Public Education Fact Sheet: De-identification and Anonymization of Personal Information

Prepared by the Public Interest Advocacy Centre

6 October 2011

What is de-identification?

De-identification is the process of removing identifiers from personal information. Personal information as defined under Personal Information Protections and Electronic Documents Act (PIPEDA) is “information about an identifiable individual.” Personal information may be de-identified, for example, when an individual’s name, address, telephone number, social insurance number, or date of birth is removed, generalized, or replaced with a made-up alternative in a data set.

De-identification is a common term in the health industry, where an individual’s personal health information may be de-identified so that the de-identified data can be used for secondary health research purposes.

What is anonymization?

The term “anonymization” is sometimes used in reference to the application of de-identification processes to personal information. “Anonymization” implies that the individual is anonymous and cannot be re-identified. Principle 4.5.3 of PIPEDA (Limiting Use, Disclosure, and Retention) states: “Personal information that is no longer required to fulfil the identified purposes should be destroyed, erased, or made anonymous.” Thus, PIPEDA seems to suggest that organizations could anonymize their data as a good data retention practice.

What is the difference between de-identification and anonymization?

Often, the terms “de-identify” and “anonymize” are used interchangeably. However, anonymization implies that an individual’s personal information has been made anonymous and cannot be linked back to the individual under any circumstances. De-identification simply suggests that a process has been applied to personal information to attempt to make the information no longer "about an identifiable individual". De-identified information can reveal a

detailed profile, however, and de-identified data may be re-identified with an individual, depending on the effort and skill of a determined data processor (see "re-identification" below).

Why is de-identified or anonymized data useful to organizations?

De-identified or anonymized data may be aggregated into databases for secondary uses. In the health context, uses of de-identified and aggregated data include clinical program management, health system management, monitoring the health of the public, and secondary health research, all of which could lead to improved patient care and health outcomes. In the commercial context, companies may use aggregated de-identified data to improve services and find efficiencies or to perform data mining. Data mining may use de-identified data for tracking consumer behaviour, performing consumer analytics, for customer segmentation, market research, consumer profiling or to serve targeted advertisements.

What is re-identification?

Re-identification is the process through which de-identified or anonymized data is matched back to the individual. Data can be re-identified by combining de-identified data with publicly available information or another database in which individuals are identified. Some researchers refer to re-identification as “de-anonymization.”

What are some examples of re-identification?

Researchers have shown that de-identified and anonymized data can be re-identified, proving the assumption that anonymized data is anonymous flawed. For example, in 2006, AOL released “anonymized” search log data comprising 36 million search queries to the public. AOL suppressed identifying information such as username and IP address, yet journalists were able to link search queries back to identifiable individuals. Computer scientist Dr. Latanya Sweeney found that in the 1990 census, 87% of the American population could be uniquely identified by their combined ZIP code, date of birth and gender. More recent studies show how individuals can be linked across various online media such as social networks and blogs through their username, or how social network graphs can be analyzed to uniquely identify a user. Location data could also be used to identify an identifiable individual based on the disclosure of home and workplace locations.

What’s the privacy risk with de-identified or anonymized data?

As mentioned above, Principle 4.5.3 (Limiting Use, Disclosure, and Retention) suggests that organizations should destroy, erase or make anonymous personal information that is no longer required to fulfil the identified purposes. This retention principle suggests that when personal information is anonymized, the fair information practices required by PIPEDA no longer apply to such information.

Thus, organizations may consider de-identified or anonymized data to fall outside the scope of PIPEDA. Little notice or transparency may be provided to consumers regarding how the industry de-identifies or anonymizes the data, or the extent to which there is a risk that such anonymized data could be matched with publicly available data to re-identify individuals.

If the organization considers de-identified or anonymized data to fall outside PIPEDA, the organization may believe that it does not require the consumer's consent to use or disclose the de-identified or anonymized data, so it may use such data for purposes unknown to the consumer or share the data freely with third parties.

Many experts are now challenging the ability of personal information to be truly anonymized, suggesting that the challenge of anonymization is compounded with advances in information technology, the amount of publicly available information through online media such as social networking sites, and the burgeoning and profitable data mining industry.

What can I do about this?

Consumers can contact organizations to request clarity about their privacy policies and practices in accordance with Principles 4.1 (Accountability) and 4.8 (Openness) of PIPEDA. Individuals should inquire about what information the organization considers to be "personal information" and whether they destroy data after it is no longer needed or whether they anonymize the data. If the organization anonymizes the data, you may wish to seek assurances that the data is truly anonymized and ask what the purposes are for the use and disclosure of your anonymized data. If anonymized data is disclosed to third parties, you might clarify which third parties are receiving such information and whether there are conditions on these disclosures. If you are uncomfortable with the information provided, ask the organization whether you can opt-out of the use and disclosure of your de-identified or anonymized personal information. You are always free, as well, to withdraw your consent to the use or disclosure of your personal information in any way by the organization, subject to certain limits.

If you are dissatisfied with the transparency of practices provided by the organization, you may file a PIPEDA complaint with the Office with the Privacy Commissioner. For guidance, please see: http://www.priv.gc.ca/complaint/pi_e.cfm#contenttop.

More Information

- **Detailed technical information about de-identification processes:** Ross Fraser & Don Willison, "Tools for De-Identification of Personal Health Information" prepared for the Pan Canadian Health Information Privacy Group (September 2009), online: https://www2.infoway-inforoute.ca/Documents/Tools_for_De-identification_EN_FINAL.pdf. Health System Use Technical Advisory Committee, Data De-Identification Working Group, "'Best Practices' Guidelines for Managing the Disclosure of De-Identified Health Information" (October 2010), online: http://www.ehealthinformation.ca/documents/de-idguidelines.pdf.
- **Benefits that can be derived from de-identified data and a framework for risk assessment for re-identification:** Ann Cavoukian & Khaled El Emam, "Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy" (June 2011), online: http://www.ipc.on.ca/images/Resources/anonymization.pdf.
- **The failure of anonymization and ease of re-identification:** Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization" (2010) 57

UCLA L. Rev. 1701, online:
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006.