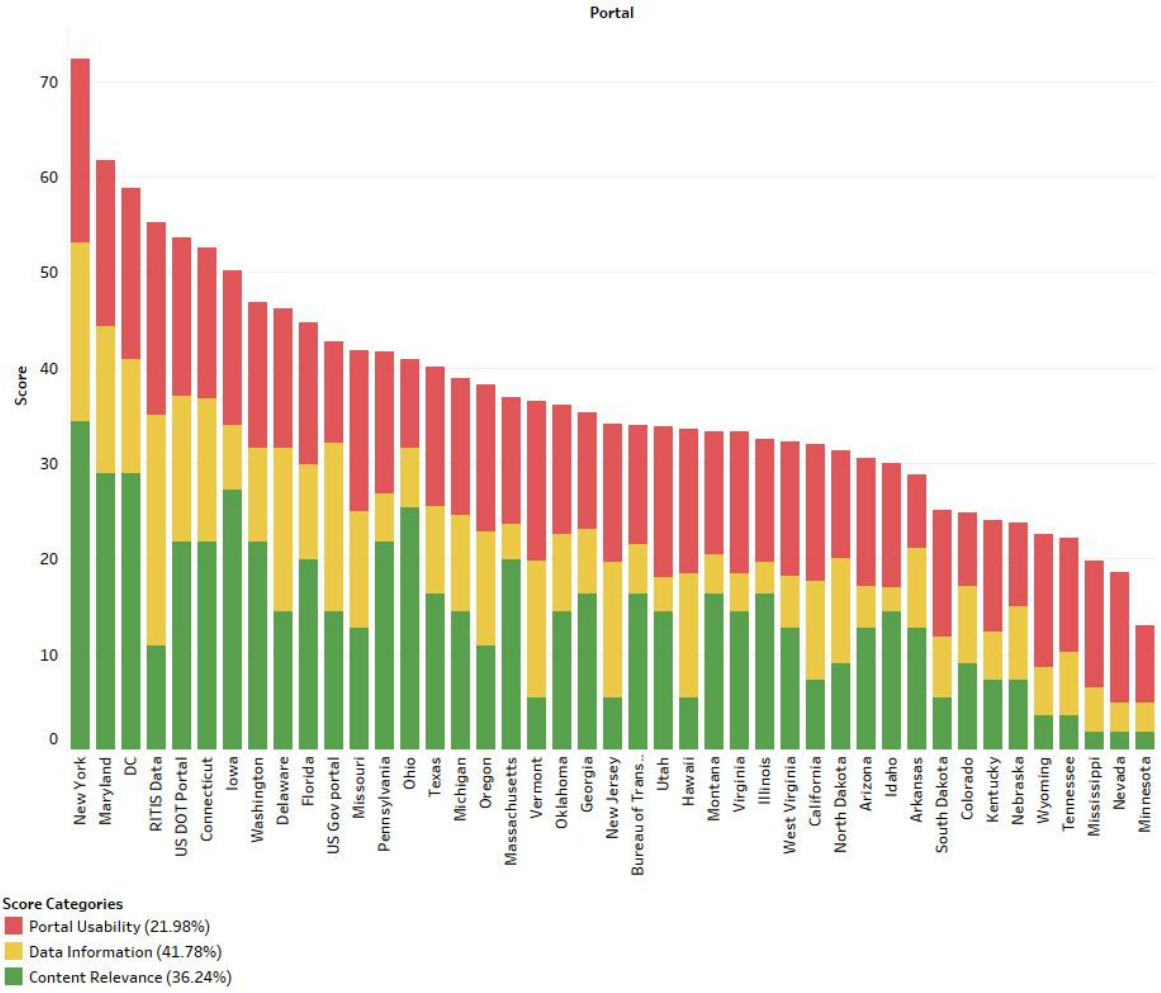

Figure 5.1 Visualization of Ranking of Portals

Figure 5.2 highlights the performance of State of New York Portal across different features when the scores are normalized to a scale out of five. The least scoring portals performed poorly because they do not provide a good user interface and had limited number of datasets. Minnesota performed poorly as it offers external links to datasets with no scope for interaction. Under the category of transportation, Minnesota portal offers only three datasets. These datasets can be downloaded in PDF and CSV formats but cannot be previewed or analyzed online. Also, these datasets are provided with only a simple description and no other details (Figure 5.3). DPER is thus effective in highlighting the areas for improvement. These portals should focus on improving their user interface as well as data content to increase usage. This rubric would serve as a comprehensive guide for these portals to improve their performance across all categories.

Although the data content is comprehensive, other features such as description and tools for analysis can help improve usage and views within an open data portal. These types of features are typically controlled by the developer designing the open data portal. Hence, the drawback of this form of open data publication is that the benefits of open data depend on the developer's style of the publication. ESRI open data platforms are widely available but Socrata platforms perform slightly better based on the DPER (Figure 5.4). Socrata provides significant features such as visualization, analytical tools, detailed data description and accessible formats of data which are the key factors enabling these portals to perform better. Every developer follows their own style and maintains this uniformity across all portals designed. Hence, DPER is useful for developers to learn from each other and build features to provide a more interactive and friendly user interface.

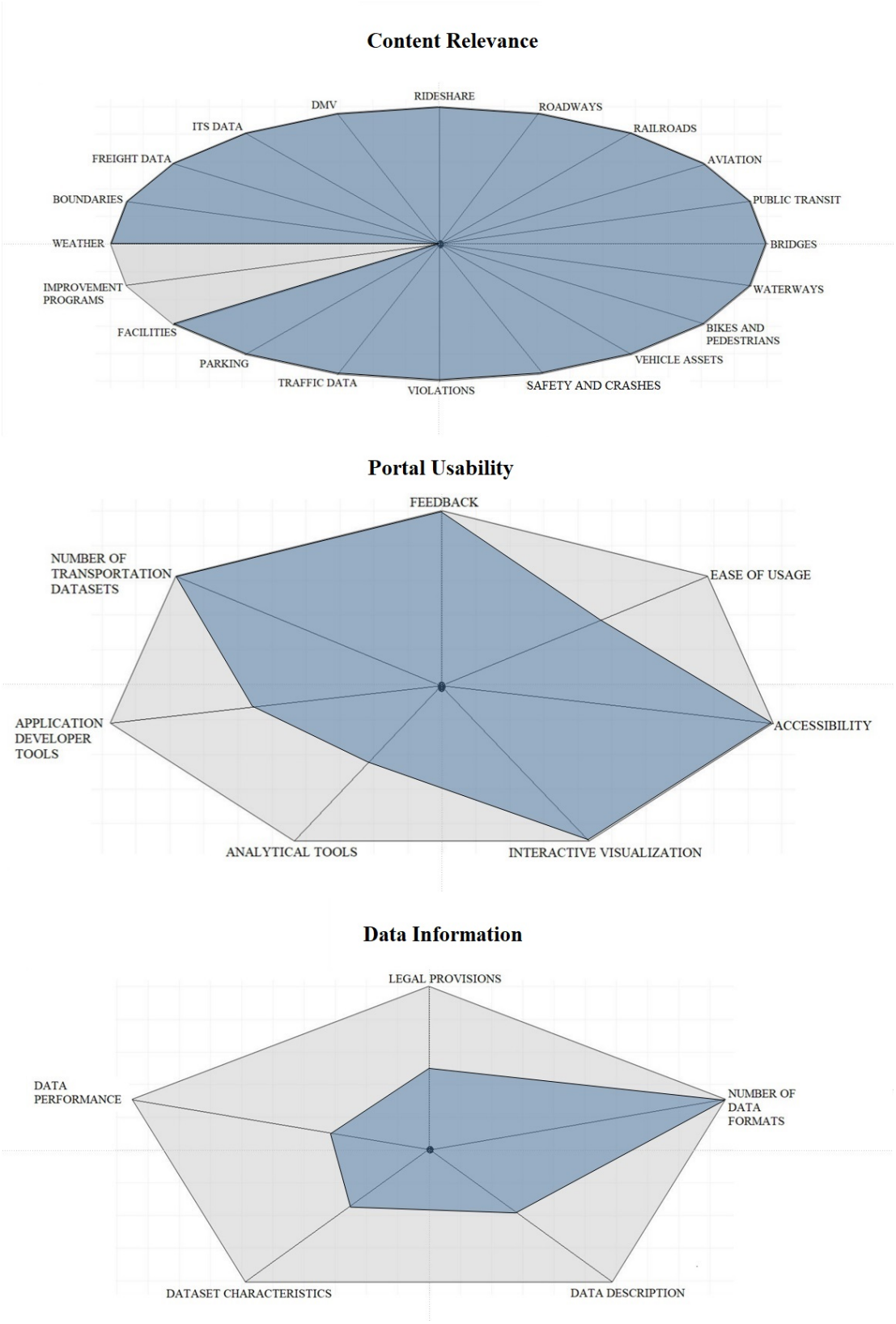

Figure 5.2 Scores of State of New York Portal across different features

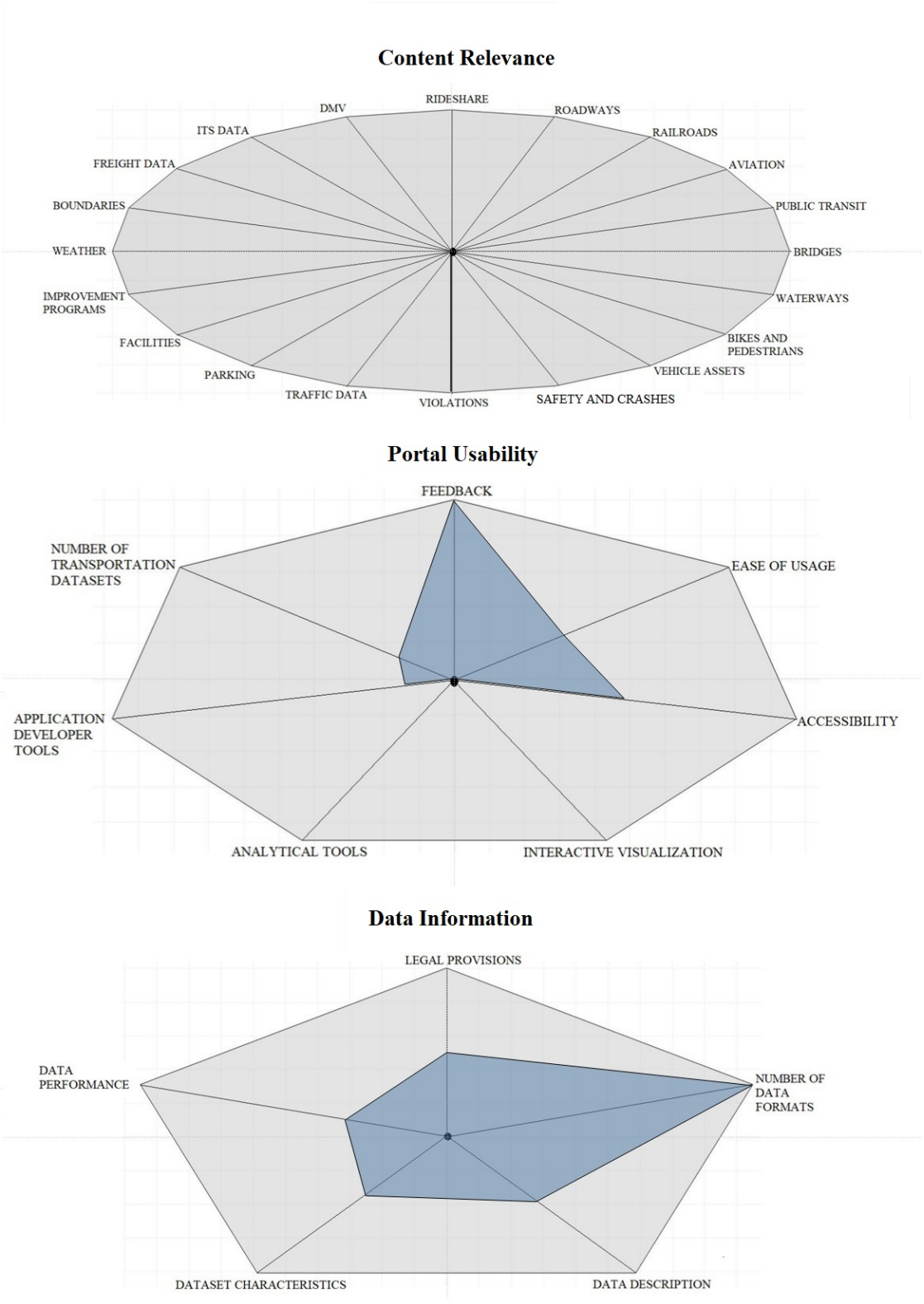

Figure 5.3 Scores of State of Minnesota Portal across different features

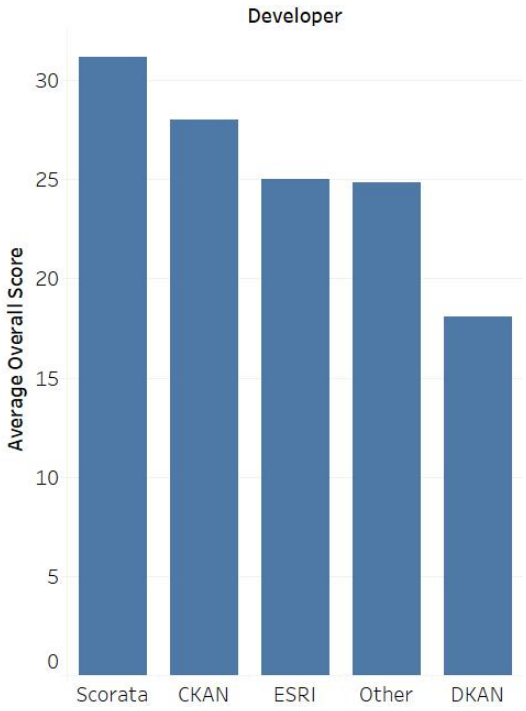

Figure 5.4 Average Overall Score for different developer Portals

This variance across different developer portals is due to the features they choose to offer. Across all portals there are certain features that are prominently absent. In case of Portal Usability none of the portals studied provide statistical tools of any kind for data analysis. In contrast features such as Search bar, Geo spatial maps and categorization data are present widely among the portals studied. For Data Information most portals fail to offer a quality or accuracy report of the data provided. Many datasets contain missing points which are not acknowledged. Spatial Resolution of datasets is not available for datasets across most portals studied. Figure 5.5 and Figure 5.6 highlights the several features absent across the portals studied. Based on their individual performance each portal can identify the weak areas and aim to improve. A good learning point would also be observing the performance of other portals. Developers can learn from each other to improve their design to include the best of the features.

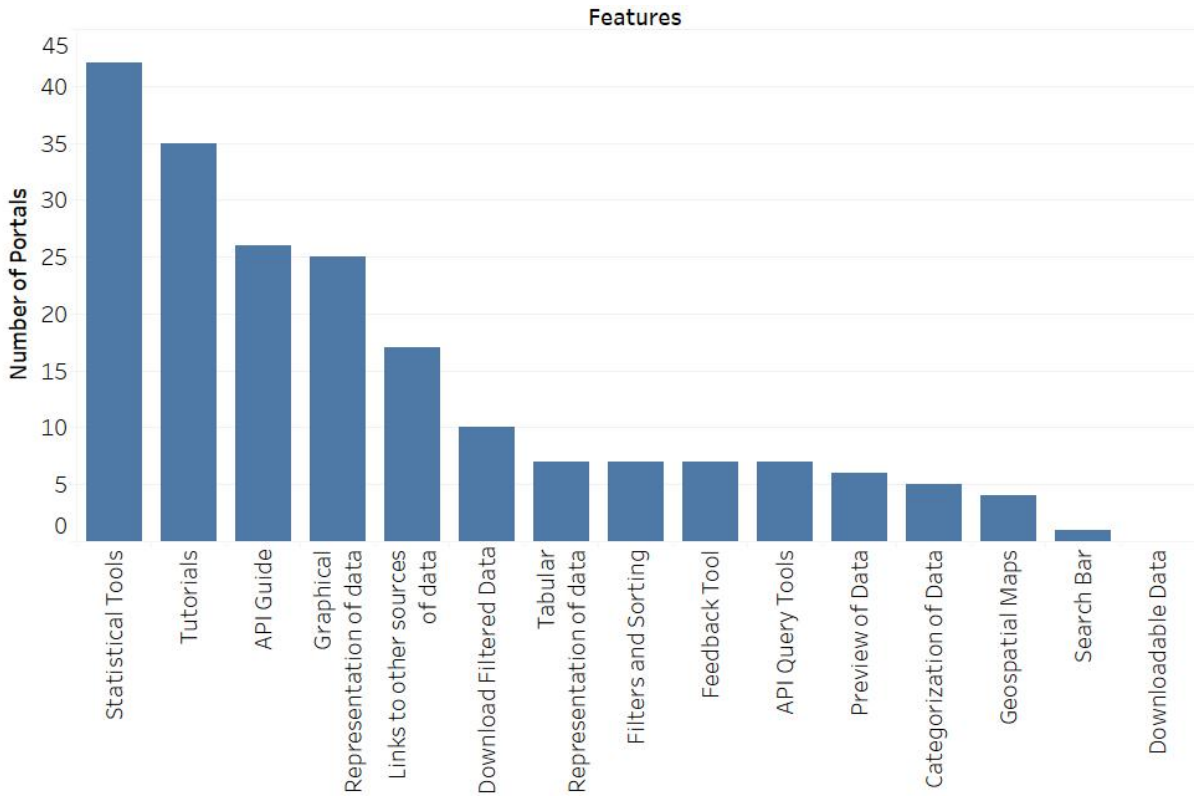

Figure 5.5 Count of Portals with absence of specified feature of Portal Usability

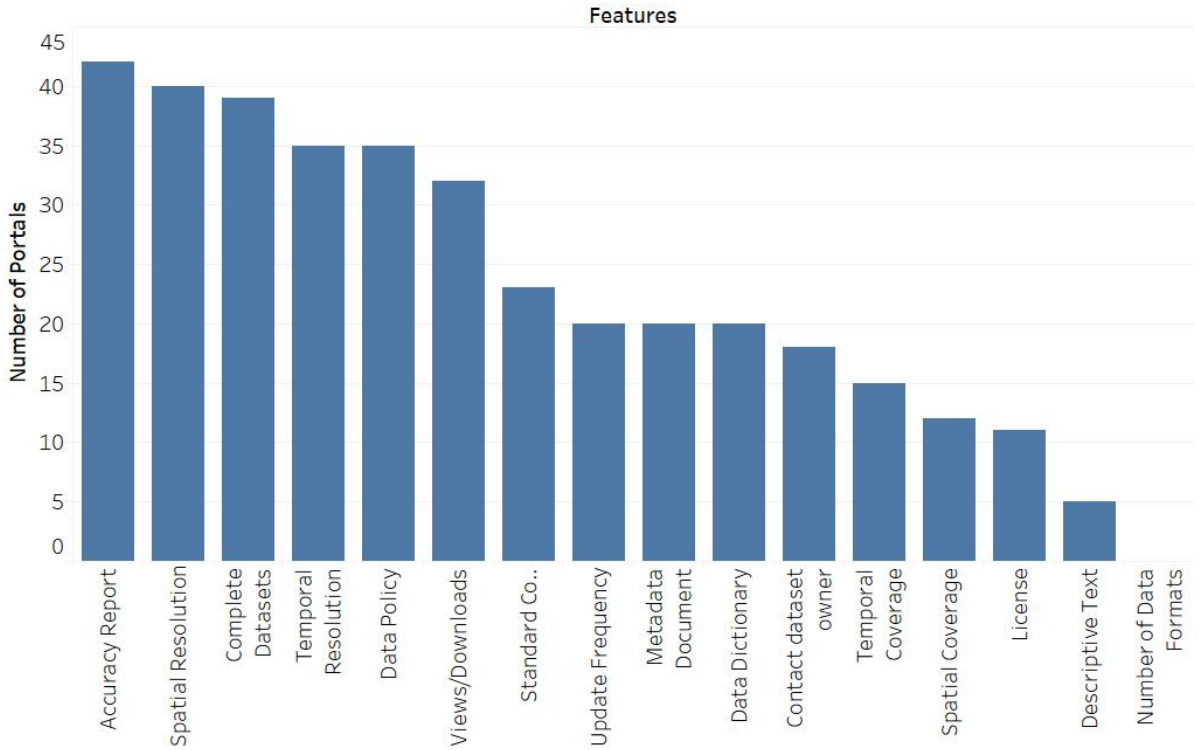

Figure 5.6 Count of Portals with absence of specified feature of Data Information

In comparison to the other portals RITIS (Transportation Big Data Portal) ranks 23rd among the portals studied. The data available is interactive with several applications that RITIS has developed for the users. (26) While downloading data, detail information such as metadata, temporal and spatial characteristics and quality confidence are provided Figure 5.7). The design of the RITIS portal is very different from the other portals to accommodate the large stream of incoming data from multiple data sources. The massive data downloader application is an effective tool to download large data for analysis. It allows the user to select roadways on the U.S. map with several specifications such as data fields, time and date range for which information is collected. The biggest drawback of the RITIS data portal is that it is not open source. However, the RITIS data portal has been highly useful to several transportation agencies in the public and private sector.

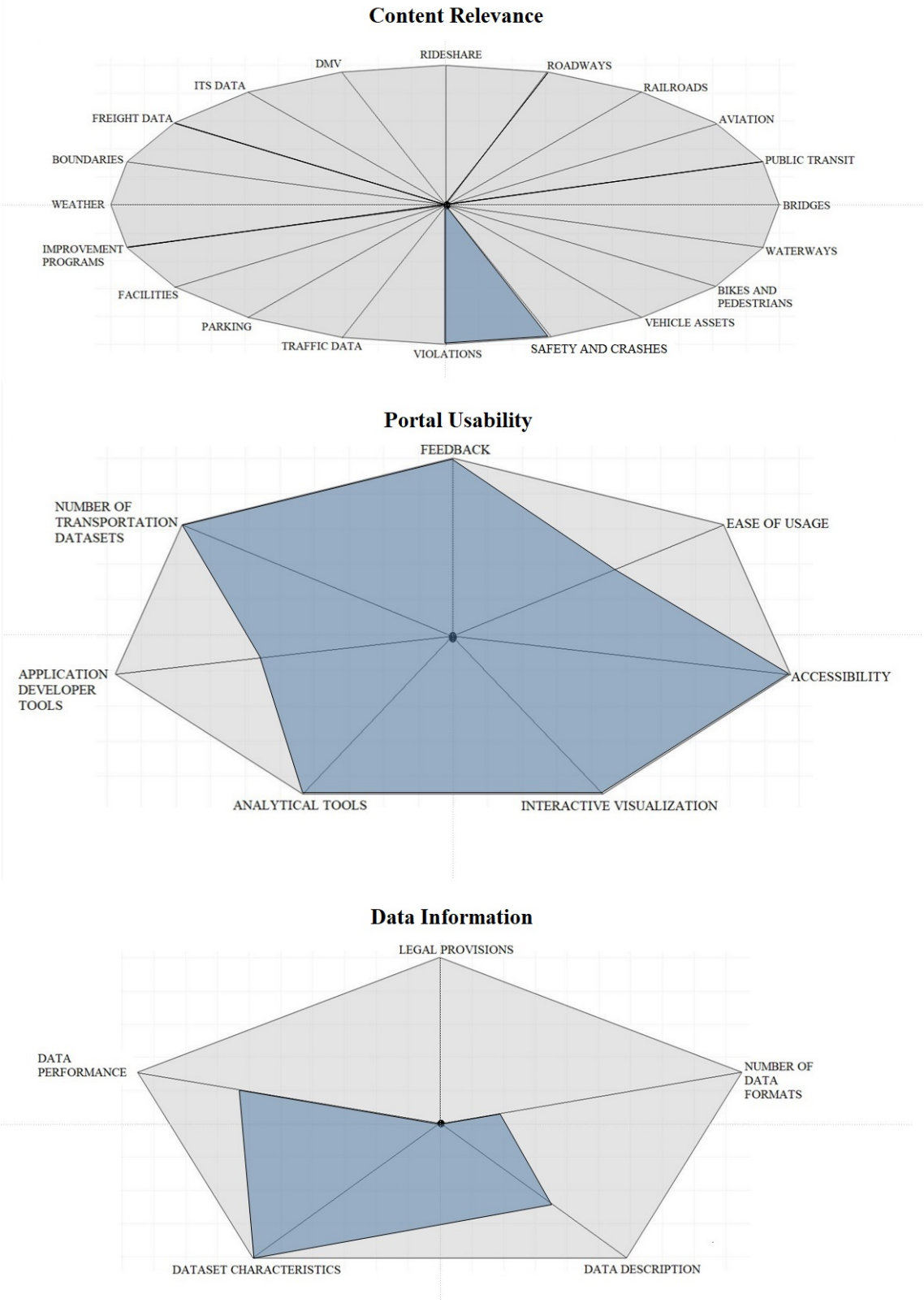

Figure 5.7 Scores of RITIS data portal across different features

CHAPTER 6. CONCLUSION

Our research objective was to design an evaluation rubric and study the current status of the different open data portals available for transportation data. The DPER designed was used to evaluate the 43 Data Portals and their data contents. The aim of this rubric was to evaluate the quality based on features of the portal, openness of data and relevant data content. The portal of the State of New York scores the highest with its rich data content and the ease of usability for end users. Key factors contributing to New York's portal's performance are the user-friendly interface with interactive visualization tools, detailed data description, relevant transportation data and useful API references for application developers. In contrast, low scoring portals fell short in providing a smooth user interface and relevant data. Socrata and ESRI are developers whose portals perform significantly better than others and also accounted for a majority of the portals evaluated.

Today, there are over 900 open data portals in the U.S. which indicates that the space for transportation datasets has highly expanded. However, there is a lack of clarity in the field of open transportation data in terms of the source, use, and application. Hence, the first step was to create a repository of Open Data Portals that provide transportation datasets. This was achieved through rigorous searches over the internet and the prevalence of open data portals created by State DOTs over the years.

The next step was to identify the datasets pertaining to transportation. There are no uniform categories of transportation data found in every portal. After manually scouring through the datasets available in each portal, the first section of the rubric was developed which should also provide guidance to other agencies about what transportation data they can work towards making open. This covers only the dataset already available and it does not

necessarily meet all user's needs. Additional efforts should be made to procure feedback from users as well as evaluate the availability of data within DOTs that can also be provided. Hence, the feedback tool included in our rubric plays an essential part in leading the design of these portals towards better serving the user's needs. This would also help verify whether the chosen categories aptly describe transportation data or should be expanded further.

Next we aimed at designing an effective and flexible rubric which can serve the purpose of quality assessment. The design was modified several times to identify a suitable one which could be extended for future work. The DPER performs well in terms of evaluation as the parameters are clearly defined to assure reliability. The rubric analyses data over different parameters and normalizes them to a uniform scale. This is an asset as it improves its extensibility and provides scope for including additional parameters in the future. The portals studied were diverse in terms of design, data content and data size. This was an advantage which resulted in designing a comprehensive rubric which could be applied to evaluate any open data portal.

The DPER designed is based on features of the open data portals and the principles of the open data policy. There are many developers for open data platforms. Each developer possesses their own style and design in developing the portal. Socrata offers a user-friendly portal with detailed information on data, contact details of data owners and interactive tools. Also, it provides a detailed data dictionary which is absent in most portals. However, they do not provide a requisite for metadata document which is only provided at the owner's discretion. In contrast, ArcGIS provides metadata for every dataset. Hence, there is variability in the performance of the portals developed by them. This imbalance across different portals indicate the need for uniformity in the user interface.

Addressing the variability has been a major part of this thesis. The rubric has provided a clear picture of the current scenario of the different open data portals for transportation. Using the rubric individual agencies can not only identify their own drawbacks but can also observe the performance of the high scoring agencies. The rubric essential becomes a yardstick in guiding the DOT and any agency which hopes to invest in open transportation data publication.

The highest score on the rubric is 72.30 secured by State of New York portal. One of the main reasons for this score is that the data information score considers the quality of all datasets that the portal offers. When the portal is scored for data dictionary, data description or data characteristics we look at how many datasets can provide clear information. Mere provision is not the standard anymore, data needs to be complete with relevant information for best utilization. This also means portals with greater number of datasets will be penalized harder for not covering enough data information. Another area of drawback is the category of data provided. New York Portal offers data across 19 of the 20 topics in the rubric where as an Arizona or California open data portals cover only 5 or 6 of the topics. This gap in data provision has highlighted the need to assess data collection and availability across different states. If there are states lagging behind in this aspect proper initiative must be in place to collect the data first and later it can be published openly. These are the challenges that the agencies look forward to tackle as they try to improve their performance in open data publication.

Exploring open data in transportation has certainly been an eye opener in identify its potential. Throughout the process of the research, we faced difficulty in finding detailed and meticulous information on this topic. Studying 43 portals has helped us in understanding its

different aspects, the portal and its content. To share this knowledge, we developed a visualization tool highlighting the critical observations of our research. We are in a time period where data is highly valued. Developments in big data analytics are rapidly growing. Similar growth can be observed in the sector of open data. The paths of Big data and open data are bound to cross, sooner than later. Transportation is a field which stands to benefit from this growth, hence, this is the right time to evaluate this data and discuss establishing quality standards for the same.

The DPER provides agencies with the ability to measure the performance of their open data portals and draw inspiration from higher ranking portals. The DPER also indicates the areas of variability which highlights the need to define a uniform format for publishing open data that can lead to beneficial results. Expanding the same idea, our DPER contributes towards exposing the areas for standardization. This uniformity can also benefit developers and researchers who want to obtain data across multiple agencies without the barrier of inconsistencies in data content. With the advent of DOTs launching their repository of transportation datasets, the time is right to explore the idea of standardizing both open data and the design of its portals.

REFERENCES

1. Porter, J. Google Doodle Celebrates the 30th Anniversary of the World Wide Web. , 2019, pp. 1–3.
2. Giacaglia, G. Data Is the New Oil. https://hackernoon.com/data-is-the-new-oil-1227197762b2.
3. The Open Knowledge Foundation. What Is Open Data? *Global Open Data for Agriculture and Nutrition*. 1–34.
4. Hörz, M. Open Science Data. http://www.michael-hoerz.de/wp-content/uploads/WS13-09-Open-Science-Data.pdf.
5. Open-Source-Software Movement - Wikipedia.
6. Simon Chignard. A Brief History of Open Data - Paris Innovation Review. http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data.
7. Obama, B. Executive Order 13642 - Making Open and Machine Readable the New Default for Government Information. *Federal Register*, Vol. 78, No. 93, 2013, pp. 28111–93. https://doi.org/10.1163/_q3_SIM_00374.
8. Burwell, S. M., S. Vanroekel, F. Chief, I. Officer, T. Park, U. S. C. T. Officer, D. J. Mancini, D. Administrator, R. Affairs, and I. Definitions. M-13-13 — Memorandum for the Heads of Executive Departments and Agencies. Vol. 2002, 2017, pp. 1–12.
9. Master, G. Open Government Data. https://www.data.gov/open-gov/.
10. Open Data Inception — OpenDataSoft. https://data.opendatasoft.com/explore/dataset/open-data-sources%40public/map/?sort=code_en&basemap=jawg.light&location=2,28.94864,2.00835.
11. Manning, P. Big Data in Transport. *Big Data in History*, 2013, pp. 1–119. https://doi.org/10.1057/9781137378972.
12. Development, P. Public Transportation Embracing Open Data. No. August, 2015, pp. 1–7.
13. Emerging Technologies in Transportation Casebook_Open Data in Transportation - Wikibooks, Open Books for an Open World. https://en.wikibooks.org/wiki/Emerging_Technologies_in_Transportation_Casebook/Open_Data_in_Transportation.
14. Gurin, J. Big Data and Open Data: How Open Will the Future Be? *I/S: A Journal of Law & Policy for the Information Society*, Vol. 10, No. 3, 2015, pp. 691–704.
15. Introduction _ Regional Integrated Transportation Information System. https://www.ritis.org/intro.
16. Burwell, S. M., S. Vanroekel, T. Park, and U. S. Chieftechnolo. Open Data Policy - Managing Information as an Asset. *Executive Office of the President*, Vol. M-13–13, 2013, pp. 1–12.

17. Huijboom, N., and T. Van Den Broek. Open Data : An International Comparison of Strategies. *European Journal of ePractice*, Vol. 12, No. March/ April 2011, 2011, pp. 1–13. https://doi.org/1988-625X.
18. M Rojas, F., D. Weil, and M. Graham. Transit Transparency : Effective Disclosure through Open Data. No. June, 2012, pp. 1–85.
19. Murphy, A. TriMet Kicks off Open Data Series, Discusses Future of Transit Apps _ TriMet News.
20. TriMet Third-Party Apps.
21. McHugh, B. Pioneering Open Data Standards: The GTFS Story. *Beyond Transparency - Open Data and Future of Civic Innovation*. 125–136.
22. OpenMBTA_ Free, Open Source Real-Time Arrival Times and Schedules for Boston Public Transit.
23. About New York City Transit.
24. Fried, B. MTA Unveils Open Data Policy, Clearing a Path for NYC Transit Apps – Streetsblog New York City.
25. America, N., and C. Urban. Transportation in New York City. *Wikipedia*, 2018, pp. 1–5.
26. Rich, S. NY Adds 100 Transportation Data Sets to Open.
27. CTA - Overview (Structure, Mission, Values, Etc.
28. CTA Developer Center - Open Chicago Transit Data - CTA.
29. CTA Transit App Center - CTA.
30. Authority, C. T. Open Data from CTA.
31. Bay Area Rapid Transit.
32. Roth, M. BART a National Leader in Real-Time Data Transparency and Development – Streetsblog San Francisco.
33. BART Apps _ Bart.
34. About - BART Data Portal.
35. Solutions, D. C. # DataReads : Open Data in Public Transportation. 2–4.
36. Berners-Lee Tim. 5-Star Open Data. *2006*. http://5stardata.info/en/.
37. The Open Data Barometer | Open Data Barometer. http://opendatabarometer.org/barometer/.
38. OKF. Global Open Data Index - Methodology. *Open Knowledge Foundation*. 1–8. http://index.okfn.org/methodology/.
39. Susha, I., A. Zuiderwijk, M. Janssen, and Å. Grönlund. Benchmarks for Evaluating the Progress of Open Data Adoption: Usage, Limitations, and Lessons Learned. *Social Science Computer Review*, Vol. 33, No. 5, 2015, pp. 613–630. https://doi.org/10.1177/0894439314560852.
40. Moskal, B. M. Scoring Rubrics : What , When and How ? Vol. 7, No. 3, 2000, pp. 3–7.

41. Perlman, C. Performance Assessment : Designing Appropriate Performance Tasks and Scoring Rubrics. *Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators*, 2003, pp. 497–506.

42. Moskal, B. M., and J. A. Leydens. Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research & Evaluation*, Vol. 7, No. 10, 2000, pp. 1–10. https://doi.org/10.1016/j.asw.2010.01.003.

43. Thorsby, J., G. N. L. Stowers, K. Wolslegel, and E. Tumbuan. Understanding the Content and Features of Open Data Portals in American Cities. *Government Information Quarterly*, Vol. 34, No. 1, 2017, pp. 53–61. https://doi.org/10.1016/j.giq.2016.07.001.

44. Pack, M. Demystifying Big Data. https://www.roadsbridges.com/demystifying-big-data.

45. Analytic, S. Analytic Hierarchy Process – Leader Example. 1–13. https://en.wikipedia.org/wiki/Analytic_hierarchy_process_–_leader_example.

46. Goepel, K. D. Implementation of an Online Software Tool for the Analytic Hierarchy Process (AHP-OS). *International Journal of the Analytic Hierarchy Process*, Vol. 10, No. 3, 2018, pp. 1–5. https://doi.org/10.13033/ijahp.v10i3.590.

47. File Types. https://fileinfo.com/browse/.

48. MapInfo TAB Format - Wikipedia.

49. GeoPDF - Wikipedia.

50. FGDC. Content Standard for Digital Geospatial Metadata. *Fgdc-Std-001-1998*, 1998.

51. ISO 19115:2003. ISO 19115 Geographic Information - Metadata. 140. http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020.

52. International Organization for Standardization. ISO/TS 19139:2007 Geographic Information -- Metadata -- XML Schema Implementation. *Iso*, 2007.

53. (CC creative commoms). About The Licenses - Creative Commons. *What our licenses do*. 1. https://creativecommons.org/licenses/?lang=en.

APPENDIX A. DATA FORMATS

The datasets across all portals were available in 24 different data formats listed in Table A.1 (*47*). These include tabular and spatial data formats. Some of these data formats are non-proprietary where as other are proprietary formats which require AutoCAD, TerraGo and MapInfo applications.

Table A.1 Data Formats used for open data publication

Data Format	**Description**
Comma-Separated Values (CSV)	It stores tabular data in plain text.
XLSX	It is a file extension for an open XML spreadsheet file format used by Microsoft excel.
Shapefile (SHP)	It is a popular geospatial vector data format for geographic information system software. This format can spatially describe vectors features such as points, lines and polygons.
Keyhole Markup Language (KML)	It is used to display geographic data in an Earth browser such as Google Earth.
KMZ	A KMZ file consists of a main KML file and zero or more supporting files that are packaged using a ZIP utility into one unit called archive.
Extensible Markup Language (XML)	It is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.
JavaScript Object Notation (JSON)	It is an open standard file format that uses human-readable text to transmit data objects consisting of attribute-value pairs and array data types.
Tab-Separated Values (TSV)	It is a simple text format for storing data in a tabular structure and a way of exchanging information between databases.

Resource Description Framework (RDF)	It is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model.
RDF Site Summary (RSS)	It is a type of web feed which allows users and application to access updates to online content in a standardized, computer-readable format.
Portable Document Format (PDF)	It is a file format developed by Adobe to present documents, including text formatting and images, in a manner independent of application software, hardware and operating systems.
Hypertext Markup Language (HTML)	This is the standard markup language for creating webpages. They are provided when data is available in an external link.
Drawing (DWG)	It is a proprietary binary file format used for storing two and three dimensional design data and metadata.
Drawing Interchange Format (DXF)	It is a file extension for a graphic image format used in AutoCAD software.
Map Info TAB	It is a geospatial vector data format for geographic information systems software. It is a proprietary format developed by MapInfo Corporation. (*48*)
GeoJSON	It is a format for encoding a variety of geographic data structures. It supports geometry types such as Point, LineString, Polygon, Multipoint, MultiLineString and MultiPolygon.
GeoPDF	It refers to map and imagery products by TerraGo software applications. They use geospatial PDF as a container for maps, imagery and other data used to deliver an enhanced user experience in TerraGo applications. (*49*)
Shape Entities (SHX)	It is a file extension for a compiled shape entities file format used by AutoCAD.

CPG	These are plain text files that describes encoding applied to create shapefile.
DBF	This is a standard database file used to store attribute data and object IDs.
PRJ	This file contains the metadata associated with the shapefiles coordinate and projection systems.
SBN	It is a spatial index file that optimizes spatial queries.
SBX	It is similar to SBN files, works alongside to speed up loading times and optimize spatial queries.
TXT	It is a computer file that is structured as a sequence of lines of electronic text.

APPENDIX B. STANDARDS FOR METADATA DOCUMENTATION

Standards are documents that describe the definition or architecture for systems involved in delivering transportation data. It prescribes a standard format to attain uniformity in describing the information across different platforms. This uniformity enables industry growth, increases compatibility and interoperability among various users of the data. There are 3 specific standards available for drafting metadata documentation which are described below.

FGDC-STD-001-1998 Content Standard For Digital Geospatial Metadata

This standard prescribes a common set of terminology and definitions for geospatial data. Metadata is a description of data provided. As specified by the standard, certain topics of information are indicated to describe data. The topics of information specified for compliance are described in Table B.1 (*50*) .

Table B.1 Topic and its Description

Topic	Description
Identification Information	It describes the content of data. It includes the sub-headings description, time period, updating frequency, keywords, spatial coordinates, contact and citation Information. This is the first set of information provided in a metadata document which helps the user understand the context of data.
Data Quality Information	It is an assessment of the quality of data. The data quality is evaluated based on accuracy of attribute information, logical consistency report (defines the relationship between datasets and the tests conducted), completeness report, positional accuracy (accuracy in terms of horizontal and vertical positions), lineage from which data has been collected and cloud cover (area of data

	obstructed by clouds).
Spatial Data Organization Information	It is used to describe the mechanism used to represent the spatial information. It indicates the indirect spatial references (names of types of geographic features and location referencing methods), direct spatial reference (the system used to represent space), point and vector object information.
Spatial Reference Information	It includes the description of the reference frame which describes the mean to encode coordinates in the dataset. It describes the horizontal and vertical coordinate system.
Entity and Attribute Information	It describes the information about the attributes, entities of data and the values assumed by attributes. Under this topic entity type, attribute label, attribute definition, attribute source, attribute accuracy value and attribute measurement frequency are described.
Distribution Information	This topic provides information about the distributor who is publishing and maintaining this data. It includes the contact information, resource description, distribution liability, and ways to receive data, technical pre-requisites and available time period.
Metadata Reference Information	It includes the information on metadata document, date created, review date, contact information, metadata standard information, time conversion, access and use constraints, and security information and extensions.

These standards are used to standardize the information described by metadata. All the topic listed above provide significant information about the data provided. These standards are used by Open Data Portals publishing data with geospatial links.

Each topic mentioned in the standard provides a set of elements whose use in metadata can be mandatory or optional. Based on the given criteria and user's policy the data elements are used to provide description about the data.

Open Data Portals are platforms for agencies to publish datasets and provide open access to all with no restrictions. ArcGIS developed portals provide a tool to view the metadata of the document. Many publishing agencies have adopted the FGDC standard in creating the metadata document. Datasets published in Open data portals of Michigan DOT, Massachusetts DOT, Idaho DOT, Washington DOT provide FGDC Standard compliant metadata.

This metadata standard generalizes the content of the data. The focus is on the geospatial content of data and methods used for its representation. Hence, irrespective of the category of data this metadata can be used. It also allows for extensions which allows the user to create elements to improve the metadata quality.

ISO 19115: 2003 Geographic Information – Metadata

This standard aims to provide a structure to digital geographic data. With the ever increasing use of digital geographic data, this standard aims to standardize the data to enhance its usage. It provides a common set of metadata terminology with extension properties to standardize the description of digital geographic data. The standard describes different packages of information to describe geographic data which are listed in Table B.2 (*51*).

Table B.2 Packages of ISO 19115

Package Name	**Description**
Metadata Entity Set Information	It is an aggregate of several entities such as identification, constraints, data quality, maintenance information, spatial representation and reference system, content information, portrayal catalogue reference, distribution, metadata extension and application schema information.
Identification Information	It is the information about the data on the topics of format of data, graphic overview of data, specific uses, constraints on the resource, keywords describing the resource, updating frequency of the data and information on aggregate parts of the dataset.
Constraint Information	These are restrictions placed on the dataset. It includes access, use or other constraints.
Data Quality Information	It is an assessment of quality of geographic data. This quality is evaluated in terms of completeness, logical consistency, positional accuracy, thematic accuracy and temporal accuracy.
Maintenance Information	This package focuses on scope and frequency of updating data.
Spatial Representation Information	It identifies the mechanism used to represent the spatial information. It includes both grid and vector spatial representation.
Reference System Information	It focuses on the reference system used for spatial and temporal data.
Content Information	It identifies the feature catalogue of datasets. It includes a description to these datasets and their content.
Portrayal Catalogue Information	It identifies the portrayal catalogue used.

Distribution Information	This package focuses on distributor information. It identifies the distributor resource, format of distributing and options of distribution.
Metadata Extension Information	This is a provision provided for user to include extended elements for a comprehensive metadata describing the dataset.
Application Schema Information	It defines the application schema used in the dataset.
Extent Information	It includes the extent of temporal and spatial entity of the dataset.
Citation and Responsible Party Information	It defines a standard format for citing a source of information or the party responsible for data.

This standard aims to provide data producers with appropriate information to characterize their geographic data properly. It facilitates the organization and management of metadata for geographic data. It enables users to apply geographic data in the most efficient way by knowing its basic characteristics. It facilitates data discovery, retrieval and reuse. Users will be able to better locate, access, evaluate, purchase and utilize geographic data. It enables users to determine the usefulness of geographic data in a holding.

The conformance requirements of the standard include using the mandatory packages in the metadata document. Any metadata claiming conformance with this standard shall pass the requirements by providing the mandatory packages.

It is intended to be used by information system analysts, program planners and developers of geographic information systems as well as others in order to understand the basic principles and the overall requirements for standardization of geographic information. The metadata of data on Iowa DOT's open data portal is compliant with this standard.

It defines the schema required for describing geographic information and services. It is applicable to the cataloguing of datasets, clearinghouse activities and the full description of datasets. It is also applicable to geographic datasets, dataset series, and individual geographic features and feature properties.

ISO 19139: 2007 Geographic Information – Metadata – XML Schema Implementation

This standard (*52*) provides the semantic content to standardize metadata for geographic information. To enhance interoperability, it provides the XML schemas based on ISO 19115 content for standardized encoding. Hence, this standard describes the rules for encoding the metadata as XML schemas by providing examples for better understanding.

This standard is a technical specification providing XML schemas that are meant to enhance interoperability by providing a common specification for describing, validating and exchanging metadata about geographic datasets, dataset series, individual geographic features, feature attributes, feature types and feature properties.

The standard itself clearly describes the implementation of XML schemas derived from ISO 19115. The Unified Modelling Language is used to define the packages and their respective XML encodings.

It is intended for use by information system analysts, program planners and developers of geographic information systems who are active users of ISO 19115. Datasets from Pennsylvania DOT open data portal, Oklahoma DOT open data portal and many such portals are published with ISO 19139 compliant metadata document.

It only defines the geographic metadata XML encoding derived from ISO 19115.

APPENDIX C. CREATIVE COMMONS LICENSE

Creative Commons develops, supports and stewards legal and technical infrastructure that maximizes digital creativity, sharing and innovation. They provide tools which enables individuals and large businesses to grant copyright permissions to their creative work. Every license is designed to provide rights to copy, distribute and make some use of the work commercially as well as non-commercially. They are valid all over the world as long as the copyright is valid.

Creative Commons provides a three layer design - Legal Code, Human Readable and Machine Readable. The first layer facilitates understanding of the license by the law making authorities. The second layer is the Human Readable layer called the Commons Deed. Commons Deed enlists most important terms and conditions not included in the legal code. The final layer of the design is recognized by software which is the machine readable version of the search engine. In Open Data portals below every data, if a CC license is used then the mage as shown below in Figure C.1. The conditions of these license types are indicated in Table C.1. (*53*)

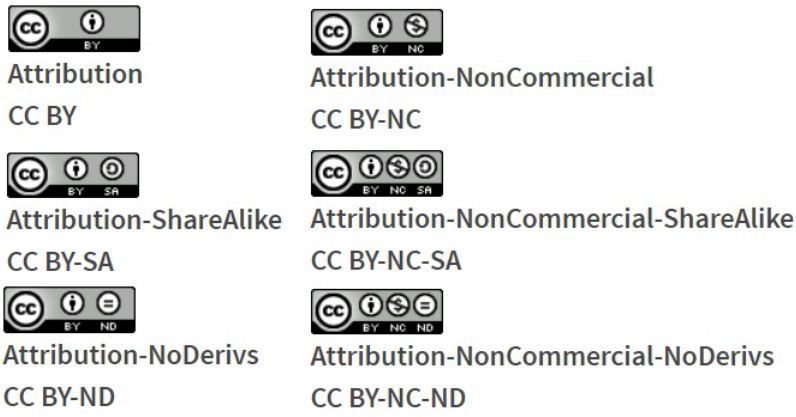

Figure C.1 Different types of Creative Commons License

Table C.1 Creative Commons License and its description

License	Description
CC - BY	**Attribution :** Allows distribution, remix, changes and extension of the work even for commercial purposes. Provide credit to original creation.
CC-BY SA	**Attribution-ShareAlike** : Allows distribution, remix, changes and extension of the work even for commercial purposes. Provide credit to original creation. Also license further work under identical terms.
CC - BY ND	**Attribution-NoDerivs** : Allows for redistribution, commercial and non-commercial use. It has to be used unchanged and provide credit to creator.
CC - BY NC	**Attribution-Non-Commercial :** Allows for remix, changes and extend the work for non-commercial purpose. New works must acknowledge the creator and be of non-commercial nature. Derivative works do not require license.
CC - BY NC SA	**Attribution-NonCommericial- ShareAlike** : Allows remix, changes and extend work for non-commercial purposes. Provide credit to creators and license new work under identical terms.
CC - BY NC ND	**Attribution-Noncommercial-NoDerivs** : It only allows for download and share the work with credit to creator. No changes can be made, cannot be used commercially.