DATA QUALITY IN PRACTICE

Why it matters and where to start

OPEN NORTH

October 2024

CONTENTS

INTRODUCTION

Before we dive in, let's take a moment to reflect:

Does your organization or project strive to implement a data-driven culture to improve its services, products, policies and decision-making?

Do your team members lose countless hours trying to access, interpret or make sense of key datasets essential to your organization's operations?

Can your team or organization afford to make decisions based on data that might be unreliable, inaccurate or irrelevant for the task at hand?

A data-driven approach to problem solving and decision-making requires actionable data. In other words, you need to easily understand and access your data while being confident that it is of sufficient quality to support your objectives.

This guide is for any person, team or organization that values data-driven decision-making and needs to ensure that their data is useful to their work.

1. WHAT IS DATA QUALITY?

Data quality is the degree to which data is fit for its intended purpose. The quality of data can be measured against several dimensions to determine if it meets the requirements to generate accurate and reliable information and insights.

8 common data quality dimensions

In this section, we highlight **8 common data quality dimensions** to take into account when assessing your data quality requirements. These 8 dimensions are not exhaustive. How you apply them will also vary depending on the data you use and its intended purpose.

Data Quality Dimension	Definition
1. Well documented	Well documented data is accompanied by comprehensive and appropriate metadata.
2. Timely	Timely data is sufficiently current for the task at hand.
3. Complete	Complete data does not contain any missing values or has an acceptable number of missing values for the task or analysis at hand.
4. Reliable	Reliable data is the result of consistent and stable data collection processes that can be repeated with the confidence that similar results will be obtained.
5. Reusable	Reusable data has the potential to serve a purpose beyond its original intent.
6. Accessible	Accessible data can be easily used and accessed by intended users.
7. Secure	Secure data is safeguarded, with access limited to individuals possessing suitable training and authorization.
8. Relevant	Relevant data is helpful and applicable for the task at hand.

Fictional case study

We will illustrate these 8 data quality dimensions using a fictional dataset containing data about food aid recipients from different neighbourhoods.

Table 1: Fictional dataset on neighbourhood food aid recipients

id	name	phone	age	annual revenue ($)	# of food baskets delivered in 2024	neighborhood	satisfaction level
1	Quinn Greenbolt		25	36374	25	a	high
2	Josef Schoen	582924564	42		44		medium
3	Rob Durgan	647314136			33	B	Medium
4	Nina Mitchell		19	17375	1	C	low

1. Well documented

Metadata is information that describes your data. It helps users understand key aspects of the data such as its source or publisher, the date of creation, the last update, the available data formats and the definitions of the variables in the dataset. Below is an example of metadata for our fictional dataset.

Food Aid Recipients for Neighbourhoods in the ABC Metropolitan Area

This dataset includes available food aid recipient data observed in neighbourhoods across the ABC Metropolitan Area.

Refresh date	**Last refreshed**	**Publisher**
Monthly	YYYY-MM-DD	Diversity and Inclusion Services
Type	**Topics**	**Formats**
Table	Food	XML | JSON | CSV

Well documented data is accompanied by comprehensive and appropriate metadata.

While this is a good start, a well documented dataset should provide users with additional information about provenance, features, and even data quality assessments.

Provenance	Describes the original source(s) of the dataset, such as data collected by various food aid organizations throughout the metropolitan area.
Features	Provide a definition for each variable in the dataset.
Data quality	Provides metrics on data quality performance and describes any known data quality issues.

To assess whether a dataset is well documented, compute how many metadata fields are empty and gather feedback on the clarity and thoroughness of the data descriptions.

2. Timely

Determining whether data is timely is not always an objective process. When there is a level of subjectivity involved, it is important to be transparent about your decisions.

Timely data is sufficiently current for the task at hand.

Example 1: Your data is objectively untimely

If your dataset dates from 2020, it is not suitable to measure the growth rate of food aid beneficiaries in neighbourhoods between 2023 and 2024.

Example 2: Determining data timeliness is subjective

You seek to understand which neighbourhoods have the highest demand for food aid to inform investment strategies and food aid policy. Some stakeholders consider any data collected more than 2 years ago as outdated. Other stakeholders disagree and find data collected within a 4 year range as relevant. There is no consensus on data timeliness.

3. Complete

> *Complete data does not contain any missing values or has an acceptable number of missing values for your task or analysis.*

In **table 2**, incomplete data values are highlighted in red. In an ideal scenario, our data would resemble **table 3**, where all data values are complete.

Table 2: Fictional dataset with incomplete data

id	name	phone	age	annual revenue ($)	# of food baskets delivered in 2024	neighbor-hood	satisfaction level
1	Quinn Greenbolt		25	36374	25	a	high
2	Josef Schoen	582924564	42		44		medium
3	Rob Durgan	647314136			33	B	Medium
4	Nina Mitchell		19	17375	1	C	low

Table 3: Fictional dataset with complete data

id	name	phone	age	annual revenue ($)	# of food baskets delivered in 2024	neighbor-hood	satisfaction level
1	Quinn Greenbolt	2196547658	25	36374	25	a	high
2	Josef Schoen	582924564	42	40000	44	A	medium
3	Rob Durgan	647314136	64	22409	33	B	Medium
4	Nina Mitchell	2382142748	19	17375	1	C	low

Completeness can be calculated by dividing the available values by the expected total number. For example, in table 2, if we expect that 100% of records have a phone number, our completeness rate would be 50% since there are only 2 available phone numbers out of a total of 4 records.

You do not systematically need a 100% completeness rate for all of your data variables. Thresholds for data completeness may vary depending on the nature of the data, its intended use and the potential consequences of using incomplete data.

4. Reliable

> *Reliable data is the result of consistent and stable data collection processes that can be repeated with the confidence that similar results will be obtained.*

Below are a few ways to interpret data reliability:

Data accuracy

Accuracy refers to how closely your data reflects reality. From a technical standpoint, it involves assessing the precision of your data collection instruments. For instance, do pedestrian sensors accurately count only pedestrians and not cyclists or cars? How precise are satellite, GPS, or Bluetooth devices in measuring distance between objects?

Data uniqueness

Certain data values must be unique, such as contacts in our fictional dataset along with their contact information. Duplicates can lead to errors in analysis and operations.

Data consistency

Consistent data follows established rules and formats and is compatible with previous data. In **table 4**, we highlight in red that *“medium”* and *neighbourhood “A”* are spelled two different ways. These inconsistencies will result in inaccuracies during analysis, as shown below.

Table 4: Inconsistent categorical data

id	satisfaction level	neighborhood
1	high	a
2	medium	A
3	Medium	B
4	low	C

faulty analysis ↓

neighborhood	satisfaction level	percentage
a	medium	100 %
A	Medium	33.3 %
A	high	33.3 %
A	low	33.3 %

Table 5: Standardized categorical data

id	satisfaction level	neighborhood
1	high	a
2	medium	a
3	medium	b
4	low	c

accurate analysis ↓

neighborhood	satisfaction level	percentage
A	medium	50 %
A	high	25 %
A	low	25 %

5. Reusable

> *Reusable data has the potential to serve a purpose beyond its original intent.*

Whether data is reusable can depend on a variety of factors. Below is a non-exhaustive list:

1. Machine readability

… machine-readable data can be automatically read and processed by computers and is typically stored in formats like CSV, JSON, or XML.

2. Data standardization

… data standardization is crucial for reuses that involve comparing your dataset with other datasets. For example, imagine you are conducting a regional study on food aid, and you want to combine the fictional dataset with similar data from other cities collected by other organizations. It is essential that all datasets use standardized methods for collecting data and have consistent data structures. This ensures that the data can be easily compared and analyzed together.

3. Intended use

…let's assume our fictional dataset dates from 2020. Considering the context of COVID-19 is important. The pandemic may have increased reliance on food aid due to disruptions in supply chains and high unemployment rates. While this dataset could be valuable for studying the pandemic's impact on local food aid demand, its specific context might make it unsuitable for other purposes, potentially skewing the data beyond usability.

4. Data quality

…the other data quality dimensions presented in this guide can significantly impact whether or not data can be reused. For example, the more data is incomplete, the less likely it can be reused.

6. Accessible

> *Accessible data can be easily used and accessed by intended users.*

Accessibility generally depicts the extent to which data is available and how easily it can be accessed by users.

Measuring accessibility can range from assessing whether the dataset is:

- available in the appropriate formats
- structured and organized in a manner that is ready for use
- whether individuals requiring access to the dataset to perform their work have appropriate access permissions.

Generally speaking, accessibility is not a quantitative measure, and we recommend using a checklist[1] that addresses any documentation and machine readability issues.

7. Secure

> *Secure data is safeguarded, with access limited to individuals possessing suitable training and authorization.*

Data quality has connections with each of the three pillars of information security (data confidentiality, data availability, and data integrity). However, the most apparent connection lies with data integrity.

Data integrity refers to data that remains unaltered without proper authorization. For instance, if a team member intentionally or unintentionally alters all phone numbers and ages within our fictional dataset, the integrity of the data is compromised.

Implementing adequate safeguards and security measures to prevent this type of data incident is essential. Maintaining and regularly updating a registry of security incidents and breaches will enable you to monitor the effectiveness of your security safeguards.

[1] Here is an example of a checklist for accessible spreadsheets provided by the UK government (Government Analysis Function website).

8. Relevant

It is important to note that data relevance can be subjective and may vary based on context or individual perspectives. What is relevant in one scenario may not be in another. That is why data relevance begins by defining clear objectives, and having a precise understanding of what information is needed and why, prior to collecting any data.

Using our fictional dataset as an example, let's explore two scenarios where our data may be considered irrelevant for its intended use.

Relevant data is helpful and applicable for the task at hand.

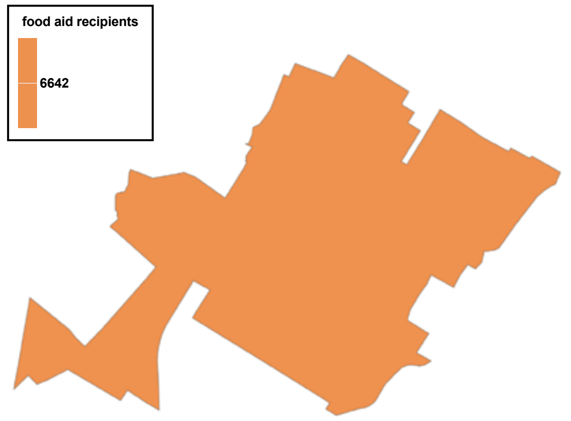
1 - Issues With Data Type

Our goal is to enhance the satisfaction of food aid beneficiaries. While our dataset includes satisfaction levels categorized as high, medium, and low, it primarily allows us to assess average satisfaction levels and identify any variations among neighbourhoods. However, this data offers restricted insights into the underlying causes of dissatisfaction or opportunities for enhancement. To better understand and address these issues, relevant data should provide explanations from beneficiaries regarding their current satisfaction levels as well as suggestions for improvement.

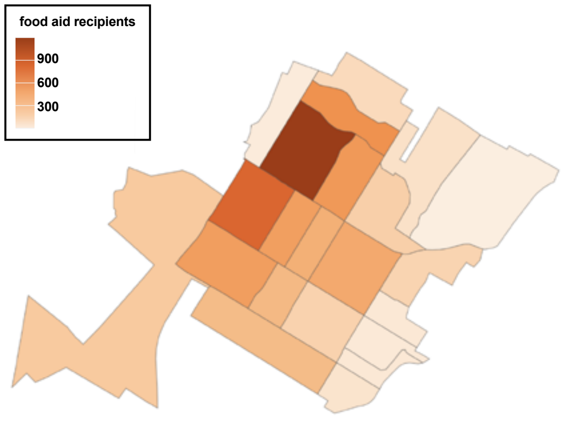
2 - Issues With Data Scale

Let's imagine our fictional dataset contains spatial data. Food aid organizations and local governments have assigned you the task of generating a heat map pinpointing neighbourhoods with the highest number of food aid recipients. This initiative aims to enhance service delivery in areas with the greatest need. However, upon examination, you realize that the data at your disposal is aggregated at the city scale, rather than the neighbourhood scale. This means that any insights aimed at targeted interventions on a neighbourhood level would be inconclusive without access to data at the appropriate spatial scale.

Map with irrelevant data at the wrong spatial scale



Map with relevant data at the appropriate spatial scale



2. WHY DATA QUALITY MATTERS

The consequences of using bad data

According to a 2017 article in the Harvard Business Review,[2] on average, 47% of newly created data contains at least one critical error. The article highlights that "bad data wastes time, increases costs, weakens decision making, angers customers, and makes it more difficult to execute any sort of data strategy".

Using bad data not only leads to operational and productivity issues, it can skew the outcomes, decisions and policies informed by the use of data, which in turn can harm individuals, groups and entire communities.

Simply put, using bad data wastes time and resources, and leads to poor and potentially harmful decision-making.

Is data quality a legal requirement?

In Canada, data quality is a legal requirement under existing privacy laws which only apply to personally identifiable information (PII). For example, adequate data quality regarding PII is implied for all commercial organizations under the **Personal Information Protection and Electronic Documents Act** (PIPEDA), where principle 6 requires that personal information must be as accurate, complete, and up-to-date as possible in order to properly satisfy the purposes for which it is to be used.

Canadian provinces have internal requirements for data accuracy; however they are very limited. For example, in the case of the province of Ontario, data quality requirements only apply to open data published by provincial ministries and agencies.

Under laws such as the **General Data Protection Regulation** (GDPR) in the European Union or the **California Consumer Privacy Act** (CCPA) in the United States, organizations must ensure that the personal data they collect, process, and store meets certain data quality requirements (such as accurate and timely data). Failure to comply with these data quality requirements can result in legal consequences, including fines and penalties.

[2] "Only 3% of Companies' Data Meets Basic Quality Standards" (Nagle, Redman, Sammon, 2017)

3. HOW TO GET STARTED ON YOUR DATA QUALITY OBJECTIVES?

You can only improve what you measure. Developing and implementing data quality standards and requirements suited to your context and setting clear goals to achieve them are necessary for your organization or team to monitor and measure its data quality.

5 key steps

Here are 5 key steps to take into consideration to define and achieve your data quality objectives:

1. Define a clear data use case	…understand your objectives and what you want to achieve.
2. Identify your data sources	…data relevance begins here. Think about what data could help you achieve your objectives.
3. Define your data quality requirements	…identify which data quality dimensions are most important for your use case and define specific requirements for each dimension.
4. Assess your data quality	…measure how your dataset(s) stack up to your data quality requirements.
5. Develop a data quality action plan	…ensure you can meet and maintain your data quality requirements over time.

Key conditions for successfully achieving data quality objectives

Define a clear data use case

Clearly defining your use case will help you identify the data that is required for achieving your goals. Collecting only the data that you need minimizes the time, effort, and resources required to manage it and maintain its quality.

Identify and engage stakeholders

Identify all stakeholders impacted by your use case. Engage with them to incorporate diverse perspectives and mitigate biases in data identification. This ensures the data collected supports the successful resolution of your use case.

Begin thinking about data quality requirements early

Consider data quality requirements before you start collecting data. Data quality is relevant at every stage of the data life cycle, including the planning and preparation phase. Assess the relevance of data when identifying the right data to support your use case.

Explore existing data quality standards

Explore existing international[3] and domain-specific standards[4] before creating new ones. Customizing data quality requirements to your specific needs is important, but starting with established standards can save time and enhance the likelihood of your data being compatible with other datasets.

Define clear roles and responsibilities

Appoint a staff member accountable for the team's data quality efforts. This individual can delegate tasks such as defining requirements, conducting initial data quality assessments, drafting action plans, and performing data quality audits as needed. Leadership should ensure the data quality team has adequate resources to engage relevant stakeholders.

Train your staff on data quality concepts

Educate your team on key data quality concepts. While this guide provides an overview, your team may need to familiarize itself with requirements, standards, and best practices specific to your field.

Develop a data quality monitoring and evaluation plan

Create a comprehensive plan for monitoring and evaluating data quality. Allocate resources for regular maintenance and audits to ensure ongoing data quality.

Utilize tools to support data quality objectives

For more complex data use cases, invest in or develop tools that support your data quality goals. Consider software that can identify and merge duplicates, detect errors or anomalies, and possibly correct them.

(3) Examples: **ISO 25012** for international standards on data quality; **ISO 8000** for data quality and enterprise master data; **ISO 8601** for international standards covering the worldwide exchange and communication of date and time-related data.

(4) For example, quality benchmarks for **Realtime Transit data**.

GET STARTED TODAY!

To begin your data quality journey, you can use Open North's data quality checklist. You will complete your data quality assessment in 5 simple steps and receive a clear set of data quality requirements and a concrete action plan.

Click on the link below to explore, use and adapt Open North's data quality checklist.

Data Quality Checklist

DATA QUALITY CHECKLIST

For data you can use and trust

OPEN NORTH

October 2024

If you require support to implement your data quality requirements or to adapt existing standards to your specific context, Open North can support you through its targeted support service: contact us at info@opennorth.ca.

About Open North

Open North is a Canadian nonprofit dedicated to advancing the common good. As an organization with expertise in data governance and digital strategy, we work alongside governments, nonprofits, and mission-aligned businesses to create transformative digital strategies and data governance frameworks.

Open North's team is made up of professionals with a wide range of expertise, including in government, strategic and operational planning, urban planning, community building, information technology, applied research, international development, and policy development. With our diverse backgrounds and skills, Open North's team members bring valuable perspectives and experience to all projects.

Open North is part of Montréal in Common, a project led by the City of Montréal as part of the Smart Cities Challenge, carried out with the financial support of the Government of Canada.

opennorth.ca

About the Smart Cities Challenge and Montréal in Common

Montréal in Common is an innovation community led by the City of Montréal, whose partners are experimenting with solutions regarding access to food, mobility and municipal bylaws, with a view to rethink the city. Montréal in Common projects are made possible thanks to the prize awarded to the City of Montréal by the Government of Canada as part of the Smart Cities Challenge.

Author: Samuel Kohn

Canada Montréal SMART CITIES CHALLENGE Montréal en commun

Data Quality in Practice