Enhancing Open Government Data Quality: A Quantitative Evaluation Assessment for Cross-Jurisdictional Open Data Programs in Waterloo Region

by

Xuxuan Li

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Science

in

Geography

Waterloo, Ontario, Canada, 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This study builds on the previous research for identifying the current issues and gaps existing for the cross-jurisdictional data quality of the open data programs in the Waterloo region, not only as the governments in the Waterloo region have a unique two-tier municipalities structure, but also how the four municipalities the City of Waterloo, the City of Kitchener, the City of Cambridge and the Region of Waterloo shares one same data portal. The goals of this study are to understand what data quality metrics are important for the quality of open data, and how an evaluation tool can be created to effectively measure the data quality for the open data in the Region of Waterloo. A quantitative approach was used for measuring individual metrics of the data quality dimensions such as completeness, timeliness, metadata, and usability. The results show there are still a lot of improvements that can be made by the lower-tier municipalities on quality assurance, regular maintenance, and updates of data policies. The results also indicated that upper-tier municipalities like the regional government of Waterloo can take the leading role in improving the overall data quality of open data programs by creating open metadata and data standards. Additionally, the results also note the insufficient of both current and previous research and provide suggestions for future studies in similar settings.

Acknowledgements

I would like to thank my thesis supervisor Dr. Peter Johnson for providing me this precious opportunity to study as a Master of Science candidate at the Department of Geography and Environmental Management of Faculty of Environment in the University of Waterloo. Without his patient guidance, quick and detailed feedback, I will not be able to finish my thesis today. I also want to thank him for being kind and caring to me when I struggle with my mental health both in my undergraduate and graduate studies, I will not reach where I am without him.

Besides, I would like to thank my committee member Dr. Rob Feick for the feedback he gave me promptly.

Additionally, I am grateful for my mother for providing ongoing supports for me to study in Canada and the encouragements she has been giving.

Lastly, I want to express my thanks to the friends and staff members in the faculty, who helped me through my time at university and made it an unforgettable memory

Dedication

I would like to dedicate this thesis to the memory of my close friend Stephanie Ye-Mowe, who devoted her life to improving student life and helping people including me during her time at the University of Waterloo.

Table of Contents

List of Figures

List of Tables

List of Equations

Chapter 1
Thesis Overview

1.1 Introduction

Open data refers to the data that is made for the public to access, use and share for free without any copyright limitation by the government and different organizations and agencies (Sadiq & Indulska, 2017) With the rapid advancement of technology, more places are adapting to the modern world by digitizing their services and opening their information as open data. As one of the major data producers, governments hold vast amounts of data, including census information, road networks, records, and other geospatial data. As the most recent COVID-19 pandemic forcefully pushed the government and businesses to open their services, speeding up the progress of digitalization towards e-government as workplaces have adapted to remote work (Amankwah-Amoah et al., 2021). With more open data becoming available, it can increase the transparency of the public sector, enable more citizen engagement and provide more resources for researchers and scientists to support their studies. However, there is also a cost of this forced acceleration of the open data movement, which was the quality of the digital information. Some issues like quality assurance, update frequency, and data availability were presented due to the different levels of resources each government allocates to their open data program.

1.2 Level of Governments in Canada

In Canada, there are three distinct levels of government—federal, provincial, and municipal—with upper-tier and lower-tier governments (Government of Canada, 2017; Legislative Assembly of Ontario, 2023). In Ontario, municipalities are categorized into four different types, which are single-tier municipalities (Toronto, Hamilton, Brantford), upper-tier municipalities (Region of Durham, Region of Waterloo, Wellington county), lower-tier municipalities (City of Waterloo, Town of Orangeville, City of Burlington etc.) and separated lower-tier (City of London, City of Kingston, City of Stratford). Regions

and Counties are defined as "upper-tier" municipalities providing regional-wide services, which include public health, transit, planning, and more. A region or a county also encompasses multiple local (lower tier) governments, such as cities and townships, responsible for services such as issuing building permits, land severances, and managing parks, among other responsibilities (Association of Municipalities Ontario, 2021).

1.2.1 Open Data Policies in Canadian Governments

Both federal and provincial governments have specific data strategies and policies guiding their open data programs, such as Ontario's Digital and Data Directive 2021, these policies are typically applicable only within their respective departments and agencies. As Roy (2014) explained, it is challenging for a single open data policy or standard to fit every municipality's unique circumstance, given differences in governance and public administration. Zuiderwijk & Janssen (2014) suggested that due to the varying governmental levels and responsibilities, some governments and organizations may view this as an opportunity for cost-saving collaborations, while others may be concerned about potential legal risks and liabilities. Furthermore, the disparities in government resources and funding for digital infrastructure contribute to the isolation of open data efforts at different governmental levels (Roy, 2014). Consequently, each municipal government currently needs to take the responsibility of creating its own data policy or by-law to guide its open data program. There is no guideline or policy for how the upper-level government can offer to assist lower-level governments and agencies in publishing and sharing higher-quality open data. This lack of standardization complicates the reuse and reproduction of open data and places an additional burden on quality assurance for the upper levels of government (U.S. Environmental Protection Agency, 2021)

1.3 Challenges

The main challenge for the open government data provision in Canada is primarily rooted in the political system, as federated countries grant more autonomy to regional and municipal authorities, leading them to prefer creating their own policies and platforms to tailor open data programs to their unique situations (Kassen, 2018). While this tendency often occurs in countries like the United States, Canada, the United Kingdom, and some other federated nations where local governments play significant roles in e-government and open government development, it does increase costs and complexity when attempting to integrate municipal-level data into broader areas (OECD, 2019). In contrast, some unitary countries maintain more unified open data platforms with consistent formats, standards, and policies due to centralized power structures. This approach has proven to be more cost-effective for cross-jurisdictional data usage. Therefore, it is important to recognize that the current disorganized situation in open data is also heavily influenced by the dynamics of political power.

1.4 Open Government Data in Waterloo Region

The Region of Waterloo is an upper-tier municipality that comes with three lower-tier municipalities: The City of Waterloo, the City of Kitchener, and the City of Cambridge, comprising the two- tiers of municipalities at both regional and local levels.

The reason for choosing the Region of Waterloo along with the City of Waterloo, City of Kitchener and City of Cambridge for this study is due to their unique situation of open government data. Typically, in a regional setting with well-established open data programs such as the Region of York, and the Region of Durham, the datasets are only focused on the regional scope with information gathered from lower-tier municipalities and services. Even when the regional data is not available, users can integrate regional-level data using data from lower-tier municipalities and settlements. These regions collect and share regional-level datasets regularly from their departments and lower-tier municipalities to ensure data quality. However, instead of having multiple data portals by separating datasets from each tier

of municipalities, the open data program in the Region of Waterloo only has one data portal, sourcing directly from each local municipality, functioning more as a massive archive by including all datasets from lower-tier municipalities. Besides, despite there being 87 unique datasets available from the regional government, more than 30 of the datasets are documentation for freedom of information requests, councillors' contact information, or attendance sheets for council meetings. Although there is only one open data portal for the four municipalities with all their datasets, some regional services such as the Waterloo Regional Police Services, Grand River Transit and Grand River Conservation Authority also operate separate open data portals and only provide their data on it exclusively.

Despite sharing one same open data portal, each municipality in Waterloo region still follows their own data policies and guidelines, creating data disparities, such as varying data classifications (e.g., varying land use categories, some datasets containing comprehensive information while others focus on single aspects). These distinctions can create confusion and add inaccuracies during the data integration process. The diverse data sources may introduce unnecessary information, as users may struggle to cross-reference between datasets and determine what information to retain. Data users might face numerous similar datasets that need to be manually filtered during the search before proceeding into the integration process, requiring a significant understanding of the data to be aware of compatibility. Unfortunately, some vital datasets can be overlooked during the search due to missing or incorrect tags or metadata descriptions (Miller, 2018). Rahm's (2016) research also suggests that even with the substantial need for data integration in numerous research and study contexts, the process remains primarily manual due to different formats and logical heterogeneity from various purposes for which datasets were originally created.

The data quality can be changed after integration with the datasets by different standards. As an example shown in Figure 1., the trails dataset provided by the Region of Waterloo has the most recent update on March 2024 compared to the trail datasets provided by the City of Kitchener and City of

Waterloo, which was last updated on November 2023 and March 2020, it can be seen that there are still a lot of trails missing after the integration.

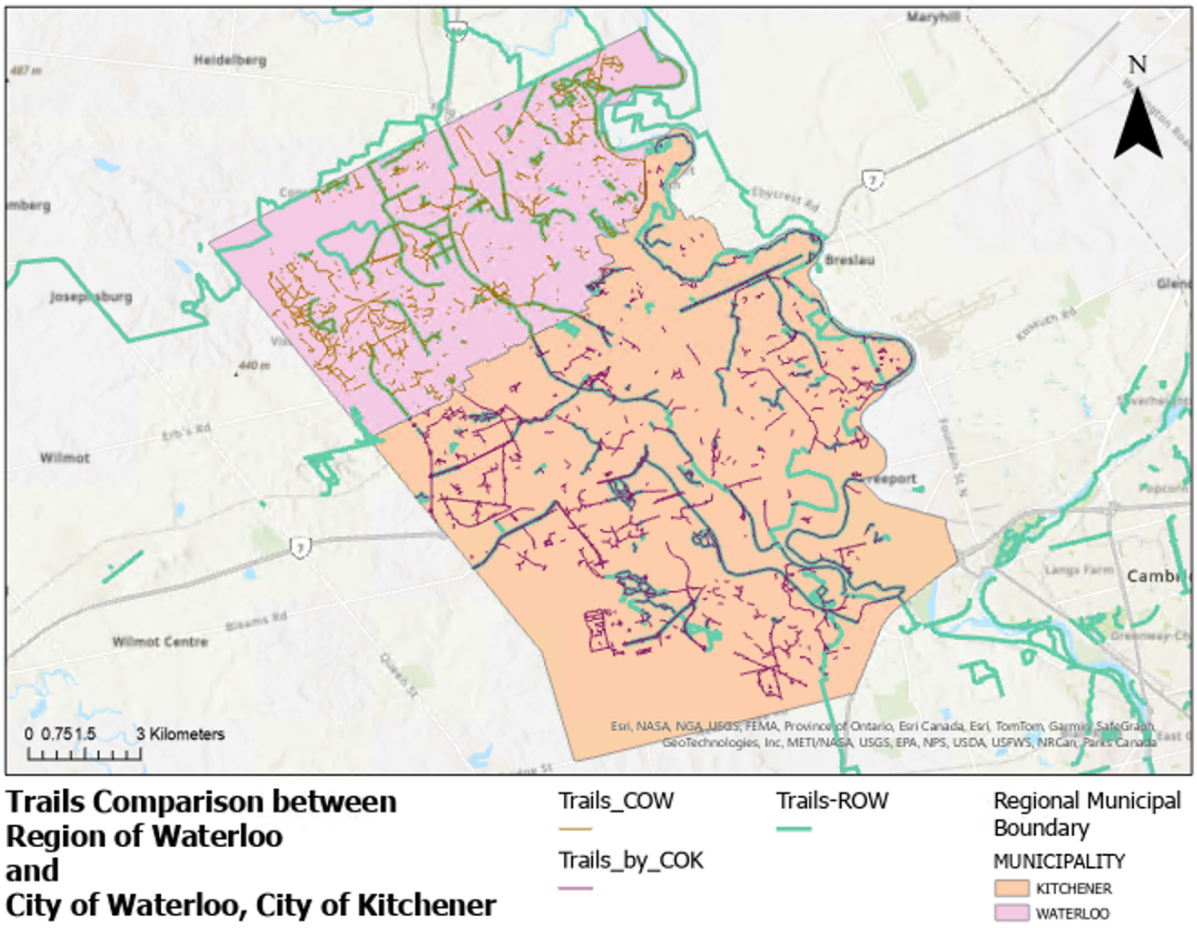

Figure 1 Map of Trails data comparison between the Region of Waterloo and the City of Waterloo, City of Kitchener

Although there are many diverse types of open data, including financial data, imaginary data and geospatial data, the research scope of this study is the geospatial datasets in the data portal of the Region of Waterloo.

Besides, there are also concerns about the out-of-date data policies in the Waterloo region. For example, the City of Waterloo still uses the Sunlight Foundation's "Ten Principles for Opening Up

Government Information" from 2010 (City of Waterloo, 2013). Although Sunlight Foundation's principles emphasize data quality, including completeness, timeliness, accessibility and machine readability, there is no detailed information about how the City of Waterloo measures each metric to ensure the quality of their data. Based on the information on the city's open data standards, the focus has been put towards accessibility and re-usability particularly. Nevertheless, reusing low-quality open data can only result in a further downgrade in data quality.

1.5 Research Goals

This study aims to understand how the open government data program works in a region with two-tier municipalities in Ontario by evaluating the range and quality of their open data programs and assessing their impact on cross-jurisdictional data initiatives. Additionally, this study seeks to determine the factors that might influence the quality of their open data and recommend potential improvements for the future. The study was conducted based on the following objectives:

1) Conduct a literature review to identify existing gaps of open data quality and the data quality evaluation system, explaining how to effectively measure open data quality using a quantitative approach.
2) Develop a comprehensive evaluation tool tailored to the municipal context for assessing the data quality of open data in the Region of Waterloo, the City of Waterloo, the City of Kitchener, and the City of Cambridge.
3) Present the results of the evaluation, identify any gaps, and formulate recommendations for enhancing cross-jurisdictional open data provision in Waterloo Region.

Chapter 2
Literature Review

2.1 Issues in Open Data

Open data has rapidly advanced in recent years, offering the potential for various industries to analyze and utilize data in diverse ways, benefiting both businesses and the general public (Mayer-Schönberger & Zappia, 2011). However, many open data programs are still measuring their success solely by the number of datasets released (Mergel et al., 2018). The strategy for quantity over quality often leads to open data being published without proper quality control, which can impact the accuracy of the dataset and hinder future research and studies (Bonaguro, 2015; Vetro et al., 2016). Poorly documented data requires more effort to process and understand, which increases the cost of interpreting the data and wastes resources (Sadiq & Indulska, 2017). On the other hand, as Vetro et al. (2016) indicated, a significant focus has also been put towards developing data-sharing platforms, misplacing the priorities of ensuring the basic quality of its products. Furthermore, most current spatial infrastructures lack meaningful connections between users and datasets, such as a feedback system, which means that these datasets miss out on opportunities for constructive criticism that could help improve data quality and accuracy (Zuiderwijk & Janssen, 2015).

Charalabidis et al. (2018) argued that the quality of open government data is just one aspect of a broader need, emphasizing the importance of government data policies as the key instruments for promoting transparency and convenience for citizens. Despite such policies existing across various government levels, their implementation often falls short of expectations. It is also worth noting that certain government departments and agencies may not fully understand the extent of open data policies when publishing their data, which is difficult for the open data implementation as they have no understanding of the value of OGD that can benefit the data users (Ubaldi, 2013). One of the main

implementation challenges lies in disclosure policies, often driven by copyright limitations, which can restrict further data reuse.

Although the focus on opening the government and sharing government data in Canada began to gain momentum in the 2010s, particularly with the Canadian Federal Government's launch of its open government strategy (Government of Canada, 2019), there is still a notable absence of an evaluation system for measuring open data quality in the Canadian context. Specifically, none of the existing systems and tools can fully cover every open data program, meaning new tools must be developed based on local cases. This lack of an evaluation system can lead to variations in the quality of open data, including issues such as outdated, incomplete, and inconsistent data. These issues, in turn, can further diminish the overall data quality and discourage users from accessing open data (Bertot et al., 2010).

2.2 Data Principles

According to previous research, there are some principles which should be followed when determining data quality metrics (Kaiser et al., 2007; Vetrò et al., 2016). The first principle is measurability, which states that metrics should be normalized or scaled so different metrics can be compared more meaningfully. For example, the score of each quality metric should be able to be converted into the same range as 0-1 to minimize the errors or the bias that might exist during the comparison. For the metrics with only two possible values, the value should also be converted into 0 or 1 to allow a quantitative comparison. The second principle suggested by Kaiser et al. (2007) is the interpretability of the metrics, meaning each metric should be easy to understand by its users, including data owners and data users. For example, since some of the metrics can be abstract and qualitative, the quantification of these metrics will input a large amount of bias into the study, which contradicts the purpose of a quantitative approach. Additionally, the data principle (Vetrò et al., 2016) argues metrics should be quantifiable not only at attribute level, but also tuple level and even dataset level. It is hard to

compare data metrics when they are not able to determine a numeric value. Moreover, the metrics for the data quality evaluation should be adaptable, fitting into the context of the datasets.

2.3 Evaluation Models for Weighted Metrics

Data quality evaluation models can be divided into two types: subjective model and objective model. For the subjective model, it is usually done in the form of interviews or surveys. Studies by Hernandez (2020), and Härting & Lewoniewski (2020) both interviewed and surveyed dataset owners and data experts from open data teams to determine the weight scale for their evaluations as they may have a better understanding of the dataset. Interviewees heavily influence the weight distribution of metrics within the quality evaluation due to their subjective judgements on the importance of each metric (Hernandez, 2020). Another subjective model used by Zuiderwijk et al. (2015) gathered feedback from data users to determine what characteristics are more important for data quality. The research conducted by Belhiah & Bounabat (2017) also suggested that the data provider, who has a good understanding of the dataset, should specify the importance of each metric.

Similarly, the Luzzu method proposed by Debattista et al. (2016) suggests allowing users to rank their own preferences. There are three perspectives in the Luzzu method, one of which is determining the data quality by allowing users to rank through user rankings on metric, dimension, or category. The metric preference by the Luzzu method requires users to set a detailed weight for each metric of data quality; while when the users do not have a specific understanding of the detailed metric, they can also provide preference for the data quality dimensions, which summarizes the data quality metric in a few more generalized areas. However, the weight of each metric under the dimension will be set equally. Lastly, the category method is comprised of multiple dimensions in one setting. Similarly, the weights will be distributed equally for the dimensions and metrics under it. The Luzzu method is one great way to evaluate the quality of datasets for data scientists or experts who have personal standards for quality

measurement. Nevertheless, this method will require a very detailed definition and classification of each metric. This requires a much deeper understanding of the metrics for not just the data team but also users, adding more complexity and not ideal for general use.

Conversely, objective models usually treat all quality metrics equally. The data quality evaluation by International Business Machines Corporation (IBM) uses a combination of pre-defined rules to monitor if the data in each cell meets the data quality metrics, while the final data quality score is computed based on the average score of all metrics (IBM, 2021). The data users or decision makers might not have specific knowledge of the data subjects to help them identify the needs of data quality, causing the results to contain bias from human errors (Haeberer, 1993). Despite the objective models can assign an equal weighting factor for all the quality metrics, providing a basic data quality score calculation for the data evaluation, the result might also lack the focus of specific data quality metrics, misplacing the priorities of the data evaluation.

2.4 Existing Research

To create a proper evaluation tool, it is crucial to identify the right metrics for measuring open data quality, as some metrics from other studies may not be suitable for every case. Many literature and previous studies proposing evaluation tools have mentioned metrics such as metadata, timeliness, completeness, consistency, accuracy, accessibility, coherence, credibility, relevance, interpretability, granularity, and usability. For instance, Vetrò et al.'s (2016) research considered traceability, currentness, expiration, completeness, compliance, understandability, and accuracy as their metrics for data quality assessment, with a focus on measurability, interpretability, aggregation, and feasibility. Similarly, the European data portal (2014) has identified dimensions for open data quality, including metadata, accuracy, consistency, availability, completeness, conformance, credibility, processability, relevance, and timeliness.

In an attempt to evaluate the quality of open data in the early years, Berners-Lee (2006) proposed a 5-star deployment scheme to rate open data, focusing on aspects like data licensing, machine-readability, format, and links. In this system, a higher-rated dataset would exhibit better structure and provide more context in an open format, characteristics considered advantageous for improving data quality during the early stages of the open data movement (Kim & Hausenblas, 2015). However, with shifting expectations regarding open data quality today, Berners-Lee's 5-star rating system no longer aligns well. Current emphasis has shifted toward factors such as completeness, timeliness, and accuracy, resulting in situations where low-quality open data can still receive a 5-star rating in his system (e.g., data that may be out-of-date but still receives a 5-star rating).

Another example of an evaluation system can be found in the Global Open Data Index, developed by Open Data International (Open Knowledge Foundation, 2021). This index assesses the openness of open data through three categories: data characteristics, aggregation levels, and time intervals, comprising 11 questions used to calculate scores for each open dataset. However, a challenge with the open data index's evaluation system is that it primarily reflects the openness of the data rather than its quality. Consequently, it is possible for low-quality data to achieve a high score in the index.

Ulbaldi (2013) also proposed a framework for measuring open data based on factors such as availability, reusability, cost, and demand. This comprehensive framework spans multiple disciplines, including policies, technical aspects, governance models, organizational efforts, and data itself. For policies, the framework examines disclosure policies, data standards, and legislation related to privacy and re-use policies. The technical aspect focuses on the accessibility for machine readability and the reusability of the data. This governance model examined the workflow of how raw data is approved and published, while organizational efforts concentrate on infrastructure and potential consequences when issues arise.

The evaluation tool developed by Viscusi & Spahiu (2014) aims to determine the quality of open government data using an empirical method. The data quality metrics were pre-defined to help select the samples to evaluate the specific situation of open data in the institutional web portal of the Italian government. Although Viscusi's model uses a ranking method for measuring metrics of completeness, accuracy, and timeliness, the score of each metric is determined by a more qualitative approach, the metric/dimension receives a certain score once they meet the condition similar to Berners-Lee's 5-star approach (2006). For example, unlike normal completeness, which evaluates the missing values in the dataset, completeness in Viscusi's tool is measured by the data availability and linkability (Viscusi & Spahiu, 2014). The accuracy dimension in Viscusi's model is a bit different than the data accuracy, which is more related to the availability of data formats such as CSV, XML, and JSON, measuring if the datasets can be processed automatically by programming and statistical tools. Timeliness is determined simply by whether the data is up to date. These three metrics were then used to calculate a mean degree of compliance, which served as an indicator of data quality for the data portal from each region in the Italian government. However, it is noteworthy that this method is primarily suitable for studies and research conducted at the portal level since it measures every dataset in an open data portal, which may not be the ideal model for all contexts as the Viscusi & Spahiu (2014)'s ranking method rates each metric by whether achieving certain characteristics, instead of reflecting the details of each metrics in a quantitative measurement.

The research done by Quarati (2023) evaluated open data differently, which was focused on the different types of platforms, total views, and other site visit-related metrics. This approach might provide some insight into how the open data programs have been running, yet it was only able to evaluate the performance of the open data government portals instead of accurately reflecting the quality of the open data programs or the actual usage of the open data.

From these approaches, it can be inferred that most data quality metrics in a proper evaluation system should ideally be quantitative and normalized. Since different open data programs might use diverse scales for their measurements, normalizing these metrics allows for straightforward and meaningful comparisons. Additionally, these metrics should be meaningful and provide enough information for understanding and interpretation. Furthermore, they should be feasible to determine, as efficiency in processing time is essential (Kaiser et al., 2007). For instance, the European Data Portal's evaluation method is based on a rating system, wherein different points are allocated to metrics such as findability, accessibility, interoperability, reusability and contextuality (European Union, 2021). The findability measures the ease of finding the dataset, including the impact of keyword usage, categories, and spatial information. Accessibility describes if the dataset is accessible through the URL provided as the European Data portal contains more than 1 million datasets from 36 countries. It is noteworthy that the European data portal focuses more on interoperability, which assesses machine-readability and format, and less on contextuality, which focuses on creation date, modification date and rights. The reusability metric is comprised of license information, access restriction, contact and publisher information, which is similar to the metadata metric in another research.

2.4.1 Open Data Quality Evaluation in Canada

Several Canadian regions and cities are testing evaluation systems for their open data quality. For instance, the City of Toronto has developed an open data master plan that incorporates an open data evaluation framework to measure data quality. They also follow Sunlight Foundation's "Ten Principles for Opening Up Government Information", much like the City of Waterloo. However, unlike the City of Waterloo, the City of Toronto has an open data evaluation framework which uses five main criteria to evaluate their open data quality: usability, metadata, freshness, completeness, and accessibility (Hernandez, 2020). The usability score weighs most of the five main criteria, accounting for 38% of the data quality score. It examines whether the field names of datasets provide meaningful names to measure

ease of use. The metadata score, weighing 25%, is the second most important characteristic in their framework and it measures if the metadata is filled out. The other three metrics of freshness, completeness and accessibility are more straightforward as they measure if the data is up to date if the amount of non-missing value, and if the data is accessible via API. They account for 18%, 12% and 7% of the score respectively. The framework then categorizes datasets into three levels—gold, silver, and bronze—based on their data quality score and the potential impact in addressing civic issues (Rayes, J., & Mahmood, S., 2020). The framework has to exclude certain data quality characteristics, such as accuracy and coherence, which rely on the data publishers (e.g., individual departments who produced the data) as the open data team may not possess the expertise or knowledge of the data itself. This is due to the decentralized data governance framework they use, where data originates from each department. This model makes it challenging to control data quality since different departments may have varying standards or opinions about what constitutes high-quality data. Consequently, there is no clear focus on the accuracy or validity of a dataset, implying that a gold-level dataset could still contain inaccuracies and logistical errors.

York Region employs a data stewardship model to address data quality issues at their roots. This model includes a data owner, a data manager, an open data board member, and a privacy officer. With this centralized data governance model, each dataset undergoes multiple checks and approvals before publishing to ensure its best quality. The stewardship model also places significant emphasis on data freshness, as data quality can change over time. To engage the public more effectively, the York Region has introduced open data into education, aiming to familiarize the younger generation with open data and gather constructive feedback. Furthermore, York region follows six data principles from the Open Data Charter, focusing on openness, timely, accessible, and usable, comparable, and interoperable data, improved governance & citizen engagement, and inclusive development & innovation (Open Data Charter, 2015). Despite York region's efforts of actively engaging with citizens and seeking feedback

from the public, there is no direct and intuitive complete feedback system apart from the comment feature under the ArcGIS Online platform, which is only available on the detail page of each map layer.

The Canada Spatial Data Infrastructure (CSDI) has also created a data quality guideline to help to identify the elements that can be measured from the geospatial data (Natural Resources Canada, 2016). The guideline indicates that geospatial data quality can have multiple characteristics for the same metrics, for example, logical consistency can be divided into conceptual consistency, domain consistency, format consistency and topological consistency (ISO, 2023; Natural Resources Canada, 2016)

Chapter 3
Evaluating Data Quality in a Quantitative Approach

To identify the current state of open data in Waterloo region, quantitative research is required to produce a more objective data quality score. Unlike some research that evaluates data quality by using a qualitative-based approach (Viscusi & Spahiu 2014), the datasets for this study were selected by employing a pre-defined data quality metric. Just like other existing research (Hernandez, 2020; Ulbaldi, 2013), the methodologies used for evaluating data quality in this study would be determined based on a series of quantitative metrics of data quality dimensions. It should be considered that some metrics like timeliness or completeness can have multiple characteristics. Some metrics only reflect unique aspects of data correspondence to the real-life object (Batini et al., 2009). Besides, with a quantitative approach for evaluating data quality in the Region of Waterloo, similar methods and metrics can be used for future studies when more datasets are available. As the main research goal of this study was to create a data quality evaluation tool tailored for the open data in the Region of Waterloo, the focus of the evaluation would be based on the metrics and the available datasets that can reflect the status of open data programs in four municipalities the best. This chapter is divided into 3 parts, including data selection, metric selection & definition, and weight determination.

This chapter is divided into 4 parts: discussing the use of ArcGIS Hub as an open data platform, the dataset determination, the metric development, and the weight configuration to describe the methodologies used for the data quality evaluation in this study and the rationale behind them.

3.1 Open Data on ArcGIS Hub

ArcGIS Hub is a data-sharing platform created by Esri that can help organizations provide data, tools, and other information to users within and outside the community (Esri Inc., 2024). The most common feature of ArcGIS Hub is to share open data, making data accessible to the users. Besides,

ArcGIS Hub allows organizations to create surveys, engage with the public and collect crowdsourcing data. ArcGIS Hub can be used to create dashboards and other infographics to highlight valuable information. Moreover, it enables collaboration between organizations and communities, allowing them to display their datasets and initiatives (Esri Inc., 2024).

ArcGIS Hub is an essential part of this study as the open data portal for the Region of Waterloo was created by using the ArcGIS Hub, which is also shared by the City of Waterloo, the City of Kitchener, and the City of Cambridge. Despite the reason to share one platform as stated by the Region of Waterloo (2024) is to make all data more accessible to the public, which could also be the excessive cost of Esri products, there were some obstacles encountered during the data evaluation due to the use of the ArcGIS Hub for open data. ArcGIS Hub provides conveniences and lowers the difficulty for smaller-sized governments to create and publish their open data program online by providing a mature framework. However, in the meantime, it also prevents further development of the open data program due to the limitation of its functionality and design compared to other data-sharing platforms such as ckan.

3.2 Data Determination

To find out what datasets are suitable for the open data quality assessment among the lower-tier municipalities in Waterloo region, it is important to decide the criteria before moving into the data selection process. Besides, how to analyze the datasets effectively is also a challenge for this study. This section is divided into two parts, the first part discusses the criteria for the dataset selection and the rationale for choosing them. The data selection and extraction process are discussed in the following sections.

3.2.1 Data Availability

It is important to know what datasets are available in Waterloo Region before diving into data selection. Despite there being around 394 datasets in total on the data portal, to make the evaluation result more objective, the datasets selected should be from all four municipalities. However, the limited

availability of the open data provided by the City of Cambridge and the City of Waterloo as shown in Table 1. (the number is provided by ArcGIS Hub), makes the sample size even smaller as the study aims to evaluate the datasets in common shared by all entities to have an intuitive comparison and objective result.

Municipality	Region of Waterloo	City of Waterloo	City of Kitchener	City of Cambridge
Number of Datasets	87	116	142	49

Table 1 Number of datasets provided in Waterloo Region

3.2.2 Data Catalogue

Data availability was not the only obstacle to the data selection process, as the Region of Waterloo, City of Waterloo and City of Cambridge did not provide any list or catalogue for the details of their datasets. To find what datasets are provided on the open data portal, a separate data catalogue had to be created as the ArcGIS Hub provides no functionality to list all the datasets in the portal either. Despite the City of Kitchener providing a publicly accessible data catalogue[1] to list all their datasets, their data catalogue was outdated (last updated in 2019) and did not contain all datasets currently listed on the open data portal. The same challenge goes for the metadata document[2] from the City of Kitchener as well, many details were left unclear or missing, especially for some important fields such as license, restrictions, date, and update frequency information. For example, in the City of Kitchener's metadata document, the update frequency is usually described as continuous even though continuous only describes the data that would be updated regularly but does not indicate how often. Information like this does not

[1] Existing data catalogue file provided by the City of Kitchener (https://open-kitchenergis.opendata.arcgis.com/datasets/aac0a59ce85b44a395e34285f9f73d12_0/explore)
[2] Metadata file provided by the City of Kitchener (https://maps.kitchener.ca/OnPointExternal/opendata/metadata/opendatahome.html))

provide a meaningful contribution to the metadata since metadata is supposed to help users to understand the dataset.

The catalogue created in this study collected the title, format, and publisher of each dataset by manually inputting from the dataset's detail page on the data portal. Despite the four municipalities sharing only one platform, there was also an issue due to the different terminology used by each municipality when naming a dataset with the same topics. On top of that, there is also an issue with the indexing. Most websites provide their search results by indexing them with page numbers, so results will not overflow on one web page (Google Inc, 2024). However, in ArcGIS Hub, the search result will only continue to load at the bottom of the web page instead of listing datasets by indexing them with page numbers, which added difficulty to the catalogue creation process.

3.2.3 Selection Criteria

It is important to note that as the focus of this study is set to the geospatial data, the non-geospatial datasets in the Region of Waterloo's open data portal would be excluded from the data selection. Comparing the data quality of datasets from different municipalities in different subjects does not provide any meaningful result, as data in different subjects might have different focuses than the other. To have a more direct and intuitive comparison of how the data quality works cross-jurisdictionally in Waterloo region, the selected datasets should all be in the same subjects and available by every municipality as the City of Waterloo, City of Kitchener, City of Cambridge, and Region of Waterloo. For example, if the data quality score from the City of Waterloo is made of the scores of address and road data only, the data quality score from the City of Kitchener should also be comprised of the datasets in the same subjects to reduce the potential differences caused by the variables. By focusing on the common datasets from all municipalities, the goal was to keep consistency throughout the evaluation process, which would provide a more objective result for the open data quality assessment. Also, the datasets should include information such as metadata, timeliness, data format, and data rules to be compared

quantitively. Moreover, the datasets should have meaningful information related to their topics. For example, datasets like regional council meeting attendance provide no meaningful information for the evaluation of open data quality.

3.2.4 Data Selection

With the criteria set for the data selection, the datasets meeting the condition were filtered out. According to the newly created data catalogue for this study, there were datasets in 7 subjects available by the four municipalities, including addresses, building outlines, municipal boundaries, road closures, trails, roads and cycling paths. However, the municipal boundaries datasets were excluded from the selection for this study because they provide no meaningful attributes besides the attribute of boundary shapes for the data quality evaluation. Thus, a total of 24 datasets from four municipalities with each dataset from these 6 subjects was selected:

- Addresses
- Building Outlines
- Cycling paths
- Trails
- Road Closures
- Roads

Then, a second data catalogue was created for these 24 datasets with more detailed information including the metadata information as name, publisher, description, published date, last update date, metadata update date, license information, update frequency and whether the dataset contains update frequency or not, URL path of GeoJSON file as it is the only format can be accessed directly through web-based API as shown in Table 2. Despite ArcGIS Hub also allowing access to the metadata files using REST API (Esri Inc., 2020), the information had to be manually collected due to the lack of metadata

files provided on the data portal as the metadata information was provided on each dataset's detail page listed in different parts of the web pages, and the information like data type rules for each dataset was only listed with the description.

Title of Dataset	Publisher	Description	Published Date	Last Update Date	Update frequency	Contains update frequency?	License information	GeoJSON URL

Table 2 Example of the detailed catalogue of selected datasets.

This more detailed catalogue would help to determine whether the corresponding metadata exists. Besides, the title and the publisher information could help to distinguish datasets in different subjects and from different municipalities. The published date, last update date and update frequency could be used for determining the timeliness-related metrics. The GeoJSON URL allows accessing the actual dataset by using API, which ensures the datasets can be up to date with any recent changes. Besides, accessing data on the server is a cost-effective approach as it removes the process of transferring files to the local machine and the spaces used for the data storage.

3.2.5 Data Extraction for the GeoJSON data

GeoJSON is an open standard file format based on the JavaScript Object Notation (JSON) format to show the geographic features, which documented information such as geometry type, coordinates, and feature type (Esri Inc, 2024).

```
{ "type": "Feature", "properties": { "OBJECTID": 648665, "ACTIVETRANSPORTID":
    393807, "STATUS": "ACTIVE", "STATUS_DATE": "2023-02-08T12:06:31Z",
    "CATEGORY": "PATHWAYS", "SUBCATEGORY": "BMUT", "TRAIL_MASTER_PLAN_CLASS":
    null, "FEATURE_TYPE": "SURFACE", "SURFACE_MATERIAL": null, "ROUTE_NAME_1":
    null, "ROUTE_NAME_2": null, "ROUTE_NAME_PROPOSED": null, "WIDTH_M": 3.0,
    "GRADE": null, "RAILING": "N", "CURBCUT": "N", "STREET": "STRASBURG RD",
    "ROADSEGMENTID": 600655, "ROADSEGMENT_SIDE": "RIGHT", "PARCELID": 55075811,
    "WARD": 5, "SOURCE": "ROAD CONSTRUCTION PLAN", "SOURCE_DATE":
    "2017-09-01T00:00:02Z", "NOTES": null, "OWNERSHIP": "KITCHENER",
    "MAINTAINED_BY": "OPERATIONS (PARKS)", "INSTALLATION_YEAR": 2020,
    "LAST_INSPECTION_YEAR": null, "SURFACE_CONDITION": "GOOD", "CONDITION_DATE":
    null, "CONDITION_SCORE": null, "MAINT_ROUTE": 10, "SW_WEIGHT_CLASS": "",
    "SW_PRIORITY": 0, "PN_PROJECT_NO": null, "PN_MAP_CLASS": null,
    "PN_OFFICIAL_PLAN_MAP": null, "PN_REGION_CATEGORY": null, "CW_WORK_AREA":
    "15-1", "WINTER_MAINTAINED_BY": "NA", "GlobalID":
    "7480a4f1-ee08-4451-870d-3f3f5d349772", "Shape__Length": 41.638483944622401
    }, "geometry": { "type": "LineString", "coordinates": [ [
    -80.466912719594205, 43.3893613772102 ], [ -80.466989630340095,
    43.389306209724801 ], [ -80.467268849475502, 43.389091039144297 ] ] } },
{ "type": "Feature", "properties": { "OBJECTID": 648666, "ACTIVETRANSPORTID":
    393803, "STATUS": "ACTIVE", "STATUS_DATE": "2020-10-27T17:55:17Z",
    "CATEGORY": "PATHWAYS", "SUBCATEGORY": "BMUT", "TRAIL_MASTER_PLAN_CLASS":
    null, "FEATURE_TYPE": "SURFACE", "SURFACE_MATERIAL": "CONCRETE",
    "ROUTE_NAME_1": null, "ROUTE_NAME_2": null, "ROUTE_NAME_PROPOSED": null,
    "WIDTH_M": 0.0, "GRADE": null, "RAILING": "N", "CURBCUT": "Y", "STREET":
    "STRASBURG RD", "ROADSEGMENTID": 600655, "ROADSEGMENT_SIDE": "RIGHT",
    "PARCELID": 55104002, "WARD": 4, "SOURCE": "ORTHO 2020", "SOURCE_DATE":
    "2020-05-01T00:00:00Z", "NOTES": null, "OWNERSHIP": "KITCHENER",
    "MAINTAINED_BY": "OPERATIONS (PARKS)", "INSTALLATION_YEAR": 2019,
    "LAST_INSPECTION_YEAR": "2023", "SURFACE_CONDITION": "GOOD",
    "CONDITION_DATE": null, "CONDITION_SCORE": null, "MAINT_ROUTE": 10,
    "SW_WEIGHT_CLASS": "", "SW_PRIORITY": 0, "PN_PROJECT_NO": null,
    "PN_MAP_CLASS": null, "PN_OFFICIAL_PLAN_MAP": null, "PN_REGION_CATEGORY":
    null, "CW_WORK_AREA": "15-1", "WINTER_MAINTAINED_BY": "NA", "GlobalID":
```


Figure 2 Example of GeoJSON file of trails dataset from the City of Kitchener

As shown in Figure 2., a GeoJSON file stores information such as data type, properties, and geometry as one feature for each entry. Therefore, to obtain the data attributes under the properties from a nested list in the GeoJSON file for the analysis process, a script had to be created in Python by using a few functions from pandas and request modules. Since loading the GeoJSON file directly into the data frame will only bring a lot of extra information with it, the Python script was set to extract only the properties from the feature and then add them into the data frame by iterating through the GeoJSON file line by line. With the GeoJSON URL provided in the second catalogue, the script could collect the attributes from the 24 datasets and analyze them.

```
for x in researchdata_file['url']:
    response = requests.get(x)
    jdata = response.json()
    features = jdata['features']
    properties = [b['properties'] for b in features]
```

```
    df = pd.DataFrame(properties)
```

Figure 3 Python script used for extracting data attributes from GeoJSON files

However, there were a few challenges encountered during the data extraction process. Originally, the study proposed to use the data from the GeoJSON link provided by the open data portal to the data server to ensure the data is kept in a fresh and up to date manner. However, by examining the queried data during the data extraction, it was found out that there is a limit for the maximum data entries that can be queried on ArcGIS Hub each time, which is set to either 1000 or 2000 entries by default to prevent large-scale scanning and potential cyber-attack. To continue to obtain all datasets by this method, it is required to obtain the object ID (There is no limit for querying the object ID only) first from the server client, then use the object ID to query every 1000 entries per time (Esri Inc., 2023), which could add unnecessary complexity and time cost. Thus, the data extraction method was transited to download each dataset by using the static data provided at the time of evaluation, resulting in the datasets having to be re-downloaded every time there was a modification to the evaluation. Another challenge was due to some cells in datasets only containing whitespace characters, which makes them considered not empty by programming definition, adding more complexity to the analysis.

3.3 Metric Development

Before determining the metric, it is important to understand what the data quality is. The International Organization for Standardization (ISO) describes data quality as a function of the characteristics of data attributes based on the consideration of data users (ISO, 2022). Meanwhile, geospatial data quality is dependent on the compatibility of the datasets to the users' demand and application for GIS (Geographical Information Systems) purposes (ISO, 2023). Based on ISO's standard, the data quality metrics of the geospatial are completeness, logical consistency, positional accuracy, thematic accuracy and temporal quality (ISO, 2023; Nature Resources Canada, 2020). Despite ISO also

suggesting not to define a minimum level of data quality as the consideration of high data quality can be different due to the fitness for use of the data to the users/organization, the value of the intrinsic characteristic of data cannot be ignored either. Thus, this study proposed to evaluate the data quality based on the characteristics of the open government data in the Region of Waterloo.

As the open data programs are different due to policies, locations, resources, and the different goals each evaluation is trying to achieve, it is important to determine what metrics should be considered during data quality evaluation as different research also utilized different approaches to measure their open data quality metrics (Vetrò et al., 2016; Nikiforov, 2018; Hernandez, 2020). For instance, in Vetrò et al.'s (2016) research, the focus was separating metrics into quantitative and non-quantitative, so the programming scripts can be set up for the calculation of each quantitative metric to automatically evaluate the data quality with the new datasets in the future. Vetro's approach emphasized the importance of normalization of each metric by converting them into standard value ranges from 0-1 with weighting factors, highlighting the comparability between metrics.

3.3.1 Metric Selection

The first step is to determine what data metrics are available for use in a data quality evaluation. As shown in Appendix A, previous studies have concluded some data metrics that are commonly used in data quality evaluation, such as metadata, timeliness, completeness, logical consistency/coherence, accuracy, accessibility, creditability/reliability, relevance, interpretability, and usability.

Following one of the important principles proposed by Kaiser et al. (2007), the selected metrics for evaluating the open data quality in this study should be measurable, in other words as being quantitative. Thus, the metrics should be divided into two categories, quantitative and qualitative, as shown in Table 3. Furthermore, to make the open data quality evaluation more feasible, metrics like accuracy should not be considered for the quality evaluation. Not only because it is difficult to find the reference data as local governments are the only one who provides these data online, but also due to the

low-accurate data is usually collected and produced by the data owner. Thus, without any specific knowledge of the subject, accuracy is a difficult metric to assess. While some studies consider usability as a qualitative metric to reflect the user experience, usability in this study is considered quantitative to evaluate whether the dataset is available in different formats. Thus, the metrics chosen for open data quality evaluation in this study are metadata, timeliness, completeness, logical consistency, and usability, which are summarized into four main dimensions: metadata, timeliness, completeness, and usability for the evaluation.

Quantitative	**Qualitative**
Metadata	Accessibility
Timeliness	Relevance
Completeness	Interpretability
Logical Consistency/Coherence	Credibility
Accuracy	
Usability	

Table 3 Quantitative metrics vs. Qualitative metrics

3.3.2 Metric Definition:

As indicated by Sebastian-Coleman (2013), certain data quality dimensions like completeness, timeliness, and logical consistency can be unclear, which would require more attention for the detailed classification and explanation to justify how each data quality metric was measured and evaluated. For example, metrics like completeness can be re-classified into column completeness, which indicates the degree of how complete a dataset is, and schema completeness or coherence, showing the consistency of data that follows the existing dataset/database rules. Table 4. offers an outline of the characteristics that were focused on each dimension and metric, and a more detailed definition is listed in the following sub-sections. It is essential to understand that the individual metrics used in this study evaluate the possible characteristics of each metric, rather than covering all characteristics of them.

Dimension	Data Characteristics
Metadata	Completeness: How many fields of metadata information are provided for the dataset
Timeliness	Currency: How recent does the dataset reflect the real-world object
	Frequency: Does the dataset have continuous or discrete updates
Completeness	Column Completeness: what portion of the data is filled with values, instead of missing values
	Schema Completeness: Does the actual data follow the dataset rules to keep them coherent or logically consistent
Usability	Accessibility (API-accessible): if the dataset can be accessed directly on the server without the need to download or transfer to local devices
	Readability (Machine-readable Formats): if the dataset provides any machine-readable formats

Table 4 Data quality dimension and characteristics

a) Metadata

Metadata plays a crucial role in open data as it describes the content of a dataset and ensures its credibility (Kubler et al., 2018). Furthermore, missing or incorrect metadata can hinder the discoverability of the dataset, creating barriers to open data quality (Neumaier et al., 2016). Metadata has many distinct aspects that can be measured like DCAT (Data Catalog) vocabularies and accessibility. For example, the metadata for the City of Toronto's open data can be accessed through web-based API. However, since the metadata for these datasets does not come with the dataset or separate file, besides the City of Kitchener's datasets, it added more complexities and limited the potential methods that can be used. According to the ISO standard 19115 (ISO, 2014), a metadata file should include the dataset title, date (publication and revision dates of the dataset), point of contact (the dataset's author or owner), description, maintenance (update frequency), category, and constraints (license). Similarly, both the government of Canada and the

government of Ontario specify metadata should include title, contact, description, keywords, update frequency and license information (Treasury Board of Canada Secretariat, Government of Canada, 2020). Thus, the metadata for this study will focus on the presentation metadata information such as title, publisher, description, creation date, last update date, metadata update data, update frequency and license. In this study, the metadata score will be calculated based on its completeness using the following function:

$$m_{Metadata} = \frac{existing\ metadata\ fields}{Total\ metadata\ fields}$$

Equation 1 metadata score calculation

Where the metadata score is calculated based on the number of metadata fields filled divided by the total required fields.

b) Timeliness

Timeliness is another vital dimension for assessing open data quality, as open data remains relevant only when it is up to date (Nikiforova, 2020). While many previous studies have focused on measuring the overall timeliness of open data portals rather than individual datasets (Nikiforova, 2020; Atz, 2014; Neumaier et al., 2016), researchers such as Viscusi & Spahiu (2014) and Candela et al. (2020) have suggested that timeliness should not be assessed solely by checking if the data is currently up to date; it should also consider the dataset's update frequency as part of the metric. Therefore, in this study, timeliness is measured by a score that combines the currency of the dataset with its update frequency. However, due to some datasets having no update frequency information provided, the study had to assume the update frequency of these datasets is the same as the other datasets in the same subject. The currency score can be calculated by modifying the timeliness equation proposed in Atz (2014), the original timeliness function is an indicator that assigns a value of 1 to the data up to date and 0 otherwise. The modified function for this study is shown in the following equation:

$$m_{Currency} = min\left(1, \quad \frac{update\ frequency}{current\ date - date\ of\ last\ update}\right)$$

Equation 2 Currency score calculation

where the currency is calculated by having the number of days of update frequency divided by the time length in days between the current date and the date of the last update. For the datasets with update frequency listed as "on demand" or "as needed", to quantify them, the study made another assumption to consider their update frequency as bi-annually (every two years).

For the frequency score, a similar approach according to the measurement by Candela et al. (2020)'s research can be adapted. As Candela et al proposed, the update frequency should be considered as four kinds as shown in Equation 3., 1) dataset with continuous update, 2) dataset with discrete but periodic update, 3) dataset with discrete but non-periodic update, and 4) other situations.

$$m_{Freq} = \begin{cases} 1 & continuous\ updates \\ 0.5 & discrete\ periodic\ updates \\ 0.25 & discrete\ non\ periodic\ updates \\ 0 & other\ situations \end{cases}$$

Equation 3 Frequency score by Candela et al. (2020)

With the consideration of open data in Waterloo region, it is difficult to find out whether a dataset has periodic updates or not because ArcGIS provides no functionality to view the previous updates. Additionally, the lack of update frequency information provided by the Region of Waterloo, the City of Waterloo, and the City of Cambridge was a major factor in making the original function impossible.

The purpose of the update frequency score is to distinguish continuous updates and discrete updates. The original formula was adjusted and simplified as the following equation to fit the case of open data in Waterloo region:

$$m_{Freq} = \begin{cases} 1 & Continuous\ Updates \\ 0 & Others \end{cases}$$

Equation 4 Frequency score calculation

In this function, a dataset with continuous update, meaning if the time range of the last update date to the current date is smaller than its update frequency, would receive a score of 1 for the frequency score. For the dataset with discrete updates or other scenarios, meaning the dataset might have updated in the past, but the time range of the last update date to the current date is beyond its update frequency, would score 0 as the frequency score.

Another reason for the frequency score is that the currentness is not the only factor contributing to the timeliness. For example, a dataset can be published recently, holding a high currency score, but its update frequency can be weekly. So, if the dataset was updated more than a week ago, it can still have a high currency score, but also not have a continuous update. By adding the frequency score, data with continuous updates will be rewarded with a higher timeliness score.

However, as the frequency score is used as a factor for offsetting the timeliness score of the datasets that do not update based on their update frequency, it should have much less impact compared to the currency score. Thus, the total timeliness score calculation should be comprised of 2 parts of the currency score and 1 part of the frequency score.

c) Completeness

Completeness is a metric that indicates the extent to which data is meaningful and not missing (Vetrò et al., 2016; Pipino et al., 2003). A study conducted by Candel et al. (2020) proposed that completeness can be broken down into three components: schema completeness, column completeness, and population completeness. Schema completeness can also be referred to as logical consistency/coherence, which assesses the proportion of the fulfilled database schema within the dataset. Column completeness is very straightforward, which measures the presence of data with missing values in each column. And lastly, Population completeness evaluates the extent to which potential missing data is covered in the dataset.

It is impractical to verify the population completeness without a deep understanding and knowledge of a specific dataset. Therefore, the proposed method in this study focuses on measuring column completeness and schema completeness for the completeness categories. The column completeness would be calculated by the following function:

$$m_{Completeness} = \frac{completed\ cells}{total\ cells}$$

Equation 5 Column completeness score calculation

Where the completeness score is calculated based on the cells filled with contents divided by the total number of cells.

The schema completeness looks to see if the rules in the dataset are followed, the study measures it based on how many columns in each dataset followed the data type rules defined for its field, which is defined by the following function:

$$m_{Schema} = \frac{number\ of\ consistent\ columns}{number\ of\ total\ columns}$$

Equation 6 Schema completeness score calculation

Where the number of columns following the pre-established data type rules would be divided by the total columns, the result will be schema completeness.

d) Usability

The last dimension is usability. In different studies, the metric of usability is defined in several ways. For instance, Osagie et al. (2017) referred to it as the measure of understandability and learnability, while Hernandez (2020) considered it a metric reflecting meaningfulness and accessibility. Slibar et al. (2018) described usability as the degree to which proper data is used within a dataset. Considering these various definitions and examples from other studies, this study defines usability as a combination of two factors: accessibility and readability. Accessibility is a simple Boolean statement that returns true/false

results depending on whether the dataset is API-accessible. To adapt this into quantitative research, a score of 1 will be given to the dataset that can be accessed directly on the server with the link provided, and a score of 0 will be given in other scenarios. Similarly, the readability is also measured by a Boolean statement as if the dataset provides machine-readable formats or not, such as CSV, JSON, and XML. A score of 1 will be given to the dataset that provides any machine-readable format, and 0 will be given otherwise.

3.4 Weights Configuration

With the metrics being set for the data evaluation, the next step is to determine the weight for each metric. Despite there being literature for evaluating specific data quality characteristics, currently, there is no standard way to determine the weight for data quality evaluation due to the fact of how each open data program is managed, and how many resources are devoted. Besides, as the focuses of each data quality evaluation are different, some might focus on one single aspect, such as timeliness (Atz,2014), or linkage (Debattista, et al., 2016). While others might focus on the perspective of overall quality (Härting, & Lewoniewski, 2020).

As mentioned at the beginning of this chapter, this study proposed to use a more subjective model to determine the weight for the open data quality evaluation as the subjective model can emphasize the importance of specific data metrics, which can be used for reflecting the missing quality dimension in the current open data program at Waterloo region. Similar studies have suggested a weighted sum model would be the best fit to calculate the data quality score, which normalizes the score from each data metric indicator and applies the pre-determined weights to them to calculate the final data quality score (Hernandez, 2020). The weight of each data metric can be summarized with the existing data from some studies. An example of the existing weights determined by the City of Toronto has put the most focus on both metadata and timeliness equally at 0.35 each. For the timeliness dimensions, the score of timeliness was calculated based on more direct and objective data of the update frequency and the last updated date

provided, earning it a more influential weighting factor of 0.4. While the metadata used in this study was collected based on the information provided on the ArcGIS Hub page, which could result in potential systematic errors that promote the completeness of metadata information, the influences of metadata should be focused less as 0.3 for calculating the data quality score. Despite the column completeness represents the degree of how information is presented within the subject, it only evaluates the missing data instead of also focusing on the accuracy or meaningfulness of the data, the metric does not hold a significant impact on the overall data quality score calculation. As the survey conducted by Hernandez (2020) suggested the column completeness should be weighted around 10%. While weighted model used by Vaziri et al. (2019) indicated that logical consistency (schema completeness) has a low impact on the overall data quality and assigned a weight of 0.1 to it. Previous studies (Hernandez, 2020; Elouataoui et al., 2022) both suggested assigning the lowest weighting factor to Machine Readability and API-accessibility as they are considered ease-to-manipulate metrics as many platforms like ArcGIS Hub would provide these functionalities to its users. Thus, earning them the least weight as 0.05.

Dimensions	**Metric:**	**Weight**
Metadata	Metadata	0.3
Timeliness	Currency	0.267
	Frequency	0.133
Completeness	Column Completeness	0.1
	Schema Completeness	0.1
Usability	Machine Readability	0.05
	API-Accessibility	0.05
Total:		1

Table 5 Weighting Factors of each data quality metric

Chapter 4
Results and Findings

This chapter discusses the results of the evaluation of the open data in Waterloo Region. These results were derived from implementing the methodology proposed in Chapter 3. This chapter first describes the result from the selected dataset by each quality metric indicator to describe the details of what was successful and what was failing, and some of the challenges encountered during the analysis process. Then, the overall result by each municipality to discuss the potential issues and efforts being put in by each municipality in Waterloo Region. Lastly, there is a data quality comparison between the three lower-tier municipalities the City of Waterloo, the City of Kitchener and the City of Cambridge and the upper-tier municipality the Region of Waterloo to conclude the cross-juristically open data quality.

Based on the data analysis, the average data quality score from the evaluation is 0.7495. As shown in Figure 4., the final data quality score can be compared by four municipalities in 6 subjects of datasets. The detailed breakdown of each individual metric score can be found in Appendix B.

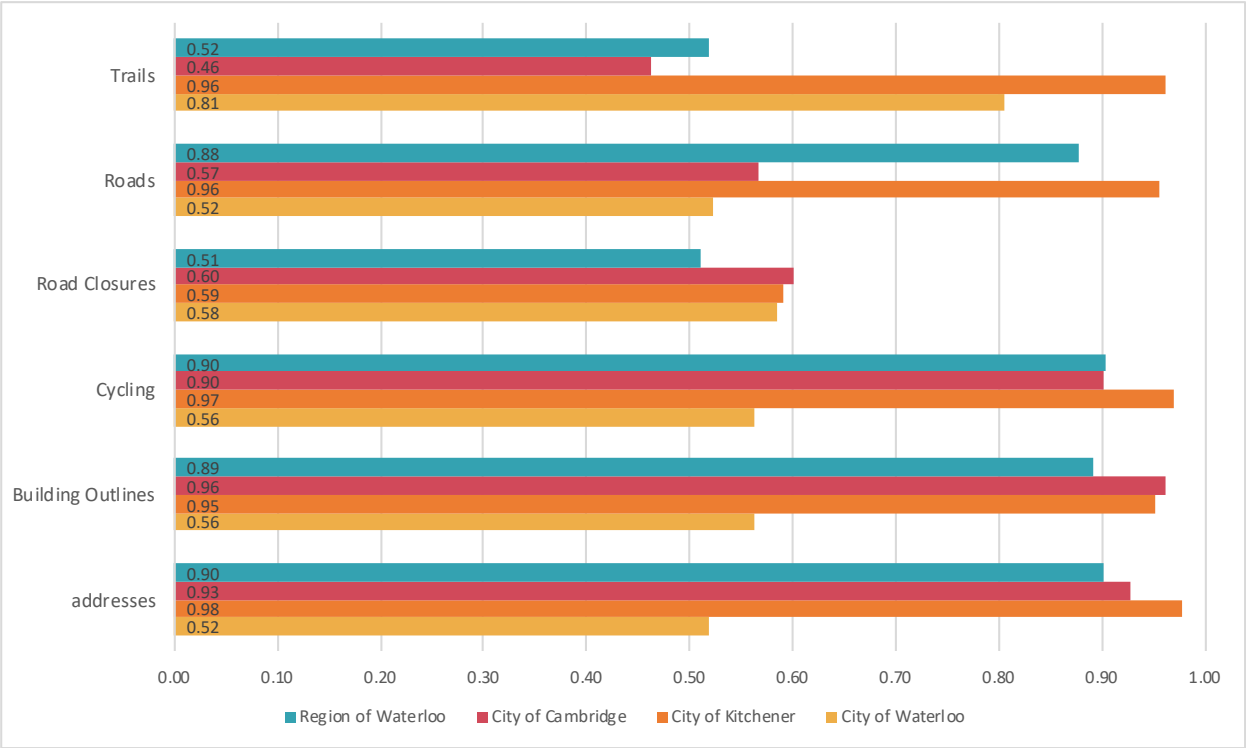

Figure 4 Final data quality score in 6 subjects by each municipality

4.1 Individual Metric Indicator

This section discusses each metric indicator, including what characteristics were measured and why each score was earned, as well as how each municipality scores in them.

4.1.1 Completeness Score

The completeness score is a combination of column completeness and schema completeness. The average completeness score in this study is 0.88.

Since the column completeness is a metric that shows the degree of showing the opposite of the missing values. The score is calculated by how much data is filled with information with the total number of cells in the dataset. The score calculation noted the issue mentioned in Chapter 3 as some datasets containing blank whitespace characters. A Python script was used to detect if a whitespace character

exists in the data value and if the data length of the value is also 1 to remove these "invisible data" from the calculation of column completeness.

As a result, the average column completeness is 0.86 in this study, which is a high score considering not every cell needs to be filled with attributes. For example, in the address data, there is usually a field for the unit number, which is only applicable to the address entries of apartments or townhouses and could lower the score of column completeness.

However, this is not the only factor that impacts the result of column completeness as previous research has discovered that data quality can decline with increasing size of datasets (Woodall et al., 2014). Just like the building footprints dataset from the City of Waterloo receiving a perfect column completeness score of 1.00, the datasets contain less information compared to the other ones. In the building footprints dataset provided by the City of Waterloo, there were only 3 fields of "Area_M", "Shape_Area" and "Shape_Length" included in the attributes. Only the data under the field "Area_M" field is filled with the actual size of each building outline, while the cells under the other two fields "Shape_Area" and "Shape_Length" were all filled with meaningless data 0 or 0.01 to make up the space. This issue usually happens after appending multiple data with different fields into one dataset without properly purging the duplicate fields. Despite the Python script used being able to detect the blank whitespace in the dataset, it is difficult to detect the meaningless data as it is an issue related to accuracy and precision.

Similarly, the building footprints data provided by the City of Cambridge only contains one field "building_footprint_ID", which is simply an identifier to help recognize each entity, even though the identifiers are potentially tied to the property information with other datasets, the City of Cambridge did not provide any documentation to explain the data and the ID number alone does not contribute any useful information to its users. This issue can only be solved by the data owner if the

department/municipality that produces these datasets could have a better-quality assurance mechanism, preventing the inaccurate and meaningless dataset from being published online.

Schema completeness is a metric used to show the logical consistency in the dataset, which was calculated by showing the percentage of columns in each dataset that follow the data type rules provided on their metadata. The result from the schema completeness can help to check the database integrity as the data type is different than the specification noted in the file, which can be caused during the data transfer or conversion process. The score is calculated by accessing the data type of each column to see if it follows the data type provided in its metadata information. The average schema completeness is 0.90, including 10 out of 24 datasets achieving a 1.00 perfect score, which shows their data strictly follows the data type rules provided in the metadata field. The lowest-scoring dataset for schema completeness is the roads dataset from the City of Waterloo scoring at 0.64.

4.1.2 Timeliness Score

The timeliness score is comprised of two parts: the currency and the frequency score. The currency score shows how current the data is, compared to what it is supposed to be. The frequency score indicates if the data is within its frequency cycle. Up to date datasets should receive a score of 1.00 in both currency and frequency scores. The result of the timeliness score across all municipalities is concerning, with an average timeliness score of 0.54. This is the lowest cumulative score of all the different metrics evaluated in this study.

The average currency score across all municipalities is 0.56, which was a result of the individual scores distributed on both the extreme low and high ends, with four datasets scoring around or less than 0.01, and 12 out of 24 datasets scoring a perfect score of 1.00. The reason for the few datasets with low currency scores is due to lack of updates. For example, the cycling infrastructure dataset from the City of Waterloo was created in 2015, yet the last update was in 2019.

The frequency score is a binary score that indicates whether the datasets are continuously updated. A dataset that is continuously updated is one where the last update date is still in range of their update frequency. The frequency score distribution is similar to the currency's, 12 datasets received a frequency score of 1.00 as well. The only exception is the trail dataset from the City of Waterloo, which received a high currency score of 0.94 but 0 in the frequency score as it was just out of its update cycle.

Several concerns have been raised with the result of the timeliness score. First, there were 15 out of 24 datasets that lacked update frequency information listed on their metadata. With update frequencies of 15 datasets missing, the update frequencies used for them were the ones from other municipalities in the same subject, for example, if the address data from the City of Cambridge has no update frequency information associated with the dataset, the frequency will be replaced with the one provided in the addresses data from the City of Waterloo during the calculation. However, this has added some uncertainties to the evaluation as the update frequency can vary based on the local situation. On top of the lack of update frequency information, 6 out of the 9 datasets with update frequency did not have continuous updates, resulting in a 0 in their frequency score and a low currency score. The result from the timeliness score exposed several issues for open data provision in Waterloo Region, including the lack of updates since around half the datasets were severely out-of-date, including some datasets like addresses, building outlines, and cycling from the City of Waterloo have not been updated for 3-4 years. Similarly, datasets from the City of Cambridge like roads, traffic closures and trails have not been updated for 2-3 years.

4.1.3 Metadata Score

Metadata score is a metric that represents how complete a dataset's metadata information is. This is important as metadata helps users to understand the use and construction of a dataset, as well as to increase the discoverability of a dataset. The metadata score was calculated based on how many pre-determined mandatory fields proposed in the method section contain information.

After conducting this analysis, most datasets in this study were determined to have enough metadata associated with them. This resulted in a high average score of 0.86. The factors that lowered the score include missing description, and update frequency, which occurred mostly among the datasets from the City of Cambridge. The datasets from the City of Waterloo and Region of Waterloo provided no license information compared to both the City of Cambridge and the City of Kitchener provided their open government license for their data. It is worth noting that all datasets from the City of Kitchener achieved a perfect score of 1.00 because some extra information was listed on their separate metadata document in addition to the information on ArcGIS Hub.

4.1.4 Usability scores

The usability metric measures two characteristics, which are machine-readability and API accessibility. This metric was meant to show data availability and accessibility. This is because some digital formats like PDF are not machine-readable. And the API-accessibility requires the data provider to enable the server connection to the public. The results from this evaluation show flaws in this metric, due to the use of ArcGIS Hub as the open data portal platform. ArcGIS Hub automatically generates formats like CSV, Shapefile, GeoJSON and KML for manual downloads. Besides, the ArcGIS Hub also automatically creates the GeoJSON linkage from the data server for users to access them live without the need to store the data on their local desktops. Moreover, even in the case when the connection to the data server from the data provider is not established, ArcGIS allows the data provider to store the dataset online with a storage credit cost (Esri Inc., 2024). Thus, since all the datasets used in this study are hosted on the ArcGIS Hub open data portal, this makes them meet the condition for this metric, earning them a full usability score of 1.00. Given this, it can be considered to result in a less useful metric. This is due to the purpose of this metric is already achieved by the technology rather than the data providers, inflating the results of the usability score.

4.2 Municipality Score

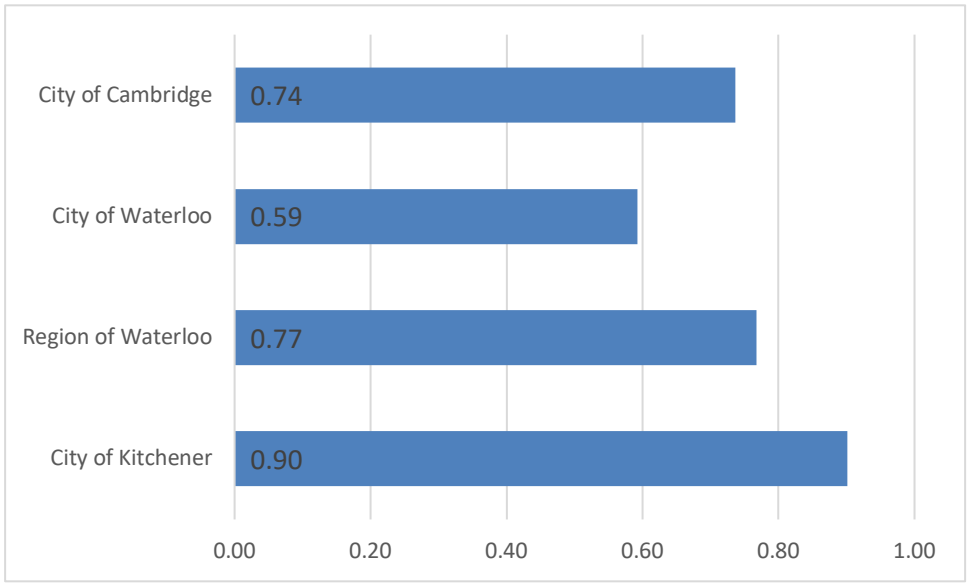

Figure 5 Average data quality score by each municipality

4.2.1 City of Kitchener

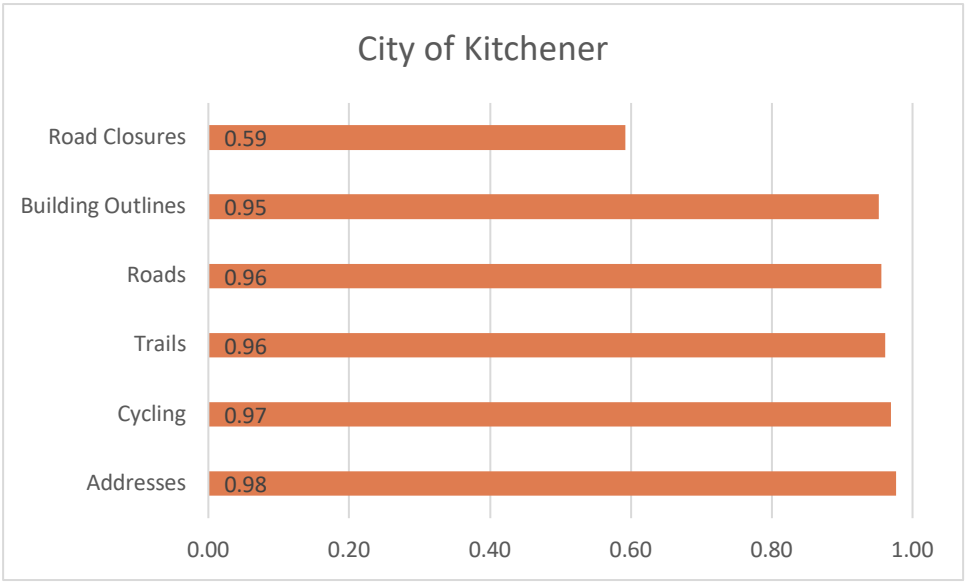

Figure 6 Data Quality Score of the City of Kitchener

Municipalities-wise, the City of Kitchener stands out as the top-quality data provider for its open data among the four municipalities. The average data quality score is at 0.90, which cannot be achieved without the support of high-quality datasets, not only because they have 5 datasets score above 0.95, but also further highlighted by its highest-scoring address dataset, achieving a score of 0.98. It is not difficult to foresee the result as the City of Kitchener provides many detailed information for their datasets such as update frequency, and metadata files. And most of their datasets are up to date as well. Besides, the City of Kitchener also provides 142 datasets, far exceeding other municipalities in the Waterloo region. While Kitchener's lowest data score is their road closures data at 0.59, which had a significant impact on lowering their average data quality score.

However, a concern has been raised regarding the credibility of the City of Kitchener's open data. Some of the information on the metadata file is self-contradictory with the information listed on their open data page. For example, the update frequency is listed hourly for the traffic closures data on their item page shown in Figure 7., while on the metadata document, it is shown as "on demand" as shown in Figure 8. This inconsistency might be caused by a script pulling data at the end of the data from the production server to the open data portal to make sure the open data stays up to date. However, without any document further explaining this, it may add confusion for the data users as all they see is the inconsistency between the information by the same provider. Thus, the timeliness score is not as indicative as it is perceived. This inconsistency would only damage the credibility of the City of Kitchener as a data producer, which could push users away. Unlike the data in the data warehouse, which usually has strict rules to address data quality such as consistency, reliability and validity, open data is built on trust and the public relies on this trust to keep themselves informed with information (Almuqrin et al., 2022). Thus, it is important to ensure that privacy, transparency, security, and reliability are there to keep the open data trustworthy as well (Meijer et al., 2014).

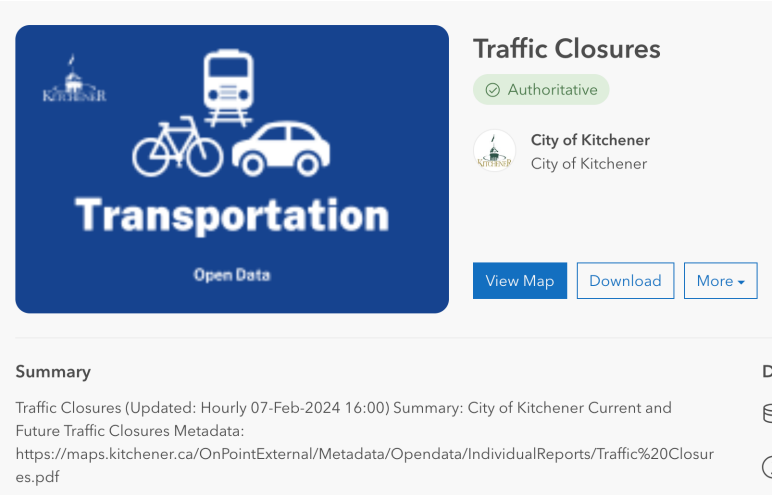

Figure 7 The City of Kitchener Traffic Closure data page with "hourly" update frequency

Source and Contraints

Source Map Label:	Traffic Closures: DTS - Transportation Planning (current
History:	Data was originally updated through iCreate but was swit
Original Source:	INS - Transportation Planning
Original Source Process:	
Maintenance:	
Current Info Source:	Stephaniein Transportation planning
Outstanding Issues:	Still need other departments to enter their road closures
Update Frequency:	On Demand
Virtual:	N

Figure 8 The City of Kitchener Traffic Closure metadata file with "on demand" update frequency

4.2.2 City of Cambridge

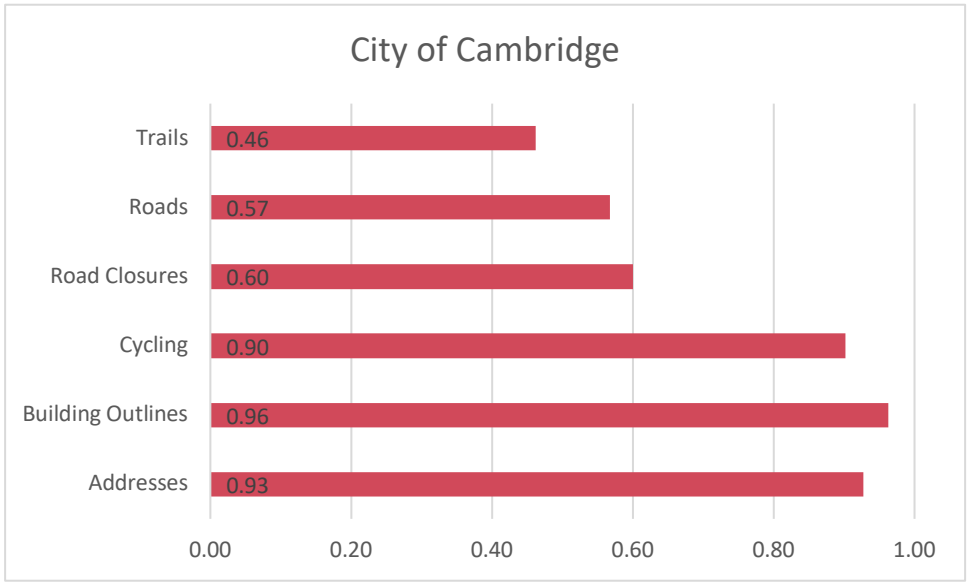

Figure 9 Data Quality Score of the City of Cambridge

The average data quality score for the City of Cambridge is 0.74, the third highest score, which was achieved with some high-quality datasets like building outlines, addresses, and cycling, scoring more than 0.90. However, they also have some low-quality datasets like road closures, roads, and trails, ranging from 0.46 to 0.60, which balanced out the high scores and made the average score lower.

The issue for datasets from the City of Cambridge is mostly related to a lack of metadata information and regular maintenance as most datasets were uploaded and last updated in 2021 in the middle of the COVID-19 pandemic. It seems like the effort of maintaining the open data program has faded away with COVID-19 as well. However, this issue was not reflected properly in the result of the evaluation as the update frequency was considered bi-annually (2-year) for the datasets like addresses or roads with “on demand” or “as needed” update frequency.

4.2.3 City of Waterloo

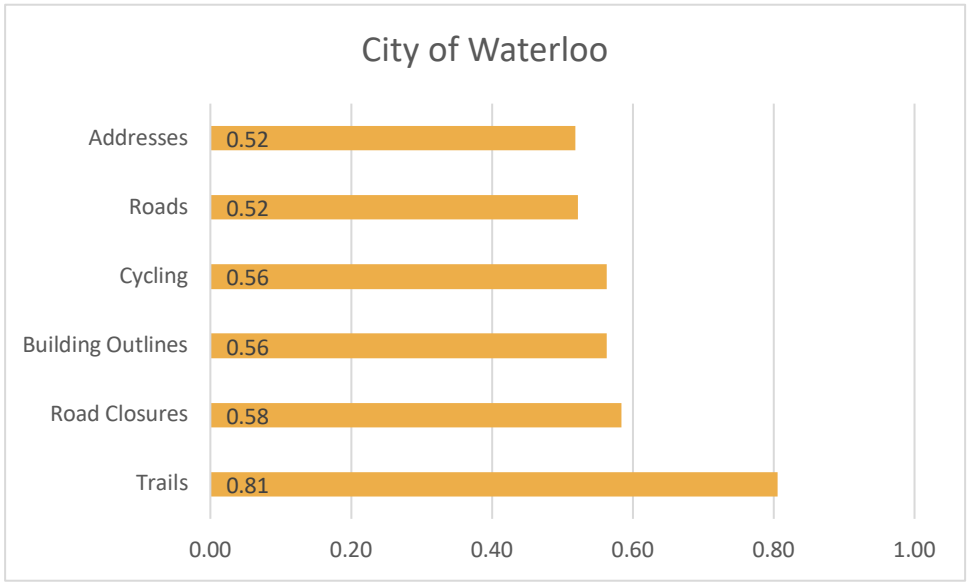

Figure 10 Data Quality Score of the City of Waterloo

The average data quality score for the City of Waterloo is 0.59, the lowest average data score provider. Unlike other municipalities, whose individual data quality score from each dataset is average distributed with some in the high-quality range and some in the low-quality range, most datasets from the City of Waterloo received a low data quality score, ranging from 0.52 to 0.58. Even the highest-scoring dataset only achieved a score of 0.81, which was significantly worse compared to the score from the City of Kitchener.

The main issue with the datasets from the City of Waterloo is that they are severely out-of-date as mentioned in the timeliness score section. Datasets like cycling infrastructure and building outlines were last updated in 2019 and 2018, while the number of both cycling infrastructure and new buildings has increased significantly in the past few years, yet there is no reflection shown in the data.

4.2.4 Region of Waterloo

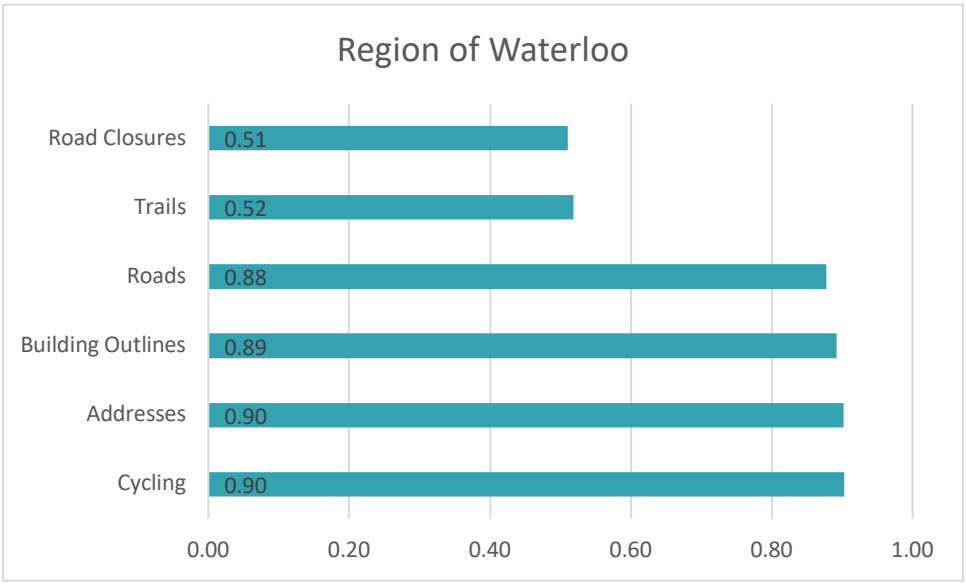

Figure 11 Data Quality Score of the Region of Waterloo

The Region of Waterloo's average quality score is 0.77, the second highest score among four municipalities. The result was achieved with datasets evenly distributed from the high to the low-quality range. Their highest scoring dataset addresses data at 0.90 while the road closures data is the lowest at 0.51. Most of the datasets from the Region of Waterloo score around 0.90, and only 2 out of 6 score around 0.51.

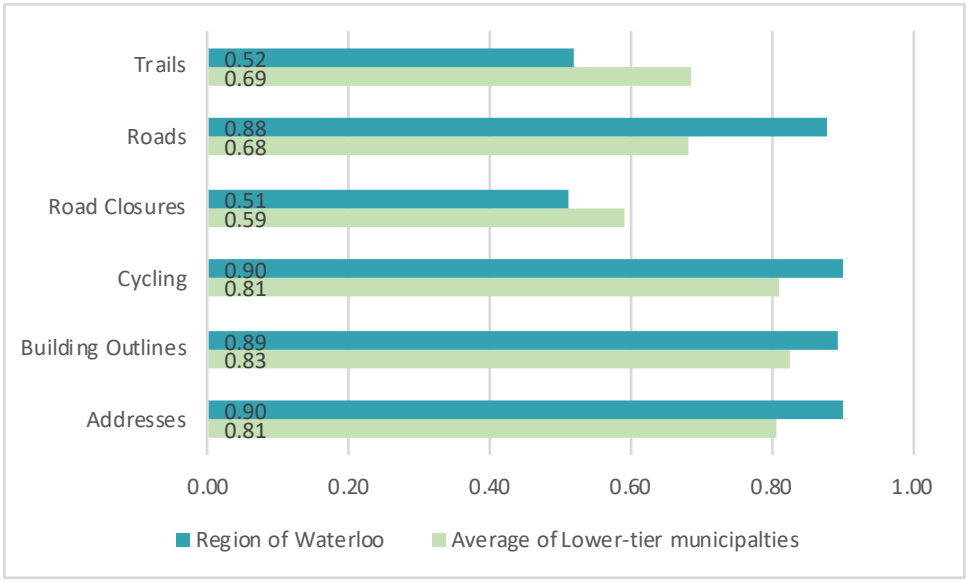

Figure 12 Data Quality Score of the Region of Waterloo vs. Average Data Quality Score from three lower-tier municipalities

Since the Region of Waterloo claims they only integrate data instead of producing most of them as an upper-tier municipality, it is worthwhile to compare their data quality score with the average score from lower-tier municipalities the City of Waterloo, the City of Kitchener, and the City of Cambridge. As shown in Figure 12, the data quality scores are similar for most datasets with the score from the three lower-tier municipalities slightly worse than the score from the Region of Waterloo. It is due to the data integration process having positive impacts on improving the data quality by removing unnecessary fields and incomplete data. It is also important to acknowledge that there might be a delay for the Region of Waterloo to update their datasets like Road Closures.

4.3 Conclusion

Overall, the results from this data quality assessment have Shown that the upper-tier government can have a positive contribution to the cross-jurisdictional data as it can oversee the data from a larger scale and help the integration process by defining some common formats. Despite some notable issues like lack of updates and availability of datasets, the open data program in Waterloo region is one of the

top open data programs compared to other regional settings such as Halton Region (does not have an open data portal), Wellington (does not have an open data portal), and Niagara Region (98 datasets). Most upper-tier governments in Ontario do not have a complete open data program. However, the open data in Waterloo is still behind on most practices ranging from policies to maintenance when compared to the other similar regional settings like in York region, whose open data program comes with detailed metadata including update frequency and the indication of continuous or discrete update, regular update and maintenance of the datasets, and most importantly their regularly updated open data policy[3], ensuring the data quality of their open data program is kept high-level and focused to the current situation.

[3] The newest open data policy from York region (https://insights-york.opendata.arcgis.com/documents/york-region-open-data-policy/explore)

Chapter 5
Discussion

Data quality is not a new topic for open data, which requires more attention and effort to ensure its success. Despite a data evaluation tool having been developed by this study for the Waterloo region, there is still a lot of work that needs to be done to improve the data quality for the open data programs with the two-tier government system. Issues like lack of data provision standards, accessibility and availability, and communication with users still exist for the government's open data programs (Johnson et al., 2022). Although open data has been a worldwide movement, providing reliable and high-quality open data remains a challenge for most open programs (Baculi et al., 2016). As noted by some studies, data production is often a by-product of technology or government services, which usually have a bigger scope of development and responsibilities other than providing data itself (Arribas-Bel et al., 2021; Charalabidis et al., 2018). The findings from the results of the open data evaluation in this study identified that there are still gaps and challenges existing in the open government data programs behind the issues reflecting on the quality metrics.

5.1 The challenges of providing open data

The evaluation results have reflected many deep-level issues that cannot be solved right away, connecting the result with the literature, a few challenges stand out the most with open data in the Region of Waterloo.

5.1.1 Decentralization of Government

One of the main challenges that most open government data programs face is the decentralization of their data governance. From the data quality evaluation in this study, there are significant differences in data quality scores between datasets in different subjects. According to the open data policy by the City of

Waterloo[4], each department should be responsible for providing its open data to the public. However, with each department and government agency working individually, it is difficult to maintain a continuous effort to the quality standard due to the lack of centralized coordination. Yet the decentralized approach was not only used at a departmental level but a municipality or even a provincial level, as indicated by Johnson et al. (2022), accessing data cross-jurisdictionally can be difficult due to the different methods and standards used by different levels of governments. Similarly, as noted by Roche et al. (2020), the decentralization of open data has imposed a further requirement for data management in both accessibility and security.

5.1.2 Lack of Data Provision Standards

The lack of data provision standards also contributes to the low-quality open data, especially the lack of data quality assurance process. Despite most open government data programs having existing data standards to improve their quality metrics like completeness, the data can remain inaccurate and outdated due to lack of quality assurance or regular updates (Johnson et al., 2020). Many quality-related issues like the duplicated field showing meaningless data mentioned in Chapter 4 can be easily detected and fixed if there is a quality assurance process implemented. The quality assurance process can identify and improve data quality issues before they impact data users. By reviewing the dataset's metadata, completeness, consistency, timeliness, and even some qualitative aspects like accuracy, the quality assurance process can ensure the quality of the dataset is improved significantly. Besides, a quality assurance process can be used for identifying the areas of improvement needed in data collection, integration, and transfer process. However, existing research done by Nikiforova and McBride (2021) suggested the lack of provision standards does not only affect data quality but also the usability of the data portal as well. According to

[4] City of Waterloo's Open Data Policy (2013)
https://www.waterloo.ca/en/government/resources/Documents/Cityadministration/Policies/Corporate-Policy/Administrative/Open-data-policy.pdf

Nikiforova & McBride (2021), the poor usability of the data portal is a result of the poor data quality, which results in the low desire to make a usable data portal for providing the contents that hurt the implementation of the open data programs. The participation of the public is an important process, as it allows citizens and industries to have different usage of the data than government and agencies, adding direct and indirect values to open data and allowing the further development of open data applications (Matheus et al., 2020).

Even though there are many benefits to having an open data standard, some challenges will also come along with it. The first one is the direct cost of building an open data standard (Johnson et al. 2017), a complex data standard model can work, but it also increases the amount of both money and human resources that need to be devoted. The unnecessary over-complication, especially applying federal settings to local ones, made it hard to define the exact types of each detail and comes with. Consequently, the maintenance and update process will be more complicated due to different update frequencies, coverage, scales, and other additional costs. In addition, the uneven distribution of the open data made this become a limitation for creating an open data standard. Not every level of government has enough resources (e.g. direct cost of servers, and platforms) to publish their open data, and not all data can be shared publicly due to security and privacy concerns. Some upper-level governments and municipalities might have more resources or funding to host their web-based service for open data such as the City of Toronto and, the City of Edmonton, who were able to build their open data website to host their data. Moreover, it is also difficult to determine the same way for datasets in different jurisdictions, as some data might not be available in certain areas, which simply makes an open standard not applicable (Roy, 2014).

5.1.3 Outdated Policies

Most open data programs are run by governments and agencies, which are heavily dependent on the corresponding policies to support daily operations and future developments (Rivera Perez, Emilsson

& Ubaldi, OECD, 2020). It is important to note the different results of the data quality score by each municipality are also tied to the different open data policies each municipality follows (Zuiderwijk & Janssen, 2014). The open data policy by the City of Waterloo notes each department should be responsible for providing their own data to the public and they should also create their own data stewardship model within each department to be accountable for managing and maintaining the datasets (City of Waterloo, 2013). On the contrary, the City of Kitchener uses a set of specific requirements to ensure the quality of its open data (City of Kitchener, 2007). Unlike the other cities, the City of Cambridge provides no policy or plans about their open data. The Region of Waterloo states each department and service within the region should be responsible for maintaining its own datasets as the regional government does not produce data (Region of Waterloo, 2022). Based on these policies, it is noticeable that most of the policies were created over a decade ago, despite there might be more updated versions in recent years, none of them are available to the public. While the development of open data has been forcefully evolving over time especially in the last few years due to the pandemic as more workplaces have started to adapt to remote work, requiring more digital data than ever (KPMG, 2020). Thus, open data policies also need to be updated to reflect the current open data state. Some successful open data programs like the NYC Open Data would ensure their policies and standards are up-to-date and accessible to the public (New York City, 2023).

5.1.4 Limited Data Availability

There is another concern about data availability for the open data program in Waterloo region. Even though the regional government emphasizes they do not produce data, there are still many datasets the regional government can integrate from the existing local municipal datasets, such as bridges, places of worship or railways, which are all available by the three local cities. However, despite the fact most of these datasets share common features, due to the different classifications and details used by different municipalities, it can also be difficult for users to integrate their own regional-scale data without prior

knowledge of data integration and classification (Johnson et al., 2020). As pointed out by Ziegler & Dittrich (2007), data integration requires adaptation and reconciliation of their different functionalities, and there is always more than one single issue like completeness, consistency, and performance need to be considered when performing a data integration as the systems are not designed to fit each other. The lack of variety of regional-level data can only push the users away further and discourage public engagement. Previous research has found lack of specific datasets users are interested in can discourage their potential use of open data in the future (Beno et al., 2017; Johnson et al., 2020). Without public participation, the open data program would not receive constructive feedback and lose its purpose of being open and transparent.

5.1.5 Measuring Success

Some existing research has indicated that it is also crucial to evaluate the impact of open data programs, as most research and studies for spatial data quality remain in academia (Devillers et al., 2010). As Devillers et al (2010) pointed out, a lot of findings from the research on spatial data quality were not adapted and integrated to be part of the GIS software and applications, leaving the users unaware and unable to understand the concept of the spatial data quality. Besides, fitness for use is another factor impacting the success of open data programs as the definition of high data quality is a subjective opinion dependent on the expectations of the users are met (Sackl et al., 2017). Yet, there has not been an interpretation of the fitness of open government data due to the expectation

5.2 Limitation

Although the evaluation tool has helped to identify the gaps and the challenges facing the open data programs in the Region of Waterloo, it is important to understand there are still some limitations in this study. The sample size of the dataset is one of the major limitations in this study due to the limited availability of datasets in different subjects, which limits the scale of the evaluation and can make the final results inadequate. Additionally, for the timeliness, the study took the assumptions on using update

frequency from other datasets and also the 2-year period for the length of "on demand" update frequency, which might create bias and impact the final timeliness scores. Moreover, there are also some issues related to the accuracy of the data, however, it is difficult to verify without any other available data sources to collect similar information like governments due to the sensitivity of some data subjects. Besides the assumption and a few limitations caused by the data itself, there are also some limitations this study was not able to verify.

Chapter 6

Conclusion

6.1 Conclusion

The research goals of this study are 1) to conduct literature reviews of open data quality and the current evaluation system to measure open data quality in a quantitative method. 2) develop a comprehensive evaluation tool for assessing the data quality in the two-tier municipal contexts in the Region of Waterloo. And 3) identify the gaps and challenges that remain for the open data programs. For the first goal, the literature was able to identify the issue with open data quality and some existing models used for the evaluation systems for measuring open data quality. The data quality evaluation tool was developed, and the quality of open data between the two-tier municipalities was compared by using the data quality score from the evaluation. The remaining gaps and challenges have been identified by comparing the results to the literature. Despite the study's objectives being achieved, there are still many underlying issues existing with the open data programs in the Region of Waterloo. The results from the evaluation can be used for identifying the gaps in the open data programs, highlighting areas where improvement can be made to enhance the cross-jurisdictional open data programs in the Waterloo region. Besides, the findings from this study can be used as a reference to improve the insufficient consideration of each data quality metric, contributing to a more complete and objective model with a comprehensive result for future studies. By addressing the identified gaps and implementing recommendations from this study, the open data programs in the Waterloo region can make greater impacts in advocating government transparency and providing more valuable information for decision-makers, researchers, and the public (Matheus et al., 2023).

6.2 Recommendation

According to the results of the data quality evaluation, there are several recommendations that can be made for the open data programs in the Waterloo region.

6.2.1 Open metadata and data standard

During the data collection process, it is noted that many datasets have missing metadata information. Without a clear metadata standard, each municipality provides their metadata info differently, for example, as previously mentioned, the City of Kitchener noted some of their dataset's update frequency as continuous. A data standard is a set of rules that determines how to describe data, process data, and store data in a consistent way (Statistics Canada, 2021). It is important to keep the metadata and the actual datasets in a uniform format. As it can be seen throughout the data selection process, the upper-tier municipality Region of Waterloo and the three lower-tier municipalities the City of Waterloo, the City of Kitchener and the City of Cambridge do not share the same/similar data standards. The different data standards have caused many difficulties in integrating the data on a larger scale. Although the reason for most municipalities to localize their own metadata and open data standards is to adapt to their unique situation, the three municipalities are connected under the same upper-level municipality the Region of Waterloo, making their situation cross-jurisdictionally. As an upper-tier municipality, the Region of Waterloo should have more resources and authorities to initiate this recommendation. By sharing an open metadata and data standard with the number of available datasets from the three lower-tier municipalities, it can save time and cost for the regional government to make the data integration process much easier, indirectly increasing the availability of datasets. The lower-tier municipalities would also be benefited from creating higher-quality datasets followed by the standard.

6.2.2 Improving Terminology of Naming and Keywords in the Data Portal

The Region of Waterloo states the reason for having all four municipalities merged into one open data portal is to make the datasets more accessible to its users. As it is also understandable that with the limited data availability by some lower-tier municipalities like the City of Waterloo, and the City of Cambridge, sharing one data portal can save a lot of software subscription costs. However, with the large number of datasets in the same portal, different terminology used by the different municipalities, and missing or incorrect tags and keywords, this decision would make the datasets less accessible and add more complexity for the users to find the datasets they want (Lopes et al., 2015). According to the study done by Miller (2018), the keyword search can return a large number of unwanted results that require manual examination, which can interfere with the discoverability of the dataset and make some vital datasets overlooked due to missing or incorrect tags or metadata descriptions. For example, when searching for addresses datasets using the keyword "addresses", the Region of Waterloo open data portal returned a total of 65 datasets as shown in Table 6. Despite the address datasets of each local municipality can be found in the top 5, the address data from the Region of Waterloo is listed as the 64th dataset on the last page of the search result. As ArcGIS Hub's searching algorithm examines not just the title of the dataset but also the metadata information, then returns the results with a summary of every dataset that includes the term (Esri Inc., 2024).

Order	Name of dataset	Publisher	Keywords
1	Addresses	City of Kitchener	Base Data, Basemap, City of Kitchener, Kitchener, Open Data
2	Address_Proximity_Directory	City of Kitchener	Planning, Base Data, Basemap, Community, Boundaries, City of Kitchener, Kitchener, Open Data
3	Plow_Priority_by_Address	City of Kitchener	Services, Infrastructure, Utilities,

			City of Kitchener, Kitchener, Open Data
4	Addresses	City of Waterloo	Base Data, city of waterloo, open data, waterloo
5	Addresses	City of Cambridge	Base Data, Basemap, Data Catalogue
6	Property_Ownership_Public	City of Kitchener	Base Data, Basemap, City of Kitchener, Kitchener, Open Data
7	Roads	City of Kitchener	Transportation, Infrastructure, Traffic, Basemap, Base Data, City of Kitchener, Kitchener, Open Data
8	City of Waterloo Building Permits	City of Waterloo	records, city of waterloo, building, open data, permit
9	Place of Worship	City of Waterloo	Points of Interest, city of waterloo, open data
10	Business_Dictionary	City of Kitchener	Community, Services, Landmarks, Points of Interest, City of Kitchener, Kitchener, Open Data
…	…	…	…
64	Addresses	Region of Waterloo	Information, Base Data, Address, Region of Waterloo

Table 6 Searching result by using keyword "addresses"

This issue is not only caused by ArcGIS Hub's searching algorithm but also by the different terminology used as the title of datasets. Since there is no collective agreement or guideline for categorizing or naming the datasets, municipalities can just make up their own terminology for the datasets. As an example, the Region of Waterloo named its cycling path data as "Cycling", and both the City of Waterloo and the City of Kitchener named their data as "Cycling Infrastructure", while the City of

Cambridge named its data as “Bikeway Network”. But using cycle or cycling as the search term, the datasets from the Region of Waterloo, the City of Waterloo and the City of Kitchener can be found. Even if the search term changes to bike or bike network, only the cycling infrastructure dataset from the City of Waterloo would be returned in the result. However, the Bikeway Network dataset from the City of Cambridge would not be promoted by using any related terms.

Thus, the datasets describing the same datasets should use the same terminology with different acronyms as suffixes such as ROW, COW, COK, and COC to allow users to distinguish the datasets from different municipalities. In addition to that, it is also important to improve the keywords tagging on the ArcGIS Hub, as shown in Table 6., the current keywords used remain generic and unclear, which does not contribute to the discoverability of the datasets. Furthermore, each municipal government should also create an individual catalogue to list what datasets are available.

6.2.3 Creation of an open data team

Even though open data movement has been adapted for more than one decade in the Waterloo region, according to the data evaluation score, the current open data status is still unclear. The datasets have various issues not only related to accuracy, and update but also missing the quality assurance process.

It is suggested that each municipality to create their own open data stewardship to manage its open data programs (Peng et al., 2016). Since most municipal governments use the decentralized model, which makes each individual department in the government the data owner, giving them a lot of freedom to manage their own data in the open data program (Cerrillo-Mártinez & Casadesús-de-Mingo, 2021). As a result, there is not a lot of supervision for each individual dataset, indirectly lowering the quality of the open data program. Therefore, this study is proposing a new open data team for monitoring and ensuring that the data each department or municipality produces matches the standards based on ISO 19157. The

addition of the open data team can create a process that requires each dataset to be reviewed through multiple processes before publishing, adding accountability and creditability to the data quality.

6.2.4 Update of Open Data Policies

As the focus on data quality keeps updating, and the demands for open data keep changing, the policies for open data should be updated regularly as well. For example, in the early period of open data, the most used evaluation model is Berners-Lee's 5-star scheme (Berner-Lee, 2006), focusing on the accessibility and openness of open data. However, with the advancement of open data movement, researchers have started to evaluate individual characteristics of data such as completeness, timeliness etc.

The federal government's action plan started as the Open Government Initiative (OGI) in 2011 and is updated regularly in a bi-annual period identifying the existing gaps and focusing on the current events to improve the open data program (Government of Canada, 2019). Similarly, the open data policies in the Waterloo region are all created around the early 2010s, shortly after the OGI. Yet unlike the federal government, the current open data policies have no regular update, municipal governments such as the City of Waterloo are still using the open data policies created back in 2010. As noted by the Government of Canada (Natural Resources Canada, 2024), despite many data policies not being updated regularly in response to the change in the status of geospatial data, the data policies are still critical for geospatial data implementation and removing obstacles for the users. Therefore, it is important for the Region of Waterloo, City of Waterloo, City of Kitchener, and City of Cambridge to update their open data policies to reflect the current open data status. Moreover, the new policies should focus on identifying the gap for the already existing programs and maintain a regular update to improve the result of quality metrics as the City of Toronto does (City of Toronto, 2023). Updating the open data policy can help municipalities ensure compliance with the data quality standards, creating a more transparent government, and adding public trust and engagement. With the provision of high-quality open data, the government would be able to demonstrate more transparency to allow more citizens to engage in governance. Citizen participation

would create a positive loop by providing constructive feedback. For the private sector, the provision of high-quality open data would allow them to create more financial value from the data, resulting in them being willing to offer more support to the open data programs (Johnson et al., 2020)

References

Amankwah-Amoah, Khan, Wood, & Knight G. (2021). COVID-19 and digitalization: The great acceleration. *J Bus Res* (136), 602-611.

Arribas-Bel, D., Green, M., Rowe, F., & Singleton, A. (2021). Open data products-A framework for creating valuable analysis ready data. Journal of Geographical Systems, 23(4), 497-514. https://10.1007/s10109-021-00363-5

Atz, U. (2014) The Tau of Data: A New Metric to Assess the Timeliness of Data in Catalogues. In Proceedings of the International Conference for E-Democracy and Open Government (CeDEM2014), Krems, Austria.

Almuqrin, A., Mutambik, I., Alomran, A., Gauthier, J., & Abusharhah, M. (2022). Factors Influencing Public Trust in Open Government Data. *Sustainability, 14*(15)https://10.3390/su14159765

Association of Municipalities Ontario. (2021, *Municipal 101* . https://www.amo.on.ca/about-us/municipal-101

Baculi, E., Fast, V., & Rinner, C. (2017). The Geospatial Contents of Municipal and Regional Open Data Catalogs in Canada. *URISA Journal*, *28*(1), 39+. https://link.gale.com/apps/doc/A559830504/AONE?u=wate34930&sid=googleScholar&xid=1e08dd48

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. *ACM Comput.Surv., 41*(3)https://10.1145/1541880.1541883

Belhiah, M., & Bounabat, B. (2017). *A User-Centered Model for Assessing and Improving Open Government Data Quality.*

Beno, K. Figl, J. Umbrich, & A. Polleres. (2017). Open Data Hopes and Fears: Determining the Barriers of Open Data. Paper presented at the *2017 Conference for E-Democracy and Open Government (CeDEM),* 69-81. https://10.1109/CeDEM.2017.22

Berners-Lee. (2006). Linked Data. https://www.w3.org/DesignIssues/LinkedData.html

Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly, 27*(3), 264-271. https://10.1016/j.giq.2010.03.001

Bonaguro, J. (2015) How to Measure Open Data. DataSF. City and County of San Francisco. https://datasf.org/blog/how-to-measure-open-data/#:~:text=Number%20of%20datasets%20is%20a,but%20data%20usability%20goes%20down.

Candela, G., Escobar, P., Carrasco, R. C., & Marco-Such, M. (2022). Evaluating the quality of linked open data in digital libraries. Journal of Information Science, 48(1), 21–43. https://doi.org/10.1177/0165551520930951

Cerrillo-Mártinez, A., & Casadesús-de-Mingo, A. (2021). Data governance for public transparency. Profesional De La información Information Professional, 30(4). https://doi.org/10.3145/epi.2021.jul.02

Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Lampoltshammer, T., & Ferro, E. (2018). Open Data Evaluation Models: Theory and Practice. In Y. Charalabidis, A. Zuiderwijk, C. Alexopoulos, M. Janssen, T. Lampoltshammer & E. Ferro (Eds.), *The World of Open Data: Concepts, Methods, Tools and Experiences* (pp. 137-172). Springer International Publishing. https://10.1007/978-3-319-90850-2_8

City of Waterloo. (2013). Open Data Policy. https://www.waterloo.ca/en/government/resources/Documents/Cityadministration/Policies/Corporate-Policy/Administrative/Open-data-policy.pdf

City of Toronto. (2023). Towards an updated Data Quality Score in Open Data. https://open.toronto.ca/towards-an-updated-data-quality-score-in-open-data/

Debattista, J., Auer, S. O., & Lange, C. (2016). Luzzu—A Methodology and Framework for Linked Data Quality Assessment. *J.Data and Information Quality, 8*(1)https://10.1145/2992786

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., & Shi, W. (2010). Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. *Transactions in GIS, 14*(4), 387-400. https://10.1111/j.1467-9671.2010.01212.x

Elouataoui, W., El Alaoui, I., El Mendili, S., & Gahi, Y. (2022). An Advanced Big Data Quality Framework Based on Weighted Metrics. *Big Data and Cognitive Computing, 6*(4)https://10.3390/bdcc6040153

Esri Inc. (2020), *Metadata*. https://developers.arcgis.com/rest/services-reference/enterprise/metadata.htm

Esri Inc. (2024), *ArcGIS Hub.* https://www.esri.com/en-us/arcgis/products/arcgis-hub/overview

Esri Inc. (2024), *Feature Layer*. https://developers.arcgis.com/rest/services-reference/online/feature-layer.htm

Esri Inc. (2024). *GeoJSON*. https://doc.arcgis.com/en/arcgis-online/reference/geojson.htm

Esri Inc. (2024). *Understand Credits*. https://doc.arcgis.com/en/arcgis-online/administer/credits.htm#

Esri Inc (2024). What's New. https://www.esri.com/arcgis-blog/products/arcgis-hub/announcements/whats-new-on-arcgis-hub-search/

European Data Portal. (2014). *Metadata Quality Assurance*. https://data.europa.eu/mqa/methodology?locale=en

European Union (2021). *Data Quality Guidelines* https://op.europa.eu/webpub/op/data-quality-guidelines/en/

Google Inc. (2024, *Indexing.* https://developers.google.com/search/docs/fundamentals/how-search-works

Government of Canada. (2017). Government. https://www.canada.ca/en/immigration-refugees-citizenship/services/new-immigrants/learn-about-canada/governement.html

Government of Canada, & Treasury Board of Canada Secretariat. (2020). Open Data Metadata Mapping. https://open.canada.ca/data/en/dataset/18bb430e-ffc8-43e6-a20e-cc4e15c3fd71

Government of Canada. (2019). Canada's Action Plan Open Government. https://open.canada.ca/en/canadas-action-plan-open-government

Government of United Kingdom (2022). Data Quality Assessment Pitfalls. https://www.gov.uk/government/news/data-quality-assessment-pitfalls

Härting, R., & Lewoniewski, W. (2020). Main Influencing Factors of Quality Determination of Collaborative Open Data Pages. *Information, 11*(6)https://10.3390/info11060283

Hernandez, C. (2020). Towards a Data Quality Score in open data . *Open Data Toronto,* https://medium.com/open-data-toronto/towards-a-data-quality-score-in-open-data-part-1-525e59f729e9

IBM. (2021). Data quality score. https://www.ibm.com/docs/en/iis/11.5?topic=results-data-quality-score

International Organization for Standardization. (2014). ISO 19115-1:2014 Geographic information: Metadata. https://www.iso.org/standard/53798.html#:~:text=ISO%2019115-1%3A2014%20defines%3A,digital%20data%20and%20services)%3B

International Organization for Standardization. (2022). ISO 8000-1:2022(en) Data quality. https://www.iso.org/obp/ui/en/#!iso:std:81745:en

International Organization for Standardization. (2023). ISO 19157-1:2023(en) Geographic information — Data quality. https://www.iso.org/obp/ui/en/#iso:std:iso:19157:-1:ed-1:v1:en

Johnson, P. A., Sieber, R., Scassa, T., Stephens, M., & Robinson, P. (2017a). The cost(s) of geospatial open data. *Transactions in GIS, 21*(3), 434-445. https://doi.org/10.1111/tgis.12283

Johnson, P. A., & Varga, C. Challenges to the Access of Government Open Data by Private Sector Companies. *THE FUTURE OF OPEN DATA*, 103.

Kaiser, M., Klier, M., & Heinrich, B. (2007), "How to Measure Data Quality? - A Metric-Based Approach" (2007). *ICIS 2007 Proceedings*. Paper 108.

Kassen, M. (2018), "Adopting and managing open data: Stakeholder perspectives, challenges and policy recommendations", *Aslib Journal of Information Management*, Vol. 70 No. 5, pp. 518-537. https://doi.org/10.1108/AJIM-11-2017-0250

Kim, & Hausenblas. (2015). 5 star open data. https://5stardata.info/en/

Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018). Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly, 35*(1), 13-29. https://10.1016/j.giq.2017.11.003

KPMG. (2020). *Digital acceleration.* (). https://kpmg.com/us/en/home/insights/2020/09/digital-acceleration.html

Legislative Assembly of Ontario. (2023, *Levels of Government* . https://www.ola.org/en/visit-learn/teach-learn-play/levels-government.

Lopes, N., Stephenson, M., Lopez, V., Tommasi, P., & Aonghusa, P. M. (2015). On-Demand Integration and Linking of Open Data Information. Paper presented at the *Metadata and Semantics,* 312-323.

Matheus, R., Janssen, M., & Maheshwari, D. (2020). Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities. *Government Information Quarterly, 37*(3), 101284. https://10.1016/j.giq.2018.01.006

Matheus, R., Faber, R., Ismagilova, E., & Janssen, M. (2023). Digital transparency and the usefulness for open government. *International Journal of Information Management, 73*, 102690. https://10.1016/j.ijinfomgt.2023.102690

Mayer-Schönberger, V., & Zappia, Z. (2011, October). Participation and power: Intermediaries of open data. In *1st Berlin Symposium on Internet and Society, Berlin, Germany. http://berlinsymposium. org/sites/berlinsymposium. org/files/participation_and_power. pdf*.

Meijer, R., Conradie, P., & Choenni, S. (2014). Reconciling Contradictions of Open Data Regarding Transparency, Privacy, Security and Trust. *Journal of Theoretical and Applied Electronic Commerce Research, 9*(3), 32-44. https://10.4067/S0718-18762014000300004

Mergel, I., Kleibrink, A., & Sörvik, J. (2018). Open Data Outcomes: U.S. Cities between Product and Process Innovation. *Government Information Quarterly, 35* https://10.1016/j.giq.2018.09.004

Miller, R. J. (2018). Open Data Integration. *Proc.VLDB Endow., 11*(12), 2130–2139. https://10.14778/3229863.3240491

Natural Resources Canada (2016). Geospatial Data Quality Guide. https://ostr-backend-prod.azurewebsites.net/server/api/core/bitstreams/f119c026-410d-4850-99ad-4e9405e6d6ad/content

Natural Resources Canada (2020). Canadian Geospatial Data Infrastructure CookBook. https://publications.gc.ca/collections/collection_2021/rncan-nrcan/M124-10-1-2020-eng.pdf

Natural Resources Canada (2024). Geospatial Standards and Operational Policies https://natural-resources.canada.ca/earth-sciences/geomatics/canadas-spatial-data-infrastructure/8902

Neumaier, S., Umbrich, J. u., & Polleres, A. (2016). Automated Quality Assessment of Metadata across Open Data Portals. *J.Data and Information Quality, 8*(1)https://10.1145/2964909

New York City (2023). *NYC Open Data Technical Standards Manual.* Office of Technology & Innovation – New York City. https://opendata.cityofnewyork.us/wp-content/uploads/NYC_OpenData_TechnicalStandardsManual.pdf

Nikiforova, A. (2018). Open Data Quality Evaluation. Lupeikiene A., Matulevičius R., Vasilecas O. (eds.): Baltic DB&IS 2018 Joint Proceedings of the Conference Forum and Doctoral Consortium

Nikiforova. (2020). Timeliness of Open Data in Open Government Data Portals Through Pandemic-related Data: a long data way from the publisher to the user. Paper presented at the *2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA),* 131-138. https://10.1109/MCNA50957.2020.9264298

Nikiforova, A., & McBride, K. (2021). Open government data portal usability: A user-centred usability analysis of 41 open government data portals. Telematics and Informatics, 58, 101539. https://10.1016/j.tele.2020.101539

OECD (2019), Making Decentralisation Work: A Handbook for Policy-Makers, OECD Multi-level Governance Studies, OECD Publishing, Paris, https://doi.org/10.1787/g2g9faa7-en.

Open Data Charter (2015). ODC Principles. https://opendatacharter.org/wp-content/uploads/2023/12/opendatacharter-charter_F.pdf

Open Knowledge Foundation. (2021). What is Open Data? Open Data Handbook. Retrieved from https://opendatahandbook.org/guide/en/what-is-open-data/

Osagie, E., Waqar, M., Adebayo, S., Stasiewicz, A., Porwol, L., & Ojo, A. (2017). Usability Evaluation *of an Open Data Platform* https://10.1145/3085228.3085315

Peng, Ge., Ritchey, N., Casey K., Kearns, E., Privette, J., Saunders, D., Jones, P., Maycock, T., Ansari, S. (2016). Scientific Stewardship in the Open Data and Big Data Era - Roles and Responsibilities of Stewards and Other Major Product Stakeholders. 22https://doi.org/10.1045/may2016-peng

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2003). Data quality assessment. *Commun.ACM, 45*(4), 211–218. https://10.1145/505248.506010

Popkin, G. 2019, May 1. Data sharing and how it can benefit your scientific career. *Nature, 569*: 445–447. DOI: https://doi.org/10.1038/d41586-019-01506-x

Quarati, A. (2023). Open Government Data: Usage trends and metadata quality. *Journal of Information Science*, *49*(4), 887-910. https://doi.org/10.1177/01655515211027775

Rahm, E. (2016). The Case for Holistic Data Integration. In: Pokorný, J., Ivanović, M., Thalheim, B., Šaloun, P. (eds) Advances in Databases and Information Systems. *ADBIS 2016. Lecture Notes in Computer Science, vol 9809*. Springer, Cham. https://doi.org/10.1007/978-3-319-44039-2_2

Rayes, J., & Mahmood, S. (2020). *New Open Data Products Support Data Access and Literacy*. City of Toronto Open Data Portal. Retrieved from https://open.toronto.ca/products_improve_data_literacy_access/

Rivera Perez, Emilsson, Ubaldi, & OECD. (2020). *OECD Open, Useful and Re-usable data (OURdata) Index: 2019*. https://web-archive.oecd.org/2020-03-10/547558-ourdata-index-policy-paper-2020.pdf

Region of Waterloo. (2024). Open Data Provided by our Partners. https://rowopendata-rmw.opendata.arcgis.com

Roy, J. (2014). *Open Data and Open Governance in Canada: A Critical Examination of New Opportunities and Old Tensions* https://10.3390/fi6030414

Sadiq, S.W., & Indulska, M. (2017). Open data: Quality over quantity. *Int. J. Inf. Manag., 37*, 150-154

Sackl, A., Schatz, R. & Raake, A. More than I ever wanted or just good enough? User expectations and subjective quality perception in the context of networked multimedia services. *Qual User Exp 2, 3* (2017). https://doi.org/10.1007/s41233-016-0004-z

Sebastian-Coleman, L. (2013). Chapter 4 - Data Quality and Measurement. In L. Sebastian-Coleman (Ed.), Measuring Data Quality for Ongoing Improvement (pp. 39-53). Morgan Kaufmann. https://10.1016/B978-0-12-397033-6.00004-3

Slibar, B., Oreski, D., & Klicek, B. (2018). Aspects of open data and illustrative quality metrics: literature review. *Economic and Social Development: Book of Proceedings*, 90-99.

Statistics Canada (2021). 2021 Census Data Quality Guideline. https://www12.statcan.gc.ca/census-recensement/2021/ref/98-26-0006/982600062021001-eng.cfm

Statistics Canada (2021). Data Governance. https://www.statcan.gc.ca/en/about/datastrategy

Ubaldi, B. (2013), "Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives", *OECD Working Papers on Public Governance*, No. 22, OECD Publishing, Paris, https://doi.org/10.1787/5k46bj4f03s7-en.

U.S. Environmental Protection Agency. (2021). Climate Adaptation Action Plan. https://www.sustainability.gov/pdfs/epa-2021-cap.pdf

Vaziri, R., Mohsenzadeh, M., & Habibi, J. (2019). Measuring data quality with weighted metrics. *Total Quality Management & Business Excellence*, *30*(5–6), 708–720. https://doi.org/10.1080/14783363.2017.1332954

Vetro, A., Canova, L., Torchiano, M., Minotas, C., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly, 33* https://10.1016/j.giq.2016.02.001

Viscusi, G., Spahiu, B., Maurino, A., & Batini, C. (2014). Compliance with open government data policies: An empirical assessment of Italian local public administrations. *Information polity*, *19*(3-4), 263-275.

Woodall, P., Oberhofer, M., & Borek, A. (2014). A classification of data quality assessment and improvement methods. *International Journal of Information Quality 16*, *3*(4), 298-321.

Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly, 31*(1), 17-29. https://10.1016/j.giq.2013.04.003

Zuiderwijk, A., Janssen, M., Poulis, K., & van de Kaa, G. (2015). Open data for competitive advantage: insights from open data use by companies. Paper presented at the *Proceedings of the 16th Annual International Conference on Digital Government Research,* Phoenix, Arizona. 79–88. https://10.1145/2757401.2757411 https://doi.org/10.1145/2757401.2757411

Ziegler, P., & Dittrich, K. R. (2007). Data integration—problems, approaches, and perspectives. In *Conceptual modelling in information systems engineering* (pp. 39-58). Berlin, Heidelberg: Springer Berlin Heidelberg.

Appendices

Appendix A - Summaries of data quality metrics used from 10 research/guidelines

	Metadata	Timeliness	Completeness	Logical consistency/Coherence	Accuracy	Accessibility	Credibility/Reliability	Relevance	Interpretability	Usability
Vetrò et al. (2016)	x	x	x	x						
Viscusi & Spahiu (2014)		x	x	x	x					
Hernandez (2020)	x	x	x			x			x	x
European Union (2021)	x	x	x	x	x	x	x	x	x	x
Open data Charter (2015)	x	x		x	x	x		x		x
Statistics Canada ()	x	x		x	x	x	x	x	x	
Charalabidis, y. Et al., 2018	x	x	x		x					
Nikiforova, a. (2018)		x	x			x				x
Berners-lee (2006)						x		x		x
(Batini et al., 2009)		x	x	x	x					

Appendix B – Detail Breakdown of the Result of Each Data Quality Metrics

Title	Publisher	Column Completeness	Schema Completeness	Completeness	Frequency Score	Currency	Timeliness	metadata	Usability	DQS
Addresses	City of Waterloo	0.93	1.00	0.96	0.00	0.00	0.00	0.75	1.00	0.52
Addresses	City of Cambridge	0.90	0.75	0.83	1.00	1.00	1.00	0.88	1.00	0.93
Addresses	City of Kitchener	0.86	0.91	0.88	1.00	1.00	1.00	1.00	1.00	0.98
Addresses	Region of Waterloo	0.84	0.93	0.89	1.00	1.00	1.00	0.75	1.00	0.90
Building Outlines	City of Cambridge	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00	0.96
Building Outlines	City of waterloo	1.00	1.00	1.00	0.00	0.00	0.00	0.88	1.00	0.56
Building Outlines	Region of Waterloo	0.96	0.71	0.84	1.00	1.00	1.00	0.75	1.00	0.89
Building Outlines	City of Kitchener	0.80	0.72	0.76	1.00	1.00	1.00	1.00	1.00	0.95
Cycling	City of Waterloo	1.00	1.00	1.00	0.00	0.00	0.00	0.88	1.00	0.56
Cycling	City of Cambridge	0.84	0.92	0.88	1.00	1.00	1.00	0.75	1.00	0.90
Cycling	City of Kitchener	0.79	0.90	0.85	1.00	1.00	1.00	1.00	1.00	0.97
Cycling	Region of Waterloo	0.78	1.00	0.89	1.00	1.00	1.00	0.75	1.00	0.90
Road Closures	City of Cambridge	1.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.60
Road Closures	City of Waterloo	0.92	1.00	0.96	0.00	0.11	0.07	0.88	1.00	0.58
Road Closures	City of Kitchener	0.90	0.87	0.89	0.00	0.06	0.04	1.00	1.00	0.59
Road Closures	Region of Waterloo	0.84	1.00	0.92	0.00	0.01	0.00	0.75	1.00	0.51
Roads	City of Cambridge	0.94	1.00	0.97	0.00	0.04	0.03	0.88	1.00	0.57
Roads	City of Waterloo	0.81	0.64	0.72	0.00	0.06	0.04	0.88	1.00	0.52
Roads	City of Kitchener	0.68	0.88	0.78	1.00	1.00	1.00	1.00	1.00	0.96
Roads	Region of Waterloo	0.68	0.85	0.76	1.00	1.00	1.00	0.75	1.00	0.88
Trails	City of Waterloo	0.93	1.00	0.97	0.00	0.94	0.63	0.88	1.00	0.81
Trails	City of Kitchener	0.79	0.83	0.81	1.00	1.00	1.00	1.00	1.00	0.96
Trails	City of Cambridge	0.73	0.90	0.82	0.00	0.04	0.03	0.63	1.00	0.46
Trails	Region of Waterloo	0.67	0.78	0.72	0.00	0.18	0.12	0.75	1.00	0.52
Average		0.86	0.90	0.88	0.50	0.56	0.54	0.86	1.00	0.75