NPD CREATION AND DE-IDENTIFICATION GUIDE

This guide explains how to create non-personal data (**NPD**) and how to complete The City of Calgary's Record of Creation of Non-Personal Data form, in accordance with the Protection of Privacy Act (**POPA**) of Alberta. It is intended to help City of Calgary employees complete the form accurately and consistently, with appropriate supporting detail, so they can be confident that NPD has been created and documented correctly.

WHAT'S IN THIS GUIDE?

1. PRIVACY BASICS
What personal information is, what NPD is, and why we de-identify data.

2. WALKING THROUGH THE FORM
Question by question, guidance for filling out the NPD record.

3. UNDERSTANDING RE-IDENTIFICATION RISK AND SCORING
How to score re-identification risk and techniques for reducing risk.

4. HOW TO DE-IDENTIFY DATA — SIX METHODS
Six common ways to de-identify data to reach NPD status for sharing or publishing.

5. WORKED EXAMPLES
A few examples to illustrate possible scorings of datasets.

APPENDIX — REFERENCE DATA DICTIONARY
Illustrative examples of common data attributes and how they're typically considered in re-identification risk assessments.

💡 **When in doubt, reach out:**

The Privacy Office is here to answer your questions. Contact them at AccessAndPrivacy@Calgary.ca

💡 **Help with De-Identification of Data:**

Deidentify@Calgary.ca.

1. PRIVACY BASICS

1.1. WHAT IS PERSONAL INFORMATION?

Personal information (**PI**) is recorded information about an identifiable individual. If a person could reasonably be identified from the data, it is considered personal information. PI can also be understood by looking at the types of information that affect re-identification risk and privacy:

- **Direct identifiers:** Information that can identify a person on its own, such as a name, email address, phone number, employee ID;
- **Indirect identifiers:** Information that may not identify someone by itself, but could identify them when combined with other details, such as age, postal code, job title, event date, or community;
- **Sensitive attributes:** Information that could cause harm, stigma, unfair treatment or distress if linked to a person, such as health information, immigration status, criminal history, financial hardship or family violence information.

Mosaic effect – occurs when separate data points are combined through data matching to create a profile that can identify an individual.

💡 POPA'S DEFINITION OF PERSONAL INFORMATION

POPA defines personal information broadly. It includes an individual's:

- name or contact info;
- race, national or ethnic origin, colour or religious or political beliefs or associations;
- age, gender identity, sex, sexual orientation, marital status or family status;
- identifying number, symbol or other particular;
- fingerprints, other biometric information, blood type, genetic information or inheritable characteristics;
- health and health care history, including information about physical or mental health;
- educational, financial, employment or criminal history, including criminal records where a pardon has been given;
- personal views or opinions, except if they are about someone else; It also includes anyone else's opinions about the individual.

⚠️ BUSINESS CONTACT INFORMATION EXCEPTION

Contact information provided in someone's capacity as an employee or agent is excluded from the POPA's definition of PI, when limited to only contact information in a business context.

1.2. WHAT IS NON-PERSONAL DATA (NPD)?

NPD is data, including data derived from personal information (**DDFPI**), that has been generated, modified or anonymized (de-identified) so that it does not identify any individual. It includes synthetic data.

Synthetic data is artificial data created to maintain the structure and patterns of real data without being linked to any individual in the original data set.

> **✏️ IN PLAIN TERMS**
>
> Take a dataset containing personal information and match it with another dataset to create DDFPI. Aggregate it, generalize it, suppress small groups, and remove identifiers until the risk of figuring out who anyone is becomes low. This is one example of a simplified approach towards creating NPD.

1.3. WHAT IS DATA DERIVED FROM PERSONAL INFORMATION (DDFPI)?

DDFPI is data created by data matching (linking personal information between two or more datasets or other electronic sources of information), and that identifies any individual whose PI was used in the data matching.

The City may only keep DDFPI while it is needed for the approved purpose it was created for under The City's records retention rules. Once the purpose has been fulfilled, it must be destroyed or transformed into NPD.[1]

> **⚠️ ONE COMMON PATH FROM PI → DDFPI → NPD**
>
> A typical DDFPI workflow looks like this: start with PI → match or combine it with another dataset (now DDFPI) → de-identify the data using accepted de-identification methods to create NPD. Each stage has its own rules. For NPD creation, you are required to document the final stage (the creation of the NPD) using the *Record of Creation of NPD form*.

[1] *If you are unsure whether a retention schedule, legal hold, or privacy requirement applies, contact the Privacy Office or Records Management before deleting or transforming the data.*

Example: Data with Personal Information, to Data Derived From Personal Information, to NPD

Dataset 1

Name	Recreation program
John Doe	Swimming
Jane Smith	Gymnastics
John Smith	Swimming
Mary Doe	Swimming

PI

Dataset 2

Name	Parking Pass
John Doe	1289
Jane Smith	6789
John Smith	4589

PI

Data matching: merge datasets on Name field – DDFPI created

Name	Parking Pass	Recreation Program
John Doe	1289	Swimming
Jane Smith	6789	Gymnastics
John Smith	4589	Swimming
Mary Doe	None	Swimming

DDFPI

De-identify – pseudonymization, partial masking

Name	Parking Pass	Recreation Program
Customer 1	##89	Swimming
Customer 2	##89	Gymnastics
Customer 3	##89	Swimming
Customer 4	##NE	Swimming

NPD

De-identify – suppression and aggregation

Total Parking Passes	Recreation Program
2	Swimming
1	Gymnastics

New Insights

Lower Risk NPD

1.4. WHY DO WE DE-IDENTIFY DATA?

- To reduce the risk of identifying individuals;
- To protect privacy while still enabling analysis and decision-making;
- To allow broader internal or external sharing where appropriate; and
- To enable broader analysis, new insights, and additional uses of data while protecting privacy.

Under POPA, a data quality assurance and process assessment in creating NPD must be documented. We have created the *Record of Creation of NPD form* for this purpose.

> 💡 **WHY THE NPD PROCESS MATTERS**
>
> POPA allows NPD to be **used and shared more broadly** than personal information or data derived from personal information. In exchange for that flexibility, public bodies must follow specific requirements, including maintaining records of NPD creation, conducting quality assurance activities, and assessing re-identification risk

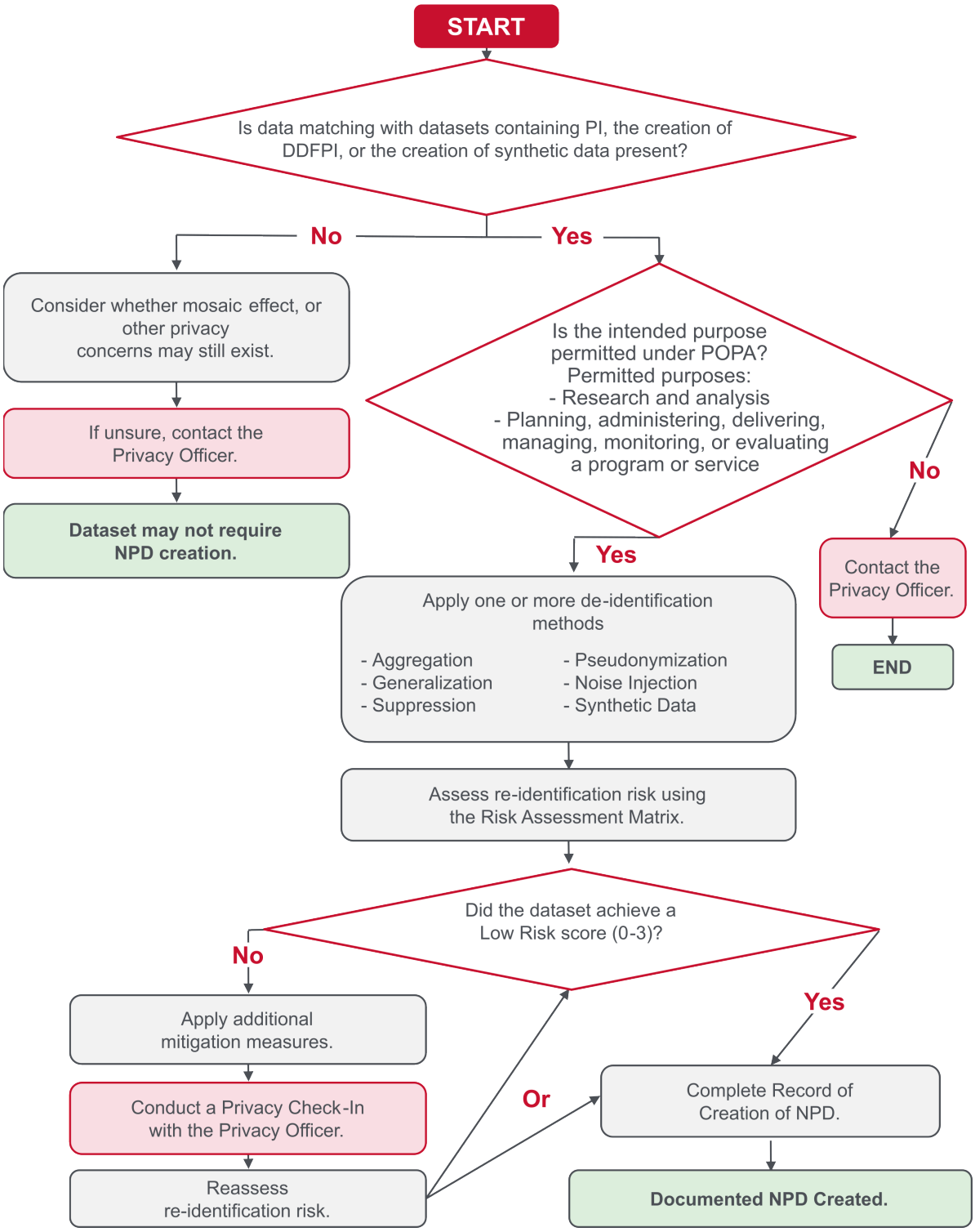
1.5 WHAT IS A PRIVACY CHECK-IN?

Sometimes a dataset doesn't score as "low risk" on the matrix alone, therefore, engagement with the Privacy Officer must be conducted to identify and implement additional risk mitigation in moving the dataset towards NPD. That's a privacy check-in.

> 🖍️ **PRIVACY CHECK-IN EXAMPLE:**
>
> The City wants to create an NPD dataset to evaluate whether translated communications are reaching residents who use languages other than English. The source data may combine service, communications, language, geography, and response information that contained personal information and was originally gathered for the evaluation of different services. Some language groups may have small counts, which can increase re-identification risk. A privacy check-in can help decide whether targeted controls - such as broader geography, larger reporting periods, grouped service categories, internal-only access, or limited suppression - can reduce the risk to low while keeping the data as useful as possible for service evaluation.

Additional risk mitigations for high-risk of re-identification data must be identified in collaboration with the Privacy Officer and must be documented.

Creating Non-Personal Data (NPD): High-Level Decision Process



NPD Record of Creation Flow

A flowchart for documenting dataset ownership, purpose, transformation, and review.

H — **HEADER FIELDS**

Who owns the dataset and when was the NPD created or refreshed?

↓

Q1 — **Q1 — DESCRIBE THE DATASET**

What is the dataset about, and what keywords help identify it later?

↓

Q2 — **Q2 — CONFIRM THE PERMITTED PURPOSE**

Why is the NPD being created?

↓

Q3 — **Q3 — IDENTIFY THE SHARING CONTEXT**

Will it be internal, external, or public body?

↓

Q4 — **Q4 — IDENTIFY SOURCE DATA ELEMENTS**

What categories of identifiers and risk-relevant elements were present?

↓

Q5 — **Q5 — EXPLAIN WHAT PI EXISTED BEFORE TRANSFORMATION**

What source data elements mattered for re-identification risk?

↓

Q6 — **Q6 — DESCRIBE THE TRANSFORMATION METHODS**

What methods were used to create the NPD?

↓

Q7 — **Q7 — ASSESS RE-IDENTIFICATION RISK**

What score was reached, and was Privacy involved?

↓

F — **FOOTER — RECERTIFICATION AND CHECK-IN**

When should the dataset be reviewed again, and was a privacy check-in needed?

2. WALKING THROUGH THE FORM

HEADER FIELDS

📝 WHAT TO INCLUDE:

- **Department/Business Unit:** the department and business unit responsible for creating and managing the dataset.
- **Division:** the division within the department where the dataset is owned or managed.
- **Contact Person:** the person who can answer questions about the dataset, and their position.
- **Contact Email:** a monitored email address that the Privacy Officer and other relevant internal staff can use to follow up about the dataset or record.
- **Dataset and/or System Name (Optional):** The name of this dataset, and the system that it is saved in, or hosted on.
- **Date of NPD Creation or Refresh:** the date the NPD was created or it was refreshed. The system automatically logs the date this form was filled in, so no need to fill that in.

QUESTION 1 — DATASET DESCRIPTION AND KEYWORDS

What is this dataset about? What keywords would you use to describe this dataset? Use complete sentences and include keywords that might be helpful when searching for this record.

📝 A GOOD Q1 ANSWER

"This dataset supports the summer sports program for youth planning by showing monthly participation totals by age ranges, and program category. Keywords include youth, recreation, sports, registrations, participation, program category, age ranges, and monthly summary."

⚠️ WATCH OUT FOR

- Listing only a system name (e.g. "SharePoint" or "SQL database").
- Pasting a full schema or column list instead of describing the data at a high level.
- Writing "no personal data" or not answering the question.

QUESTION 2 — PURPOSE

POPA only allows NPD to be created for two purposes:

- Research and analysis; and
- Planning, administering, delivering, managing, monitoring, or evaluating a program or service.

WHAT THESE PURPOSES MEAN:

The Government of Alberta has published the following definitions for the terms used in these two purposes.

I - RESEARCH AND ANALYSIS

- **Research:** means systematic investigation and analysis or study of materials or sources in order to establish facts or to verify theories.
- **Analysis:** refers to the process of examining and interpreting collected information to identify patterns, relationships, and trends. It often involves breaking data down into smaller parts, comparing findings, and drawing conclusions.

The creation of NPD for research and analysis is to ensure public bodies can modify, anonymize, or generate NPD to make accurate and informed decisions without putting individuals' personal privacy at risk.

II - PLANNING, ADMINISTERING, DELIVERING, MANAGING, MONITORING OR EVALUATING A PROGRAM OR SERVICE

- **Planning:** means to think about and decide what to do or how to do something.
- **Administering:** means to control the operation or arrangement of something.
- **Delivering:** means to provide a service.
- **Managing:** means to be responsible for controlling or organizing something.
- **Monitoring:** means to watch and check something carefully over a period of time.
- **Evaluating:** means to judge or calculate the quality, importance, amount, or value of something.

If your purpose does not clearly fall within these options:

Stop and contact the Privacy Office at **AccessandPrivacy@calgary.ca**.

QUESTION 3 — DATA-SHARING CONTEXT

How will this data be shared? Select every option that applies:

☐ Internal – within the City of Calgary only
☐ External – other public bodies
☐ External – non-public bodies (e.g. vendors, researchers)
☐ Public sharing

> 💡 **REMEMBER**
>
> The sharing context directly affects re-identification risk. The same dataset can be safe in one context and higher risk in another.

> ⚠️ **SHARING DATA**
>
> This guide does not cover sharing specifics under POPA. Please contact the Privacy Office for guidance when considering sharing PI, DDFPI, or NPD.

QUESTION 4 — WHAT WAS IN THE SOURCE DATA?

What categories of identifiers and associated risk-relevant elements were present in the source data, requiring consideration or transformation?

You are describing what types of data were present in the source data, not how risky you think the dataset was to share. (You'll score risk in Q7.)

PART A: CATEGORIES OF IDENTIFIERS

Select every category that was present in the source data.

- **Direct identifiers:** data that could identify someone on its own (name, email, employee ID);
- **Indirect (quasi-) identifiers:** data that could identify someone when combined with other data or context (postal code, age, and gender);
- **Sensitive attributes:** data that could cause harm or stigma if disclosed, even if you can't identify the person directly (health conditions, immigration status).

> 💡 **SOME ATTRIBUTES FALL INTO MORE THAN ONE CATEGORY**
>
> Ethnic identity is sensitive because of its association with real-world discrimination risk and can also be a strong indirect identifier in small populations. Select all relevant categories when more than one apply.

PART B: ASSOCIATED PRIVACY RISK ELEMENTS

Select every group that was present in the source data. These categories help reviewers quickly understand what types of data were involved. If something doesn't fit, select "Other" and describe it.

Detailed examples for each category are in the appendix, but common categories for City of Calgary records include:

- **Identity and contact information:** names, addresses, phone numbers, emails;
- **Demographic or personal characteristics:** age, sex, gender, marital status;
- **Geographic or location-based information:** postal codes, wards, neighbourhoods;
- **Dates or time-based information:** event dates, dates of birth;
- **Behavioural or transactional data:** service usage, registrations, complaints.

Question 4 identifies what types of data were present, while Question 5 explains how those elements contributed to re-identification risk.

QUESTION 5 — WHAT PERSONAL INFORMATION EXISTED BEFORE TRANSFORMATION?

Focus on what existed in the source data and why it mattered for re-identification risk. Do not describe how the data was transformed – that is Question 6.

🖍️ A GOOD FREE-TEXT EXPLANATION

"Source data included participant age (in single years), full address, and event dates.

These attributes functioned as direct and indirect identifiers, particularly when combined (e.g., age, location, and event timing).

Risk was primarily related to potential re-identification through these identifiers in combinations with small groups of records, although no direct identifiers or sensitive attributes were present."

⚠️ A POOR FREE-TEXT EXPLANATION

"We found some personal information that was risky. All personal information was removed."

WHY THIS FAILS: it doesn't say what specifically existed in the source data, or what made you think there was re-identification risk. Reviewers need this information.

QUESTION 6 — METHODS USED: HOW WAS THE DATA TRANSFORMED?

Describe the methods you used to create the NPD. Most NPD records rely on a combination of aggregation, generalization and suppression. *More detail on each methodology is in Section 4 below.*

AGGREGATION / GENERALIZATION

You reduced detail by grouping records or broadening values.

- Exact dates → months;
- Exact ages → age bands (e.g., 20–24);
- Addresses → wards or neighbourhoods;
- Counts reported by neighbourhood totals instead of by person.

> 💡 **WHEN TO USE AGGREGATION / GENERALIZATION**
>
> When identifiers are too precise or groups of values are too small.

SUPPRESSION (REMOVING CELLS, COLUMNS OR ROWS)

You removed values or records that would otherwise create uniqueness or small counts.

- Removing direct identifiers entirely;
- Removing rare categories or rows that represent unique cases;
- Suppressing cells, rows, or records where the count of these groupings is small (often <5 or <10).

> 💡 **WHEN TO USE SUPPRESSION**
>
> When aggregation alone still leaves unique or identifiable cases.

OTHER TECHNIQUES YOU MIGHT USE

- **PSEUDONYMIZATION:** replacing identifiers with codes or hashes;
- **NOISE INJECTION:** small random changes to prevent exact matching (like adding plus or minus a few days to an exact date, still useful for planning broadly, and better for de-identification);
- **SYNTHETIC DATA:** generated records that mimic patterns without mapping to real people, when generated correctly. Care must be taken in anonymizing synthetic data during creation.

QUESTION 7 — RE-IDENTIFICATION RISK ASSESSMENT

This is where you score the dataset. Calculate the final score using the Risk Matrix (Section 3 of this guide), then record it on the form.

Select which assessment method(s) you used, fill in the score, indicate if you reached out for assistance to the Privacy Officer before filling in this form, and optionally say if you used a method other than The City methods contained in this guide for evaluating your re-identification risk (high profile, high sensitivity datasets might need extra care). When in doubt, reach out.

RECERTIFICATION, REVIEW, AND CHECK-IN

- **RECERTIFY:** when the dataset structure, linkage context, or release model changes, or every three years, whichever comes first;
- **PROVISIONAL NEXT RECERTIFICATION DATE:** date planned by which to recertify this dataset again;
- **PRIVACY OFFICER REVIEW:** optional unless required by the release context (e.g., release of borderline dataset as NPD);
- **PRIVACY CHECK-IN:** applies when the score is above 6, where engagement and risk mitigation discussions with the Privacy Officer must be held.

3. UNDERSTANDING RE-IDENTIFICATION RISK AND SCORING

This section explains the Risk Assessment Matrix, the City's structured assessment tool in creating NPD. The matrix consists of five questions, scored 0–2, and then adds up to a final score. Perspective here is of the average person with average means at their disposal.

THE 5 RISK FACTORS

1. IDENTIFIABILITY

WHAT IT ASKS: How easily could a person be singled out? By unique combinations of attributes with a small number or records? Or by direct identifiers?

EXAMPLES: in a dataset of five people in a neighbourhood, even "age + gender" might single someone out. Larger groups (20+) are safer.

2. SENSITIVE ATTRIBUTES

WHAT IT ASKS: Would the data reveal something that could cause harm or distress if a person were re-identified?

EXAMPLES: health conditions, immigration status, criminal history, financial vulnerability, family violence.

3. LINKAGE RISK

WHAT IT ASKS: Could this dataset realistically be combined with other available data to single people out?

EXAMPLES: if your dataset has neighbourhood plus birth year, and another dataset linked in the same dashboard has the same fields plus name, those can be joined. Linkage requires a plausible pathway, not just "some other data exists."

4. DATA GRANULARITY

WHAT IT ASKS: How detailed is the data, and how many possible values exist for each attribute?

EXAMPLES: "exact age, exact date, exact address" is very granular. "Age band, month, ward" is much less granular and far lower risk.

5. CONTEXT AND CONTROLS

WHAT IT ASKS: Who has access to this data, how is it shared, and what safeguards exist?

EXAMPLES: internal-only with role-based access = strong controls. External sharing with vendors that have broad access controls = some controls. Stored data with weak access controls = few controls.

THE RISK ASSESSMENT MATRIX — SCORE EACH FACTOR 0, 1, OR 2

The matrix is intended as a practical screening tool for most NPD creation activities. More complex datasets may require additional review.

FACTOR	0 (GOOD)	1 (OK)	2 (WARNING)
IDENTIFIABILITY	Large groups (20+)	Mid-size groups (10–20)	Small groups (<10)
SENSITIVE ATTRIBUTES	None	Mild	High sensitivity
LINKAGE RISK	No known way to join	Possible to join	Strong linkage ecosystem
DATA GRANULARITY	Aggregated	Moderately detailed	Highly detailed
CONTEXT & CONTROLS	Strong controls	Some controls	Broad access or few controls

Note: It is unnecessary to show there is zero risk of re-identification; however, the assessment results must show that the residual risks of re-identification are low. A "2" in any single category should be looked at with the intention to reduce the risk, if possible.

ADDING IT UP — THE SCORE BANDS

Total = sum of all five factors (0–10 possible).

0–3 LOW RISK → recommended risk level for NPD
Aggregated, low-sensitivity, low linkage potential, strong controls. Document the assessment and proceed.

4–6 BORDERLINE → apply more protections
Some risk factors are elevated. Reduce granularity, suppress small groups, restrict access, or revisit your context and controls, then re-score.

7–10 HIGH RISK → privacy check-in with the Privacy Officer
Re-design the dataset (de-identify, aggregate further, change the context or controls), work with the Privacy Officer on a documented privacy check-in.

4. HOW TO DE-IDENTIFY DATA — SIX METHODS

> 💡 **NEED HELP WITH DE-IDENTIFICATION?**
>
> For assistance with de-identification approaches, contact Deidentify@calgary.ca.

These are some of the practical techniques, with some examples, you might use to bring a dataset into the "low risk" band. Most NPD records use a combination of two or three de-identification methods.

4.1. PSEUDONYMIZATION

WHAT IT IS: Replacing direct identifiers (names, emails, phone numbers) with random codes, IDs, or hashes.

HOW TO DO IT:

- Replace names with generic terms;
- Replace parts of phone numbers with symbols;
- Hash email addresses (e.g., SHA-256) to produce unique but non-identifying values, if salted and governed properly, otherwise hashed values might be vulnerable to guessing or other attacks, which might lead to re-identification. Best used in combination with other methods to reduce reidentification risk.

> 🖍️ **EXAMPLE**
>
> "Jane Smith" becomes "Person A" or a long hash like "a92f9c1d…".

> ⚠️ **NOTE**
>
> Pseudonymization on its own may not be sufficient to create NPD with a low risk of re-identification. The codes can often be reversed or linked back to people. You may need to combine it with aggregation, suppression, or generalization.

4.2. AGGREGATION

WHAT IT IS: Grouping multiple records into summaries so individual-level records are no longer shown.

HOW TO DO IT:

- Person-level registrations → total registrations by program;
- Individual service requests → monthly counts by community;
- Trip records → average trip distance by ward.

> 🖍️ **EXAMPLE**
>
> Instead of raw trip distances per person in a community, publish "total daily trips per community."

4.3. SUPPRESSION

WHAT IT IS: Removing values (cells, rows, or columns) that are highly identifying, usually direct identifiers, outliers or rare cases.

HOW TO DO IT:

- Remove rows for extremely rare cases (e.g., the only 101-year-old in a dataset);
- Remove individual cells where the value could single someone out;
- Suppress small groups (counts under a threshold) in tables and dashboards.

> 📝 **EXAMPLE**
>
> If only one to five people in your dataset has a particular rare condition, remove those rows or roll the condition up into a broader category.

4.4. GENERALIZATION

WHAT IT IS: Making specific values less precise while keeping useful information.

HOW TO DO IT:

- Date of birth → year of birth → age band (e.g., 20–24);
- Full Postal code → first three characters (e.g., T2V) → community name;
- Exact event time → hour → time-of-day band.

> 📝 **Example**
>
> Instead of "1993-06-17," change it to "1993" or "age band 30–34."

4.5. NOISE INJECTION

WHAT IT IS: Randomly applying small changes to values to prevent exact matching but keep it close enough to the original data to be useful for analysis.

HOW TO DO IT:

- Add or subtract a random number of days, for example, between 1 and 7 days, to event dates;
- Randomly add or subtract between 1–5% to continuous values like sensor readings.

> 📝 **EXAMPLE**
>
> An event date of "2023-05-18" might be stored as "2023-05-14." You randomly subtracted or added between 1 and 7 days, which ended up being minus four days for this record. These random changes need to be applied to all values in that field or column.

> ⚠️ **NOTE**
>
> Use carefully. Too much noise damages the data's usefulness; too little doesn't protect privacy.

4.6. SYNTHETIC DATA

WHAT IT IS: it uses real world data to derive a model, such as a simulation or artificial intelligence (**AI**) model, which is then used to create an artificially generated but statistically equivalent dataset.

HOW TO DO IT:

- Train a model on the real dataset;
- Generate new records that share the same statistical structure of real data without containing any information from real-world sources;
- When it involves the use of AI or Large Language Models (LLMs), engage the Privacy Officer by completing a Privacy Risk Questionnaire (PRQ).

> **EXAMPLE**
>
> Useful for analytical dashboards, prototyping machine learning (ML) models, or sharing data externally for development purposes.

> **NOTE**
>
> Synthetic data still requires careful evaluation. It can leak information about real individuals if not generated correctly.

5. WORKED EXAMPLES

Below are six examples spanning the score range to calibrate your own assessments. The scores are illustrative and your own dataset's score will depend on its actual structure, contents, and context

LEGEND:

I = Identifiability	**S** = Sensitive Attributes	**L** = Linkage Risk	**G** = Granularity	**C** = Context & Controls

EXAMPLE 1: AGGREGATED 311 SERVICE REQUESTS

DATASET: Aggregated monthly counts of 311 service requests by community and issue category. Some groups of records between 10-20, the rest are over 20.

RELEASE MODEL: External, non-public body.

SCORING: I=1, S=0, L=0, G=0, C=1 → **TOTAL** = **2 (Low Risk)**

> ✅ **BAND: LOW RISK**
>
> **WHY:** No direct identifiers. Mid-size groups, no sensitive content, fully aggregated, and no plausible way to identify individuals even by combining with other sources.

EXAMPLE 2: AGGREGATED 311 DATA WITH SMALL ROW COUNTS OF SPECIFIC GROUPS

DATASET: Same as above, but a rare subcategory occasionally produces counts of less than five in some communities.

RELEASE MODEL: External, non-public body.

SCORING: I=2, S=0, L=0, G=1, C=1 → **TOTAL** = **4 (Borderline)**

> ⚠️ **BAND: BORDERLINE**
>
> **WHY:** Small groups can enable singling out an individual. Granularity of data is less aggregated because of the smaller grouping.
>
> **WHAT WOULD MAKE IT LESS IDENTIFIABLE:** Suppress small counts and roll up rare subcategories or sparsely populated communities until groups are consistently large. Engagement with Privacy Officer if purpose necessitates release at an elevated risk level, ensuring context and controls or linkage risk has not increased either.

EXAMPLE 3: INTERNAL DASHBOARD FOR RECREATION PROGRAM MONITORING

DATASET: Weekly totals by ward and age band. No person-level rows.

RELEASE MODEL: Internal only

SCORING: I=0, S=0, L=1, G=0, C=0 → **TOTAL = 1 (Low Risk)**

✅ **BAND: LOW RISK**

WHY: Aggregated, no sensitivity, strong internal controls, no meaningful way to single anyone out. Possible linkage to other datasets, but nothing obvious or deliberate.

EXAMPLE 4: EXTERNAL EVENT DATA RESEARCH SHARING UNDER CONTRACT

DATASET: De-identified event data with age bands, broad geography, and event date generalized to month. Some moderately aggregated data. No direct identifiers. Some record groupings between 10 and 20. Linkage possible with other datasets shared with external party under the same contract, but not likely.

RELEASE MODEL: External, non-public body (research) with controls

SCORING: I=1, S=0, L=1, G=1, C=1 → **TOTAL = 4 (Borderline)**

⚠️ **BAND: BORDERLINE**

WHY: Contractual terms help, but row-level data with moderate detail can still enable linkage.

WHAT WOULD MAKE IT LESS IDENTIFIABLE: Reduce granularity further (quarter instead of month, broader geography), increase group sizes, suppress rare combinations. Engagement with Privacy Officer if purpose necessitates release at an elevated risk level.

EXAMPLE 5: PERMIT OR SERVICE DATA WITH ADDRESSES

DATASET: Dataset listing permit or service activity, including exact address or parcel identifier.

RELEASE MODEL: External – public body

SCORING: I=2, S=0, L=2, G=2, C=1 → **TOTAL = 7 (High Risk)**

🛑 **BAND: HIGH RISK - PRIVACY CHECK-IN**

WHY: Exact address effectively identifies a household and enables easy linkage with other exiting sources available. Not aggregated. External public body has some controls in place.

WHAT WOULD MAKE IT LESS IDENTIFIABLE: Remove addresses, generalize to community or ward, publish aggregated totals by permit or service type only.

APPENDIX — REFERENCE DATA DICTIONARY

Illustrative examples of common data attributes and how they're typically considered in re-identification risk assessments. Use this appendix to support a common understanding during privacy review. Adapt the examples to the specific dataset and context. A common mitigation strategy would be to use broader categories, combine small groups, or suppress small counts where needed.

> ⚠ **THIS IS REFERENCE MATERIAL, NOT A POLICY**
>
> Final determinations of re-identification risk are made using the Risk Assessment Matrix; and, where necessary, in collaboration with the Privacy Officer.

IDENTITY AND CONTACT INFORMATION

ATTRIBUTE	DESCRIPTION	TYPICAL ROLE	COMMON MITIGATION
NAME	Individual's full or partial name	Direct identifier	Remove or replace with non-identifying code
EXACT ADDRESS	Precise residential or mailing address	Direct identifier	Remove or generalize to a larger area
PHONE OR EMAIL	Personal contact information	Direct identifier	Remove or replace with non-identifying code

DEMOGRAPHIC OR PERSONAL CHARACTERISTICS

ATTRIBUTE	DESCRIPTION	TYPICAL ROLE	COMMON MITIGATION
AGE	Age in single years or age groups	Indirect identifier	Group into age bands, suppress small groups
GENDER	Gender identity	Sensitive attribute	Broaden categories; suppress small groups; careful contextual review
MARITAL STATUS	Legal or relationship status	Indirect identifier	Broader categories

SOCIODEMOGRAPHIC OR POPULATION ATTRIBUTES

ATTRIBUTE	DESCRIPTION	TYPICAL ROLE	COMMON MITIGATION
ETHNIC IDENTITY	Self-reported ethnic identity	Sensitive + indirect	Generalize categories; suppress small groups
RACE / VISIBLE MINORITY	Racialized population groups	Sensitive attribute	Use broader categories; suppress small groups; review context carefully
IMMIGRATION STATUS	Immigrant, non-permanent resident, or Canadian-born	Indirect + sensitive	Use broad categories; suppress small groups
LANGUAGE AT HOME	Most-spoken language at home	Indirect identifier	Group languages

GEOGRAPHIC OR LOCATION-BASED INFORMATION

ATTRIBUTE	DESCRIPTION	TYPICAL ROLE	COMMON MITIGATION
PROVINCE / TERRITORY	Broad geographic region	Indirect identifier	None typically required
CENSUS DIVISION	Regional administrative area	Indirect identifier	Roll up to broader region
SMALL CENSUS OR SURVEY AREA	Small, relatively stable geographic unit	Indirect identifier	Combine into larger areas or hide small counts
EXACT COORDINATES	Precise location or address point	Direct or strong indirect identifier	Remove or generalize to larger areas

DATES OR TIME-BASED INFORMATION

ATTRIBUTE	DESCRIPTION	TYPICAL ROLE	COMMON MITIGATION
REFERENCE YEAR	Calendar or reporting year	Low-risk indirect	None typically required
EXACT EVENT DATE	Full calendar date	Indirect identifier	Convert to month or quarter
REPEATED TIMING PATTERNS	Longitudinal or recurring events	Indirect identifier	Generalize to broader time windows

EMPLOYMENT OR ORGANIZATIONAL ATTRIBUTES

ATTRIBUTE	DESCRIPTION	TYPICAL ROLE	COMMON MITIGATION
OCCUPATION	Job or role	Indirect identifier	Broaden role categories
INDUSTRY	Sector of employment	Indirect identifier	Broaden industry categories
EMPLOYMENT HISTORY	Employment changes over time	Indirect + sensitive	High-level summaries; remove unique career paths

BEHAVIOURAL OR TRANSACTIONAL DATA

ATTRIBUTE	DESCRIPTION	TYPICAL ROLE	COMMON MITIGATION
SERVICE USAGE	Participation in programs or services	Indirect identifier	Aggregate counts; suppress small groups
INTERACTION FREQUENCY	Number or rate of interactions	Indirect identifier	Time or volume bands; suppress outliers
COMPLAINTS/ INVESTIGATIONS	Formal complaints or escalations	Sensitive attribute	Aggregate or suppress; remove case details

SMALL POPULATION OR UNIQUENESS INDICATORS

ATTRIBUTE	DESCRIPTION	TYPICAL ROLE	COMMON MITIGATION
RARE CATEGORIES	Low-frequency values	Identifiability driver	Combine with broader categories or suppress
SMALL COUNTS	Counts under an accepted threshold	Identifiability driver	Aggregate or suppress
UNIQUE	"Only one" or extreme outliers	Identifiability driver	Remove, combine or mask outliers

NEED HELP?

> 💡 **PRIVACY OFFICE | Email**: AccessandPrivacy@calgary.ca
>
> 💡 **HELP WITH DE-IDENTIFICATION | Email**: Deidentify@calgary.ca

*This guide should be used in conjunction with the Record of Creation of Non-Personal Data (NPD) form. **The form is the official record**; this guide explains how to fill it out, what the terms mean, and how to assess and reduce re-identification risk.*