CANON Canadian Anonymization Network

Frequently Asked Questions

(Last Updated: January 15, 2020)

The Canadian Anonymization Network (CANON) (described further below) has created these FAQs to help address some of the questions that individuals, organizations, governments and the media have expressed in relation to de-identification. CANON believes that de-identification can play a key role in the use and sharing of data for economic- and socially-beneficial purposes, but recognizes that a key factor in enabling this will be to educate the community at large about the effectiveness of de-identification methods, and to meaningfully contribute to discussions around both risks and opportunities.

This is intended to be a living document, and new information and/or clarifications will be added as appropriate.

Q: What is de-identification?

A: De-identification means different things to different people. Recognizing that terminology can vary, particularly between terms such as de-identification and anonymization, CANON is in the process of developing a lexicon to clarify these and other terms relative to one another, and will update these FAQs accordingly. However, for the present purposes, de-identification is understood broadly as a process that transforms personal information into data for which there is no serious risk of re-identifying an individual given the context in which it will be processed.

This generally includes, but also goes beyond, the process of removing and/or modifying of both “direct identifiers” (attributes that alone enable unique identification of an individual, such as name, address, or unique numeric identifiers) and “indirect identifiers” (attributes that, when combined with other data, enable unique identification of an individual).

Where identifiers are modified, rather than deleted entirely (because of the importance of maintaining the informational value of the dataset), various masking techniques may be used including, but not limited to, pseudonymization (e.g. replacing identifiers with codes), randomization (e.g. modifying attributes such that their new value differs from their true value in a random way) or aggregation (e.g. grouping values into ranges).

In addition to transforming the data, de-identification techniques take into account contextual considerations which impact the risk of re-identification of a dataset. These contextual factors are further discussed below.

Complementing de-identification are other emerging data transformation techniques, such as the creation of synthetic data (a fictitious dataset which mimics the patterns and qualities of the original).

***

Q: What are the benefits of sound de-identification practices?

A: Data is increasingly recognized as a key driver for innovation. As stated by *Innovation, Science, and Economic Development Canada*, "Digital and data-driven technology is already empowering science, supporting innovation, and driving economic growth." The availability of data is critical to the health researcher working to prevent or cure disease, the government agency seeking to better serve its citizens, and the company creating innovative products and services which create social value and drive the Canadian economy.

These and other benefits of data can be derived when organizations are able to make fuller use of the data in their custody and control, and even more so when data can be combined and made more widely available for use by multiple parties. For example, the extent of these benefits is such that in the health sector there has been a recognition that the duty to share information can be as important as the duty to protect patient confidentiality (the Caldicott Principles, Principle #7).

Of course, in order to leverage these benefits, individuals must be able to trust that their data are being used responsibly; just as the increased availability of data is critical, so too must the protection of individuals' privacy underpin a robust digital ecosystem.

CANON believes that de-identification provides promising opportunity to achieve both data availability and privacy protection, allowing for a privacy-respectful means of responsibly leveraging data for economic and socially beneficial purposes.

***

Q: What are some of the typical use cases for de-identification?

A: De-identification supports a broad range of use cases for data, including:

- **Public release**: An organization de-identifies data before releasing it publicly.
- **Disclosure to a third party**: Data is put through a de-identification process before being provided to a third party, subject to a confidentiality agreement.
- **On-site access by a third-party**: Data is put through a de-identification process and then made available by the original data holder to a third party either on-site or through a secure, audited portal – subject to a confidentiality agreement.
- **Trusted Agent**: Data from multiple organizations are put through a de-identification process and provided to a trusted agent (such as a data trust), which either analyzes, aggregates, or pools the data on behalf of the third party or makes it available to other parties (including the original organizations).
- **Internal**: An organization puts data about its customers through a de-identification process and makes it available to internal research and development staff, separated from its business lines behind a physical, virtual and administrative firewall.

Organizations that opt to de-identify data – rather than, for instance, deleting it – have made the decision that (i) the data in question has value, and (ii) that value can be realized without the inclusion of personal information.

CANON intends to develop a document elaborating on the above and other use cases which focus on the context and mechanisms by which data are made available for use in a privacy-protective manner, as well as the benefits that they enable, from the training of machine learning models to the development of insights about populations.

***

Q: If de-identification must be considered in context, which contextual factors should be taken into account when evaluating the risk of re-identification?

A: Saying that a dataset is "de-identified" is equivalent to saying that it has no serious risk of being re-identified. This is in part a function of: (i) the susceptibility of the data to re-identification (e.g. the level of specificity, or the existence of other data sets that can be linked or joined to it); (ii) the transformation applied to the data; and, (iii) the context in which the data is shared and used.

Thus, in addition to factors such as the levels of cost, effort, and expertise required to re-identify a dataset, contextual factors which can impact the risk of re-identification include:

- Release environment and data accessibility (e.g. public release; release to only limited parties; restricted access through a secure environment; etc.);
- Re-identification restrictions and how they are enforced (e.g. contractual terms between parties prohibiting re-identification, supported by significant consequences if this prohibition is violated; technical and administrative separation of datasets, particularly if an organization holds both the de-identified and original source data; etc.);
- Nature and capacity of the data recipient (e.g. known and trusted actors with professional and/or legal obligations of confidentiality; organizations that are certified to known privacy or security standards, or that use only similarly certified data processors; etc.);
- Incentives to re-identify the data (e.g. data of high intrinsic value or high sensitivity; data for which re-identification would bring an attacker acclaim and/or notoriety; etc.); and,
- Security protections around de-identified data (e.g. weak protections can lead to a data breach, changing each of the prior contextual factors)

CANON intends to further elaborate on these contextual factors – as well as the various techniques which measure re-identification risk based on both the dataset itself as well as its context.

***

Q: What is the legal context for de-identification in Canada?

A: Canadian privacy law only applies to "personal information"; information which has been effectively de-identified is not considered personal information and falls outside the scope of these laws.

The specific legal thresholds of "personal information" tend to vary across jurisdictions. For example, under PIPEDA – Canada's federal private sector privacy law – personal information is defined as "information about an identifiable individual." Canadian courts have stated that this must be given a broad and expansive interpretation, and have found that information will be about an identifiable individual "where there is a serious possibility that an individual could be identified through the use of that information, alone or in combination with other information." Notably, the "serious possibility" test

recognizes that the risk of re-identification does not need to be *completely* eliminated for information to be considered non-personal.

Other legal thresholds for identifiable information include "information that identifies an individual or for which it is reasonably foreseeable in the circumstances could be utilized, either alone or with other information, to identify an individual" and information from which "the identity of the individual who is the subject of the information can be readily ascertained."

The binary nature of personal information can create challenges for public-, private-, and health-sector organizations seeking to use data for innovative purposes while simultaneously protecting and enhancing personal privacy. Given efforts in many jurisdictions to modernize privacy regulations, this situation may change in the near future. Among the changes being contemplated, legislators are currently consulting on concepts of identifiability and de-identification, contemplating including clearer, more explicit definitions, and possibly introducing whole new categories of data along a broader spectrum of identifiability (example, "pseudonymized data").

CANON has submitted its preliminary views on some of these questions. For CANON's letter to Innovation, Science and Economic Development (ISED) Canada about de-identification related considerations in ISED's proposal to modernize PIPEDA is available here.

Current definitions of Personal Information

Public Sector

Federal	BC	AB	SK	MB	ON	QC	NS	PEI	NB	N&L	YK	NWT	Nun
s.3	Sch.1	s 1(n)	s.24(1)	Defns	s.2(1)	s.54	s. 3	s. 1	s. 1	s. 2	s. 3		s. 2

Private Sector

Federal	AB	BC	Quebec
s. 2(1)	s. 1(1)(k)	s. 1	s. 2

Health Sector

BC	AB	SK	MB	ON	QC	NS	PEI	NB	N&L	YK	NWT	Nun
NA	s.1(1)(k)	s. 2(m)	s. 1(1)	s. 4(1)	NA	s.3(r)	s. 1(t)	s. 1	s.2(1)(p)	s.2(1)	s. 1(1)	s.2

***

Q: I've heard of instances in which individuals have been re-identified from supposedly de-identified datasets. Does this mean that de-identification is ineffective?

A: In short, no, it does not mean that de-identification is ineffective. If anything, these instances show that de-identification is an evolving and dynamic field which will benefit from greater consistency in both terminology and process – both of which CANON intends to support.

Broadly speaking, studies that claim to have successfully re-identified individuals in a "de-identified" or "anonymous" dataset tend to fall into one of two categories: 1) re-identification of datasets which should not have been described as "de-identified" in the first place, or 2) exposure that a particular de-identification technique is prone to an attack. Neither of these types of findings bring into question the effectiveness of the field of de-identification as a whole.

There are many instances of the first category of study; for example, researcher (and CANON member) Khaled El Emam et al. undertook a systemic review of re-identification attacks in the health sector, finding that in 12 of 14 instances that the target dataset was not de-identified to then-current standards.

In the second category, consider by way of example a 2019 study by Luc Rocher et al., published in *Nature Communications*. For their study, the researchers developed a model which established that, if an individual can be matched to a record based on 15 demographic attributes, there is a 99.98% chance that the record belongs to that individual – even if the dataset is "heavily sampled" (i.e. even if there is a low probability that the individual's record is in the dataset). This is an important finding which suggests that the risk of re-identification is higher than previously recognized for publicly-released datasets which depend on sampling as a key part of their de-identification process.

Studies such as these, which probe the effectiveness of de-identification techniques and/or question assertions made about the re-identifiability of particular datasets are important and CANON fully supports their continued contribution to the field. They highlight the importance of the proper application of de-identification protocols, of a consistent understanding by all parties of what is meant by "de-identification" and related terms, and of researchers examining claims made about the de-identified nature of a dataset. They also expose previously unknown flaws in and/or attacks against de-identification techniques. Organizations which rely on de-identification should maintain an awareness of such developments, and adjust their practices accordingly.

However – much like a study which exposes a potential exploit in, or flawed deployment of, a particular encryption method does not bring into question the effectiveness of the field of cryptography – in neither case is de-identification as a whole shown to be ineffective. When done effectively, and in proper context, CANON believes de-identification is an important privacy-enhancing method that should continue to be promoted and developed.

***

Q: Who is CANON?

A: CANON is a not-for-profit corporation that supports a network of some of Canada's largest data custodians across the public, private and health sectors. Its mission is to promote and enhance de-identification as an effective privacy-enhancing technology.

Among its objectives, CANON intends to develop a technology- and sector-neutral framework or Code of Practice that would, among other benefits:

- Increase the public's confidence and trust in de-identification as a means of allowing responsible data use and sharing for innovative economic and socially beneficial purposes;

- Enable organizations to effectively minimize re-identification risks while still preserving the utility of data; and,
- Allow for clarity and common understanding when:
    - Organizations describe their personal information management practices to individuals;
    - Organizations enter into contractual arrangements with processors based on an undertaking of de-identification; and,
    - Organizations seek to demonstrate to regulators that they are acting in an accountable manner and cite de-identification as a mitigating factor.

We believe this proactive initiative will help organizations across sectors that are seeking to innovate for economic and social prosperity and leverage their data for public good while increasing the overall level of privacy protection in Canada.

CANON and its members look forward to the opportunity to work with technical experts, regulators, civil society, and/or any other relevant group towards the creation of an accessible, operational resource that supports and enhances Canada's position as a leader in both privacy protection and innovation through the use of de-identification.

***