City of Cambridge
Executive Department

YI-AN HUANG
City Manager

Cambridge Data Quality Guide

Introduction

This guide helps you ensure that open and internal datasets in the City of Cambridge have a sufficient level of quality for analysis and decision making. High quality data is an important strategic asset for the City because it leads to better decision making and ultimately improved service delivery for residents. Alternatively, data quality problems can lead to costly or inefficient decision making and policymaking. **There is no such thing as perfect data quality,** but the practices and tools laid out in this guide can help maximize the usefulness of Cambridge's municipal data.

We hope that this data quality guide will provide useful advice for small departmental data projects and clear instructions for open datasets and interdepartmental data efforts. Cambridge's Data Governance Committee and Data Analytics & Open Data Program will regularly update this document.

Executive Summary

High quality data is an important tool for making better operational decisions and municipal policies. Data quality is a measure of the usefulness, accuracy, and correctness of data for the task at hand. In Cambridge, that means how well-suited a dataset is to inform decisions about operations and policy. Data quality is important because high quality data leads to better decisions, lower costs, more rapid innovation, and reduced risk.

Each department, division, group, or program in the City has a mission and makes decisions or policies aimed at furthering that mission. Data at least partly informs many of these decisions and policies. Therefore, data quality should not be an afterthought but should be systematically integrated into the decision-making process. To do this, staff need to identify the key datasets relevant to each decision, the key data quality and data format requirements for each dataset, and the staff-person charged with enforcing these requirements.
Data quality has several dimensions. Some of the most common data quality dimensions include:

- Accuracy: How correct is the data?
- Completeness: What portion of the data is missing?
- Timeliness: How current is the data? How often will it be updated in the future?
- Consistency: How uniform are data formats throughout a dataset and how do field names and formats conform with other, similar datasets?
- Metadata: How well explained and contextualized is the data?

The appendices of this guide contain tools for identifying data quality requirements, drafting metadata, and choosing data standards for common field types.

Data Quality and City Business

High quality data is an important tool for making better operational decisions and municipal policies. This section explains how data fits into the overall business of the City and why it is important to have sufficient data quality.

What is Data Quality?

Data quality is a measure of the usefulness, accuracy, and correctness of data for the task at hand. In Cambridge, that means how well-suited a dataset is to inform decisions about operations and policy. Data quality is important because high quality data leads to better decisions, lower costs, more rapid innovation, and reduced risk.

How does Data Quality fit into City Business?

Each department, division, and program in the City has a mission. Staff make decisions each day to accomplish those missions. Most decision should be at least partly informed by data, and the quality of that data will impact the effectiveness of each decision. Some data users, uses, and applications will have a high tolerance for data quality problems while others (particularly data applications like analysis software) will require nearly perfect data. Similarly, some data producers will be able to generate very high quality data while others will face numerous challenges collecting good data. There may also be technological, legal, or operational nuances that affect data quality. The required level of data quality for each application will depend on the needs of the data users and the constraints of the data producers.

While it might seem abstract to map data to decision making and ultimately to departmental missions, this process is crucial for properly ensuring data quality. You and your colleagues can help ensure data quality by asking a cascading series of questions:

- What is the overall mission of our department, division, group, or program?
- What decisions do we and our stakeholders need to make to achieve that mission?
- For each decision, what data would be informative for making the best decision possible?
- For each dataset, what level of data quality is necessary within each of the dimensions discussed in this guide? (Accuracy, completeness, consistency, timeliness, and metadata.)
- Who is responsible for ensuring sufficient data quality?

A worksheet to facilitate this mapping of data quality to decision making is included in Appendix A.

City of Cambridge
Executive Department

YI-AN HUANG
City Manager

Maximizing Data Quality

Each dataset is unique and requires a different process to review and assess quality (QA). However, certain concepts apply to all datasets. Data quality has several dimensions. Some of the most common data quality dimensions include:

- Accuracy: How correct is the data?
- Completeness: What portion of the data is missing?
- Timeliness: How current is the data? How often will it be updated in the future?
- Consistency: How well do data formats and fields conform with other, similar datasets?
- Metadata: How well explained and contextualized is the data?

Accuracy and Completeness

Cambridge's data analytics & open data program suggests the following accuracy standards during the dataset QA process:

- The dataset is the most complete, accurate, and current version appropriate for public release or internal sharing across City departments.
- The data have been spot checked for common errors such as missing and misplaced values.
- Any missing data points are accounted for—missing data are either filled, registered as NA, or explained in the metadata documentation.
- Columns use the most appropriate format for the data they contain.

Accuracy Thresholds

The City has a minimum data accuracy and completeness threshold of 95 percent. This means that datasets will be accepted for open data publication if fewer than 5 percent of values in the dataset are erroneous, missing, or outdated. Accuracy thresholds for data used in internal analytics projects are set on a case by case basis depending on the intended use for the analysis.
Data coordinators and data analysts are encouraged to use their involvement in the Data Analytics and Open Data Program as a platform for process improvement related to data creation and management. However, no dataset is perfect. Departments should not hesitate to publish, share, or use datasets that may contain minor problems, so long as limitations and uncertainties are well documented. Furthermore, neither data coordinators nor the core open data team will be held responsible for any dataset's deficiencies so long as those deficiencies are documented during the data review process.

Tools for Ensuring Accuracy and Completeness

The following tools and techniques can help ensure that datasets are accurate:

- **Audits:** Sample or spot check a small portion of a new dataset to identify accuracy issues that might be endemic to the dataset. You should audit all key datasets periodically.
- **Continuous identification and improvement:** Use Cambridge's Data Follow Up form to identify datasets that have missing or erroneous information.
- **Summary Statistics:** Calculate summary statistics such as mean, median, mode, outliers, number of missing values, etc., to understand patterns in the dataset.
- **Data Visualization:** Identify trends in the dataset visually through data visualization software.

- **Peer Review:** Ask a coworker or subject matter expert to review your dataset for accuracy.
- **Validation:** Use systems that force data producers to enter correct and complete information. These systems will not accept data inputs that are incomplete or outside a given range.

Consistency

When possible, the structure and format of datasets and the values within them should remain consistent across datasets. For example, any dataset about individual parcels in Cambridge should contain a parcel identifier like a Map-Lot ID. Likewise, datetime formats within a dataset should formatted the same between fields. This consistency can make your dataset more valuable by enabling the linking of it with other municipal datasets.

Tools for Ensuring Consistency

- **Use Cambridge Data Standards when possible.** Appendix C lists several common standards.
- **Automated tools** such as ITD's geocoding software can help ensure that any dataset containing an address field also contains other geographic identifiers such as Map-Lot and neighborhood. Contact ITD if you'd like access to geocoding tools (some of which are still in beta).
- **Searching Cambridge's open data portal** can help you discover similar datasets so that you can ensure that your dataset remains consistent with previously published similar datasets.
- **Validation:** Use systems that force data producers to enter properly formatted data.

Timeliness

Data driven decision-making requires up-to-date data. When creating a dataset to be published to the portal, analyzed, or shared with other departments, data coordinators should also create an update plan specifying how often the dataset will be refreshed. Data coordinators are responsible for ensuring that datasets are updated as often as specified in dataset's metadata.

Tools for Ensuring Timeliness

The following tools can help ensure that dataset remain timely:

- **Automation and ETLs.** Datasets updated more frequently than quarterly are candidates for automation. The IT department can help create Extract, Transform, and Load scripts (ETLs) to help you update your dataset on the open data portal or in another destination.
- **Calendar Reminders and Smartsheet Automations.** Set a reminder on your calendar to update your dataset periodically. Even better, use a smartsheet with automated alerts that remind you or another data coordinator to update a dataset as often as necessary.

Metadata

Metadata – which is documentation about a dataset's structure and meaning – plays an important role in helping City staff, residents, and other stakeholders use municipal data more efficiently and accurately. Data coordinators should fill out a metadata form for all new open datasets and many datasets shared or analyzed internally. Metadata should meet the following standards:

- Metadata is **complete, concise, and free of jargon.**
- Metadata **explain the process used to create** the data and summarize any changes.

- Metadata clearly **explain any limitations** or omissions for each dataset.
- Metadata clearly **identify an update frequency** and plan.

Tools for Ensuring Good Metadata

Appendix A includes Cambridge's metadata guide, a tool for creating standardized metadata pages.

City of Cambridge
Executive Department

YI-AN HUANG
City Manager

Appendix A: Data Quality Planning Guide

Use this worksheet to connect the mission and important decisions of your department, division, group, or program to the datasets and data quality standards necessary for good decision making. Use a new worksheet for each significant decision.

Mission: What is the mission of your department, division, group, or program?

Decision: What is one significant operational or policy decision that needs to be made to help achieve that mission? Use additional copies of this worksheet for additional decisions

Data Narrative: What are the main datasets relevant to this decision? How, generally, will data be used to aid in decision making or policymaking? How will data-driven insights be combined with expert judgement and qualitative analysis?

Data User Needs and Producer Constraints: In generally, will data be used to aid in decision making or policymaking? How will data- driven insights be combined with subject matter expertise and qualitative analysis? What challenges do data producers face? What are the technical and legal nuances that affect the use or production of this data?

City of Cambridge
Executive Department

YI-AN HUANG
City Manager

Datasets: For each dataset, what are the key fields and the required format for each of those fields? Which staff person oversees ensuring data quality compliance?

Dataset Name	Relevance to Decision	Data Owner	Key Fields	Required Format
e.g., Crash Data	Crash Data can help us determine what areas to conduct speed studies	Jane Smith, Traffic Analyst	Crash Location	(Latitude, Longitude)
			Crash Timestamp	YYYY-MM-DDTHH:MM:SS TZ
			Severity	Number 1-4
			Involved Pedestrian?	True/False

City of Cambridge
Executive Department

YI-AN HUANG
City Manager

Data Quality: What are the datasets and data quality requirements necessary to facilitate good decision making? What tools and tactics will help ensure that the dataset meets its data quality requirements?

Dataset Name	Data Quality Dimension	Data Quality Requirement	Tactics to Meet Data Quality Requirements
e.g., Crash Data	Accuracy/Completeness:	All Mandatory fields correct and complete	Expert review before each update
	Timeliness:	Updated quarterly	Semi-automate updates using script
	Consistency:	Standard format for location & time fields	Use default Cambridge standards
	Metadata:	Required	Use default metadate template
	Other:	None	
	Accuracy/Completeness:		
	Timeliness:		
	Consistency:		
	Metadata:		
	Other:		
	Accuracy/Completeness:		
	Timeliness:		
	Consistency:		
	Metadata:		
	Other:		
	Accuracy/Completeness:		
	Timeliness:		
	Consistency:		
	Metadata:		
	Other:		
	Accuracy/Completeness:		
	Timeliness:		
	Consistency:		
	Metadata:		
	Other:		

City of Cambridge
Executive Department

YI-AN HUANG
City Manager

Metadata Guide for New Open Datasets

Appendix B: Metadata Template

Please use this form to provide contextualizing information for new datasets that you are submitting for publication on Cambridge's Open Data Portal. This form contains three parts and should take less than 30 minutes to complete for most datasets.

Part 1: Dataset Logistics and Sensitivity Review
Part 2: General Information about the Dataset
Part 3: Information about the Dataset's Fields

If you have any questions as you fill out this form, please contact Josh Wolff in Cambridge's IT Department at 617-349-9447 or jwolff@cambridgema.gov
Thanks for your time and for your participation in the Open Data Program!

Dataset Logistics and Sensitivity Review

Please provide IT with some information about logistics and about whether the dataset contains sensitive information.

Deadlines

Please list any hard deadlines by which this data needs to be made public on the open data portal:

Sharing

Please list any other City of Cambridge departments or divisions that might use this data:

Origin

Please briefly describe who created this dataset and how:

Location

If applicable, please indicate the system or database in which the dataset resides:

Sensitive Information

Please indicate whether this dataset contains any of the following sensitive information fields:

City of Cambridge
Executive Department

YI-AN HUANG
City Manager

- ☐ Resident names and/or initials
- ☐ ID numbers (e.g., Social Security)
- ☐ Birthdate or Age
- ☐ Gender
- ☐ Home address
- ☐ Personal telephone numbers
- ☐ Personal e-mail address
- ☐ Employee personnel records
- ☐ Drivers' license number
- ☐ Information on medical or health conditions
- ☐ Financial information (credit cards, billing info, account info)
- ☐ Health information
- ☐ Student information
- ☐ Minor/Youth/Student information
- ☐ Copyrighted or proprietary information
- ☐ Information that could jeopardize public safety
- ☐ Marital status
- ☐ Nationality
- ☐ Sexual behavior or sexual preference
- ☐ Physical characteristics
- ☐ Racial or ethnic origin
- ☐ Religious, philosophical or political beliefs
- ☐ Trade union membership
- ☐ Biometric data
- ☐ Household information
- ☐ Consumer purchase or billing history
- ☐ Unique device IDs (IP/ MAC addresses)
- ☐ Location (e.g., GPS) info (including that provided by mobile devices)
- ☐ Criminal information
- ☐ Civil justice information

Please explain any selections from the list above:

Please indicate whether the dataset contains any other potentially sensitive fields not listed above:

City of Cambridge
Executive Department

YI-AN HUANG
City Manager

General Information about the Dataset

Please provide the public with some overall information about the dataset. Please use plain language and avoid technical jargon.

Dataset Title
Name of the dataset in ten words or less:

Brief Description
A few sentences or paragraphs describing the dataset:

Related Datasets
Tell the public about related datasets on Cambridge's open data portal or GIS website:

Limitations
Tell the public about any gaps or uncertainties in the dataset, including when appropriate any sensitive data held back from public dissemination:

Keywords
Three to five search terms that will help people find the dataset:

Estimated Update Frequency
How often you update the underlying data:

Civic Innovation Problem Statements Related to this Dataset
What problems could users help solve using this data:

Attachments
Name documents or files that should be included on the portal to provide more context:

Hyperlink for More Info
Where users can go to learn more about your department:

City of Cambridge
Executive Department

YI-AN HUANG
City Manager

Information about the Dataset's Fields

Please use the table below to list a name and description for each field/column in your dataset. Feel free to add more rows to the table if necessary.

Field Name	Field Description

City of Cambridge
Executive Department

YI-AN HUANG
City Manager

Appendix C. Data Quality Format Standards for Common Fields

Attribute	Format	Example	Reasoning/Notes
Date	YYYY-MM-DD	2020-03-05	This is an official format called ISO-8601. It is an unambiguous way to write dates and logically moves from largest to smallest unit.
Time	HH:MM:SS TZ	16:35:26 EDT	This is the ISO-8601 format for timestamps.
Datetime	YYYY-MM-DDTHH:MM:SS TZ	2020-03-05T16:35:26 EDT	This is the ISO-8601 format for datetimes
Cambridge Address	### Street Name	795 Massachusetts Ave	Street names fully spelled out.
Zip code	##### or #####-#### *##### or *#####-####	02140 or 02140-3910 *02140 or *02140-3910	The * tells Excel not to drop the leading zero. Zip code should be a separate field from address when possible.
Latitude	##.###	42.367	Most mapping and analysis tools use this decimal degrees version of latitude. It is best to have separate fields for latitude and longitude when possible.
Longitude	-##.###	-71.105	Most mapping and analysis tools use this decimal degrees version of longitude. It is best to have separate fields for latitude and longitude when possible.
Combined coordinates	(##.###,-##.###)	(42.367,-71.105)	
Map-Lot	###-## or ###-##a	118-33 or 118-33a	