MOBILITY DATA ANONYMIZATION

A practical introduction to responsible data anonymization decision-making

OPEN NORTH

January 2026

CONTENTS

INTRODUCTION: MOBILITY DATA AND PRIVACY

Most, if not all, organizations providing services in the mobility space collect some amount of personal information from their customers. Whether you run a transit agency, a bikeshare fleet, or a ride-hailing app, knowing who your users are and where they are going is often essential to the service.

When personal information is held and used solely by the organization that collected it, it generally does not pose significant privacy issues - provided that the information was collected with the informed consent of individuals, for a reasonable purpose (e.g., to process a payment or enable navigation), and that appropriate safeguards are in place to protect the data.[1]

However, the landscape is shifting. It is increasingly common for organizations to use the personal information they have collected for additional purposes, or to pool their data with other organizations. These initiatives - often aimed at generating new insights, improving city planning, or creating new data products - introduce new complexities. When data leaves the original "silo" or is repurposed, the privacy risks change, and so must our approach to managing them.

> Data anonymization is about **reducing** risk while ensuring the data is still useful.

There are a variety of tools available to an organization to reduce that risk. However, it is important to remember that the elimination of this risk is not possible. A holistic approach to risk mitigation includes establishing privacy management roles, policies, and processes; implementing appropriate security safeguards to protect personal information; and leveraging various privacy-enhancing technologies (PETs), including data anonymization and de-identification techniques.

One of these tools is data anonymization, in which datasets are modified to make it more difficult to identify individuals using the available data.

[1] P-39.1 - Act respecting the protection of personal information in the private sector

KEY TERMS AND DEFINITIONS

Spectrum of identifiability

When thinking about how identifiable a given dataset is, it is useful to consider a spectrum of identifiability.[2] This spectrum of identifiability defines three states of information:

- **Personal information**
  - **Identified information:** Information which, by itself, directly identifies an individual
  - **Identifiable information:** Information for which there is a serious possibility that it could be associated with an identifiable individual.
- **Non-personal information**
  - **Non-identifiable information:** Information for which there is no serious possibility that it could be associated with an identifiable individual.

In general, the more identifiable the information, the greater its utility for a range of purposes. However, this added utility comes with increased risks. The benefit of adopting a spectrum approach to defining identifiability is that it allows for a broad range of innovative uses of information while accounting for and mitigating a reasonable level of residual risk.

Anonymization vs. deidentification

Quebec's Law 25 (the Act to modernize legislative provisions as regards the protection of personal information) makes an important distinction:

De-identified information is data from which direct identifiers (such as names) have been removed, but where it may still be possible to identify individuals with additional effort or information. For example, replacing a person's name with a random code is a form of de-identification. Under the law, de-identified data is still considered personal information, meaning that all privacy obligations continue to apply.

Anonymized information, by contrast, has been processed so thoroughly that it is virtually impossible to identify individuals, either directly or indirectly. When done properly, anonymized data is no longer considered personal information, which allows for much broader use and sharing.

[2] Canadian Anonymization Network, “Spectrum of Identifiability,” 2020, https://deidentify.ca/wp-content/uploads/2020/10/CANON-States-of-Data-One-Pager.pdf.

Threats and risks to mobility data

Personal information that directly or indirectly identifies individuals carries an inherent risk of misuse. Under Quebec's private sector privacy legislation, organizations that collect and hold personal information are required to take appropriate measures to prevent unauthorized access, use, and disclosure.

In the mobility sector, these risks are heightened by the highly granular nature of location data. Even small fragments of information—when combined over time—can reveal detailed insights about individuals' lives. Different types of mobility datasets therefore present different levels of privacy risk. Let's consider two common mobility data types: trip records and customer account data.

Trip records, for example, typically include elements such as:

- a customer or account identifier;
- origin and destination points;
- timestamps; and
- distance or duration.

While trip records may not contain names, combinations of location and time can be highly identifying, particularly when trips are repeated and patterns emerge.

Customer account data often includes:

- first and last names;
- address and contact details;
- payment information; and
- subscription or usage history.

When customer account data is linked with trip records, the risk of re-identifying individuals increases significantly.

Taken together, these characteristics give rise to several common privacy threats in mobility data, including:[3]

- Identifying individuals, by inferring home or work locations from regular travel patterns;
- Predicting behaviour, such as identifying when a person's home is likely to be unoccupied;
- Revealing sensitive information, for example through repeated visits to medical or treatment facilities; and
- Connecting the dots, where ostensibly "anonymous" trip data is matched with other information sources, such as social media posts tied to specific times and locations.

Given these risks, anonymization is often considered as a way to enable the use or sharing of mobility data while managing privacy concerns. However, deciding whether and how to anonymize data requires looking beyond the dataset itself. It requires understanding the **data situation**: the relationship between the data and the environment in which it exists, including other data it may be combined with, the people who interact with it, their roles and responsibilities, and the systems used to process and protect it.

The next section introduces this concept and explains how evaluating the data situation provides a foundation for sound anonymization decisions.

[3] Henry Laville. « Rapport : Symposium « Anonymisation des données » » (OBVIA, 2024). https://doi.org/10.61737/XAFK3054.

A STRUCTURED APPROACH TO DATA ANONYMIZATION DECISION-MAKING

Anonymization is both a risk-management process and a decision-making exercise. At its core, it helps answer a key question: *Should we share or release this data, and if so, in what form and under what conditions?*

Crucially, you cannot determine whether data is safe to share or release by looking at the dataset alone. A structured approach is necessary to ensure that anonymization decisions are based on a holistic assessment of risk arising from the interaction between data, people, legal frameworks, IT systems, and organizational culture and governance practices.

This decision-making framework comprises three main activities:[4]

- evaluate the data situation (that is, the relationship between the data and its environment);
- assess and control disclosure risks; and
- manage impacts.

This guide focuses primarily on evaluating the data situation. This step is foundational, as it sets the context for assessing and controlling disclosure risks and for managing the impacts that may result from those risks.

Note on examples:

Throughout this section, we will reference two realistic scenarios to show how the process works in practice:

Scenario A: Public Transit Data

The City of Verdeville's public transportation agency collects ridership data to support ongoing service planning and operational improvements. The agency's objective is to better understand how residents use the transit system so that schedules, routes, and transfer points can be adjusted to better meet demand.

The data collected for this purpose includes information such as wait times, frequently used stops and routes, and common transfer locations. This data is analyzed in aggregate to identify patterns and trends rather than to examine the behaviour of individual riders.

Access to the data is strictly limited. The dataset is stored on the agency's internal servers, and only three authorized staff members are permitted to access it as part of their job responsibilities. The data is used exclusively for internal analysis and is not shared outside the organization.

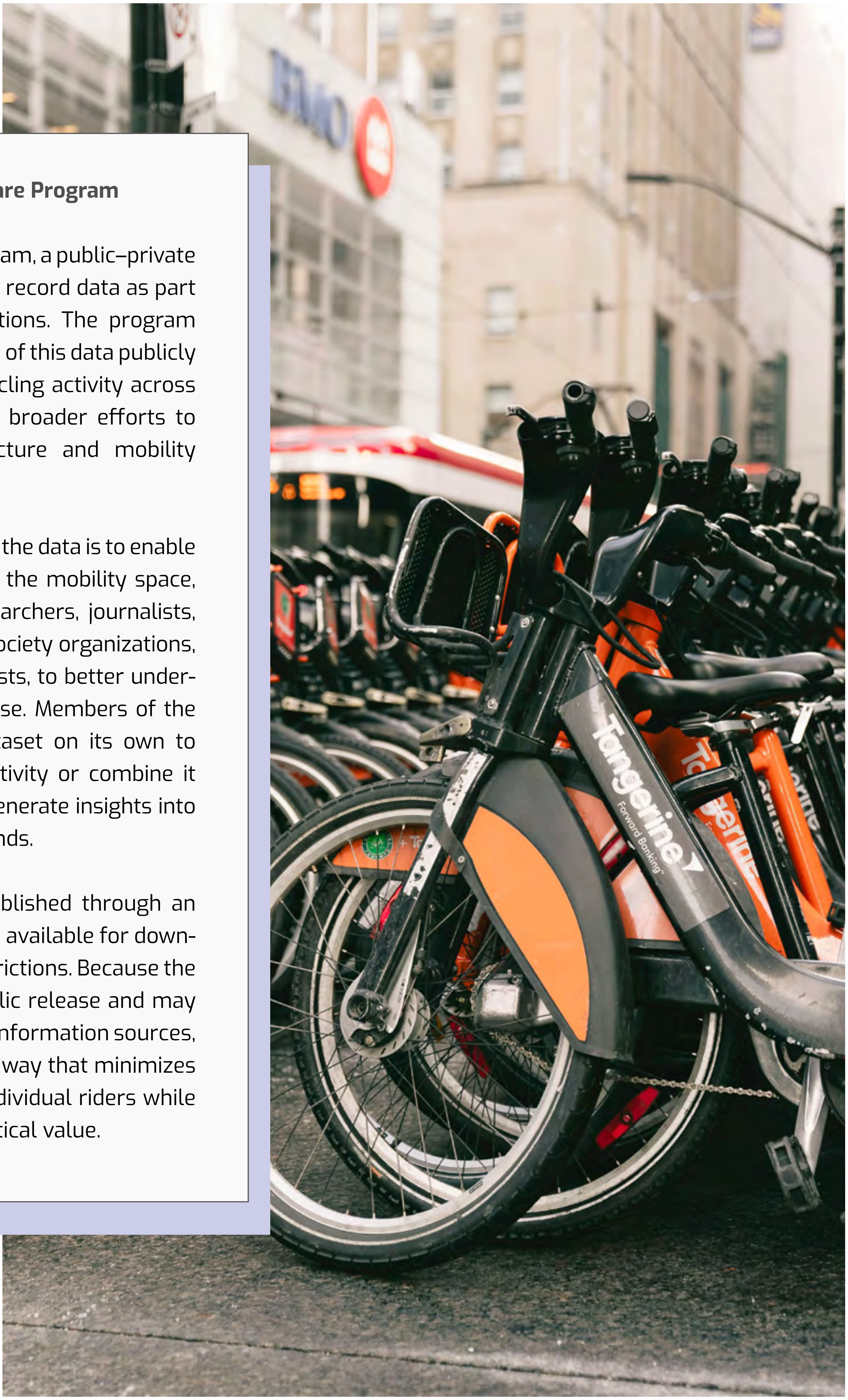
[4] This approach is based on the Anonymisation Decision-making Framework developed by the UK Anonymisation Network (UKAN).

Scenario B: City Bike Share Program

The City Bike Share Program, a public–private partnership, collects trip record data as part of its day-to-day operations. The program wishes to make a portion of this data publicly available to highlight cycling activity across the city and to support broader efforts to improve bike infrastructure and mobility planning.

The purpose of releasing the data is to enable public actors working in the mobility space, including academic researchers, journalists, government staff, civil society organizations, and private sector analysts, to better understand patterns of bike use. Members of the public may use the dataset on its own to report on bike share activity or combine it with other datasets to generate insights into wider transportation trends.

The trip records are published through an open data portal and are available for download without access restrictions. Because the data is intended for public release and may be combined with other information sources, it must be prepared in a way that minimizes the risk of identifying individual riders while still preserving its analytical value.



Activity 1: Evaluate the data situation

This activity will help you to identify and frame those issues relevant to your data situation. You will encapsulate and systematically describe the data, what you are trying to do with them and the issues thereby raised. A well-conducted data situation evaluation is the basis for the next core activity.

Specific tasks

- Describe the data situation
- Know your data
- Understand the use case
- Understand the legal issues
- Understand the issue of consent and your ethical obligations

Describe the data situation

A data situation describes how a dataset exists in the real world and how it interacts with its surrounding environment. That environment includes the people who access the data, the systems that store and process it, other datasets it may be combined with, and the rules and agreements that govern its use.

In practice, data rarely stays in one place. Depending on the use case, it may move through several systems or teams, and sometimes across organizational boundaries or into the public domain. Each movement or transformation introduces new privacy considerations.

Mapping the data flow from the point of collection onward helps make these dynamics visible. It allows you to see where the data travels, how it changes, and where risks may be introduced. Just as importantly, it helps clarify who is responsible for the data at each stage.

Key questions to ask:

- Who currently controls or manages this data?
- Who will receive or access it next?
- Will the data be used only by internal staff, shared with partners, or made public?
- Where will the data be stored or accessed (for example, secure internal servers, cloud platforms, or public websites)?
- What governance controls are in place, such as contracts, access restrictions, or data sharing agreements?

In the Verdeville Transit scenario, ridership data moves from an encrypted internal database to a secure analyst workstation. Access is limited to a small number of authorized staff, and the environment is closed and well controlled.

In the City Bike Share scenario, trip data moves from an internal database through a processing environment and is ultimately published on an open data portal. This open environment allows anyone to access and reuse the data, which significantly increases exposure and risk.

Know your data

Once you understand the data situation, the next step is to look closely at the data itself. This involves understanding who the data is about, what form it takes, what kinds of variables it contains, and what properties make it more or less sensitive.

Data subjects

Data subjects are the people the data relates to, such as transit riders or bike share users. Understanding who they are and how they interact with the service helps set expectations about appropriate data use.

Key questions to ask

- Who are the individuals represented in this dataset?
- Is the service optional or essential from their perspective?
- What expectations might they reasonably have about how their data is used?

Data types

Data can take different forms, ranging from detailed individual-level records to highly aggregated summaries. The level of detail has a direct impact on privacy risk.

Key questions to ask

- Is this dataset made up of individual records (microdata) or aggregated statistics?
- Could the same analytical goals be met using more aggregated data?

Variable types

Most mobility datasets contain a mix of variable types. Direct identifiers can identify a person on their own, such as a name or social insurance number. Indirect identifiers do not identify someone by themselves but can do so when combined with other information, such as locations, dates, or times. Target variables are attributes that are sensitive or revealing and that someone might want to infer about an individual.

Key questions to ask

- Does the dataset include any direct identifiers?
- What indirect identifiers are present, such as precise locations or timestamps?
- Does the dataset include sensitive attributes that could be revealing even if identities are not explicit?

Data properties

Dataset properties can increase or decrease disclosure risk, but at this stage they are used only as general indicators to flag areas that require closer analysis later, not as a substitute for a full risk assessment. Relevant key data properties can include:

- **Data quality:** High-quality, accurate data is more useful but can increase disclosure risk, while small amounts of error may unintentionally make identification harder.
- **Age of the data:** Older data is generally less risky because people's locations, routines, and circumstances change over time, although it may also become less accurate.
- **Level of detail:** Highly granular data increases the likelihood that individuals can be singled out within a dataset.
- **Hierarchical or grouped data:** Data that links individuals within groups or shared locations can increase risk because group-level combinations may be unique.

- **Time-stamped or longitudinal data:** Data collected over time increases risk by revealing unique patterns or changes in behaviour.
- **Population versus sample coverage:** Population-level data is generally riskier than sample data because it is clearer who is included.

Key questions to ask

- How precise is the data, particularly for time and location?
- How recent is the data?
- Does the dataset cover an entire population or only a subset?

Understand the use case

A clear and shared understanding of the use case is essential. When you know exactly why the data is being used and by whom, it becomes much easier to decide what data is truly necessary and how much risk is acceptable.

Key questions to ask

- Why do we want to share, release, or reuse this data?
- Who is expected to use it, such as internal teams, partners, researchers, or the public?
- What kinds of analysis or reuse does the data environment realistically support?

In most cases, the use case will be one of the following four types, depending on how reuse is framed.

1. Sharing of data with another party
2. Dissemination of data (either publication or release)
3. Continued use of data beyond a previously defined retention period
4. Re-use by the same organization of data for a different purpose other than that for which they were collected.

> In the Verdeville Transit scenario, the use case is internal analysis to improve service delivery, with limited access and no external sharing.

> In the City Bike Share scenario, the use case is public dissemination, with the expectation that many different users may combine the data with other sources.

Understand the legal context

Before anonymization is complete, you are still processing personal information. This means you must understand and comply with the privacy legislation applicable to your organization, depending on whether your organization is a public body or a private organization.

While public bodies and private organizations are subject to different privacy statutes, both must comply with the Regulation respecting the anonymization of personal information when anonymizing data. The regulation requires organizations to assess and reassess re-identification risks over time and to keep records of how anonymization decisions are made.

Key questions to ask

- Which privacy law applies to our organization?
- Have we clearly defined a serious and legitimate purpose for anonymization?
- Has a competent person been designated to oversee the anonymization process?
- Have direct identifiers been removed before conducting risk analysis?
- Are we documenting techniques and decisions in an anonymization register?

Understand the issue of consent and your ethical obligations

Legal compliance is not the same as earning and maintaining public trust. Even when data is anonymized, organizations have ethical responsibilities to the people whose data they collected.

People's expectations are shaped by what they were told at the time of collection and by how essential the service is to their daily lives. Mobility data, in particular, can reveal sensitive patterns that individuals may not fully appreciate.

Key questions to ask

- How closely does this use of the data align with what users originally expected?
- Are we being clear and transparent about how mobility data is analyzed and shared?
- Could this data use disproportionately affect certain groups, such as people with disabilities?
- Do users have meaningful choice, or are there power imbalances due to the nature of the service?
- Does this data use clearly serve a public benefit?

Even when not legally required, providing opt-out options where feasible and regularly reviewing practices with community stakeholders can help ensure that data use remains responsible, understandable, and aligned with public values.

This guide is accompanied by a **Data Situation Evaluation Checklist** that you can use to ask the right questions about your data situation.

NEXT STEPS

Once you have evaluated your data situation, you are in a much stronger position to decide what to do next. The next phase focuses on understanding the level of disclosure risk associated with your data and applying appropriate technical measures to reduce that risk.

This phase is divided into two main activities. First, you assess the disclosure risks in more detail. Second, you select and apply technical approaches to control those risks. These steps build directly on the contextual understanding developed in Activity 1 and should not be approached in isolation.

Activity 2: Assess and control disclosure risks

The purpose of this activity is to bring together the processes you will use to both measure and manage the disclosure risks associated with your data situation. At this stage, the goal is not to find a single "correct" anonymization technique, but to select a combination of approaches that are appropriate for your specific context.

Assess disclosure risk

Once you have determined that there is more than a negligible risk of disclosure, you need to take a closer look at how that risk might materialize. This involves examining the data in light of its environment, its intended users, and the ways it could reasonably be combined with other information.

Risk is not measured in the abstract. It depends on who might access the data, what they could do with it, and how much effort would be required to identify individuals. A dataset released internally to a small, trusted team presents very different risks than the same dataset released publicly.

Key questions to ask

- Could an individual be singled out within this dataset, even if no names are present?
- Could this dataset be linked with other available data to identify individuals?
- Could someone infer new or sensitive information about individuals from the data?
- Who are the most likely adversaries, and what capabilities would they realistically have?
- What would the potential consequences be if re-identification occurred?

Answering these questions helps you move from a general sense of risk to a clearer understanding of where the main vulnerabilities lie.

Control disclosure risk

Once disclosure risks are understood, you can decide which technical approaches are most appropriate to reduce them. This is the step that many people associate most closely with anonymization, but it only makes sense after you have carefully considered the data situation and risks.

Common anonymization approaches

There is no single way to anonymize data. In practice, anonymization usually involves combining several techniques, applied differently across variables within the same dataset. The goal is to reduce risk to a very low level while preserving as much analytical value as possible.

DESCRIPTION	KEY CONSIDERATIONS	EXAMPLE
Removal Deleting variables entirely so they are no longer available in the dataset. This typically includes direct identifiers and high-risk indirect identifiers.	› Is this variable essential for the use case? › Would removing it significantly reduce risk?	› Dataset includes: Name, Address, Age, Colour. › Action: Remove Name and Address.
Generalization Replacing precise values with broader ranges or categories. This reduces uniqueness while preserving useful patterns.	› Can precise values be replaced with ranges? › How much precision is actually needed for analysis?	› Distance: "10–50 km" instead of "18 km". › Location: "Montreal" instead of "4388 Rue Saint-Denis".
Suppression Removing specific records (rows) that are distinct outliers or have very few matching peers, making them easy to identify.	› Can precise values be replaced with ranges? › How much precision is actually needed for analysis?	› Trip Data: Removing a single trip that occurred at 3:00 AM in a remote area because it is unique.
Pseudonymization Replacing identifiers with alternative values, such as random codes or hashes. (Note: This is a security measure, not full anonymization).	› Is longitudinal analysis (tracking over time) required? › Is the environment strictly controlled?	› User ID: "John Smith" is replaced with "User_8923".

Activity 3: Manage impacts

Even after data has been anonymized, your responsibilities do not end. This activity focuses on ensuring that risks remain low over time, that stakeholders are appropriately informed, and that you are prepared to respond if something goes wrong.

Specific tasks

- Maintain stakeholders' trust
- Plan what to do if things go wrong
- Monitor the data situation

Maintain stakeholders' trust

Engaging stakeholders early and communicating clearly about how data is used can significantly reduce the impact of any future issues. When people understand why data is being shared and how risks are being managed, they are less likely to feel surprised or harmed if problems arise.

Key questions to ask

- Have we explained, in plain language, how and why the data is being anonymized?
- Are users aware that their data may be analysed or shared in this form?
- Do stakeholders have a way to ask questions or raise concerns?

Transparency is especially important for mobility data, which can feel highly personal even when anonymized.

Plan for things going wrong

Despite taking every precaution, there is always a possibility that something will go wrong, leading to a data breach. In this scenario, your organization needs a clear process in place to manage the crisis. While crisis management policies differ depending on the organization and the type of data held, there are several core areas yours should cover. Consider various breach scenarios when crafting your policy.

Key questions to ask

- **Breach management:** What are the immediate next steps? What are the roles of staff members, and who is responsible for decision-making?
- **Notification:** Who should be notified and when? For example, when staff discover a disclosure has occurred, whom do they notify and how?
- **Review:** How will you determine what went wrong, and what measures can be put in place to mitigate the risk of recurrence?
- **Communication:** How will you communicate with your stakeholders regarding the breach? Include your legal requirements for communicating with stakeholders, the public, and the government.

Note: Under Quebec's Law 25, in the event of a confidentiality incident involving personal information, organizations are generally required to notify the Commission d'accès à l'information du Québec as well as the individuals impacted if the breach presents a "risk of serious injury."

Monitor & maintain

Anonymization is not a one-time exercise. Over time, new technologies, new datasets, or changes in law can increase re-identification risks.

Organizations should have a plan to regularly review their data situation and anonymization decisions.

Key questions to ask

- Do we have a process for tracking changes that could affect risk, such as new datasets or analytical techniques?
- How often do we reassess anonymized datasets?
- Do we maintain a register of data that has been shared or released?
- Are our systems and communication channels still functioning as intended?

Your systems that support your data anonymization need to be maintained and the avenues for stakeholders to communicate with you regarding the data need to remain active.

CONCLUSION

Decisions about anonymization are always contextual and may change over time as data, technology, and expectations evolve. There is no single threshold that applies to all datasets or use cases.

What matters most is following a structured, well-documented process. By evaluating context, assessing and controlling risk, and actively managing impacts over time, you place yourself in a strong position to justify your decisions, explain them to others, and responsibly unlock the value of data.

Get started today

To begin documenting your process, you can use the Data Situation Evaluation Checklist that accompanies this guide. This tool will help you to ask the right questions as you evaluate your data situation and determine how data anonymization can help you to protect the privacy of your data subjects while ensuring the data is useful.

If you require support to implement an anonymization decision-making framework or to adapt these guidelines to your specific context, Open North can assist you through its targeted support service.

TOOL: DATA SITUATION EVALUATION CHECKLIST

Describe the data situation

- □ Who currently controls or manages this data?
- □ Who will receive or access it next?
- □ Will the data be used only by internal staff, shared with partners, or made public?
- □ Where will the data be stored or accessed (for example, secure internal servers, cloud platforms, or public websites)?
- □ What governance controls are in place, such as contracts, access restrictions, or data sharing agreements?

Know your data

Data subjects

- □ Who are the individuals represented in this dataset?
- □ Do these individuals have a choice when it comes to using the service through which their data was collected?
- □ What expectations might they reasonably have about how their data is used?

Data types

- □ Is this dataset made up of individual records (microdata) or aggregated statistics?
- □ Could the same analytical goals be met using more aggregated data?

Variable types

- □ Does the dataset include any direct identifiers?
- □ What indirect identifiers are present, such as precise locations or timestamps?
- □ Does the dataset include sensitive attributes that could be revealing even if identities are not explicit?

Data properties

- □ How precise is the data, particularly for time and location?
- □ How recent is the data?
- □ Does the dataset cover an entire population or only a subset?

Understand the use case

- □ Why do we want to share, release, or reuse this data?
- □ Who is expected to use it, such as internal teams, partners, researchers, or the public?
- □ What kinds of analysis or reuse does the data environment realistically support?
  - › Sharing of data with another party
  - › Dissemination of data (either publication or release)
  - › Continued use of data beyond a previously defined retention period
  - › Re-use by the same organization of data for a different purpose other than that for which they were collected.

Understand the legal context

- □ Which privacy law applies to our organization?
- □ Have we clearly defined a serious and legitimate purpose for anonymization?
- □ Has a competent person been designated to oversee the anonymization process?
- □ Have direct identifiers been removed before conducting risk analysis?
- □ Are we documenting techniques and decisions in an anonymization register?

Understand the issue of consent and your ethical obligations

- □ How closely does this use of the data align with what users originally expected?
- □ Are we being clear and transparent about how mobility data is analyzed and shared?
- □ Could this data use disproportionately affect certain groups, such as people with disabilities?
- □ Do users have meaningful choice, or are there power imbalances due to the nature of the service?
- □ Does this data use clearly serve a public benefit?

About Open North

Building trust in data, for the common good

Open North is a not-for-profit organization dedicated to advancing the common good. Working alongside governments and civic-minded organizations of all sizes, we provide data expertise to enhance decision making, drive innovation, improve public and civic services, and address society's most pressing challenges.

Our work focuses on building the capacity of organizations to make better decisions about managing their data so that it is useful, actionable, secure, and trustworthy throughout its entire lifecycle. At Open North, we combine deep expertise in data with a multidisciplinary approach. Our team includes urban planners, software engineers, community organizers, data scientists, cybersecurity and IT audit specialists, sociologists, geographers, and technology lawyers, bringing diverse perspectives to every project.

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: John Griffin and Steven Coutts

Canada
Montréal
SMART CITIES CHALLENGE
Montréal en commun

Mobility Data Anonymization