KRIS KLEIN'S PRIVACY SCAN WITH SHAUN BROWN AND MURRAY LONG

May 17, 2013

De-identification may be more complicated than you think

By John Wunderlich[1]

Any organization that has the custody and control of personally identifiable information (PII)[2] will regularly receive requests, either internally or externally, for access to that data. Any such 'secondary' use or disclosure of PII that a) does not fulfill a purpose identified with consent from the data subject, or b) is not based on legislative authority, is likely to be violation of one or more fair information practices, codified in Canada by the CSA *Model Code for the Protection of Personal Information* (Schedule 1 of PIPEDA[3]). However, if the requestor is given data that has been 'de-identified' so that it is no longer PII, then that data can be supplied to the requester with a reduced risk to privacy. The purpose of this article is to attempt to set out some guidelines that privacy professionals can use to ensure that their privacy program has the ability to manage the risk associated with de-identifying data for secondary uses.[4] We do not discuss the use or disclosure of PII, such as is the disclosure of personal health information to researchers, where the data is identifiable and different safeguards are used to ensure compliance with privacy requirements. Similarly, aggregate data with no 'small cells'[5] is not personally identifiable and is therefore privacy safe and out of scope of this article.

In the private sector, a marketing department might want access to a customer contact database to build a campaign for a new product. In the public sector, a department may want to link data they have with another department to reduce duplicated efforts and generate efficiencies. And in the health sector, researchers will want access to patient data from a health registry. Typically, the data requestor will frame their request as a 'need'. "I need this information to do my job" will be the refrain. From the point of view of the data custodians, then, these may be worthwhile requests that should be met if possible.

[1] John Wunderlich is an information privacy and security evangelist. Recent decisions by various regulatory authorities have made it clear that having a security or a privacy policy is insufficient to the task of mitigating risk. Because of his background in IT/Operations and Process Improvement, John brings a unique and practical perspective to organizations or individuals that are responsible for managing sensitive information.

[2] In Ontario, I find that people will distinguish between personal information (PI) as defined by the Freedom of Information and Protection of Privacy Act (FIPPA) and PHI as defined by the Personal Health Information Protection Act (PHIPA) for a variety of reasons. From a de-identification risk management perspective, using personally identifiable information has the advantage of focusing on identifiably and is the term used internationally in ISO/IEC standard.

[3] *Personal Information Protection and Electronic Documents Act*, SC 2000, c 5.

[4] The author would like to acknowledge the generous assistance of Dr. Khaled El Emam, who gave generously of his time to discuss the subject matter of this article. Individuals interested in a deeper dive into this subject should refer to Dr. El Emam's publications. Any errors in this article are the author's.

[5] Small cells refer to spreadsheet cells where the count contained in the cell is, for example, less than five. For example, a spreadsheet containing aggregate data on crimes committed might have rows specifying crimes, and columns containing police divisions. If the count of burglaries in division 55 is 3, this would be a small cell and prompt re-consideration of the spreadsheet.

De-identification should be considered as a solution to a request for data when the data custodian wants, or is required, to grant the request for data, and when aggregated data will not meet the requestor's stated analytical needs. De-identification should not be considered where the risk tolerance for re-identification is zero. This might be the case, for example, where the data custodian is a rehabilitation clinic for high profile individuals. It is also the case that privacy officers should consider offering to do the analysis for the requester, and provide only the results. If the requestor cannot articulate the analysis they require, this might be an indication that the request is a 'fishing expedition'.

De-identification is a process by which a data set containing personally identifiable information is modified such that it can be used or disclosed without substantially increasing the risk of a privacy breach. For example, a pharmaceutical company conducting a clinical trial will require de-identified patient data from the doctors conducting the trial to analyze and report on the trial as part of the approval process in bringing the study drug to market. From the point of view of the data custodian, de-identification should be viewed as a risk management tool and as a safeguard that is part and parcel of the toolkit of a privacy program.

According to Ontario's *Personal Health Information Protection Act* (PHIPA): "identifying information" means information that identifies an individual or for which it is reasonably foreseeable in the circumstances that it could be utilized, either alone or with other information, to identify an individual.[6] In the United States, the *Health Insurance Portability and Accountability Act* (HIPAA) Privacy Rule establishes national standards to protect individuals' medical records and other personal health information. The Privacy Rule is located at 45 CFR Part 160 and Subparts A and E of Part 164[7]. In its summary of the HIPAA Privacy Rule, the U.S. Department of Health and Human Services states that there "are no restrictions on the use or disclosure of de-identified health information. De-identified health information neither identifies nor provides a reasonable basis to identify an individual."[8]

These definitions, which we take to be a general approach to the reduction of risk for personally identifiable information of any type, not just health data, contain a two-part test. The first test is to ensure that the record or data set does not directly identify an individual. Removing direct identifiers such as name or address is the FIRST step in de-identifying a record or data set. The second test for identifiability is a reasonableness test. PHIPA uses the phrase 'reasonable in the circumstances', and HIPAA refers to a 'reasonable basis'. Therefore the SECOND step in de-identifying a record or data set is to ensure that safeguards are in place that reduces the risk of re-identification below the reasonableness threshold, however that might be defined in the data custodian's context. The good news is that this means that risk does not need to be reduced to zero. The bad news is that reasonableness is a vague test and, with respect to the technologies applied to data analysis, what is reasonable will change over time as re-identification techniques and technologies evolve.

In the practice of privacy and security, it is often the case that data custodians and data processors apply their own internal standards in respect to the utility of data to determine risk. This is incorrect, and a systemic problem leading to security and privacy breaches. For example, an HR practitioner will look at a spreadsheet of employee information (excluding names and addresses) and see a tool for workforce analysis with a low risk to privacy. An adversary might see that same information as a link with other information for the purpose of identity theft. Should such an adversary obtain the HR practitioners unencrypted laptop with the spreadsheet on it, the loss

[6] *Personal Health Information Protection Act, 2004*, SO 2004, c 3, Sch A, s. 4 (2).

[7] U.S. Department of Health & Human Services, *The HIPPA Privacy Rule*.

[8] U.S. Department of Health & Human Services, *Summary of the HIPAA Privacy Rule*.

should be classified as a privacy breach. The correct view of 'reasonable' for determining the risk of re-identification, therefore, is to apply the views and capabilities of a potential adversary to the record or data set in question. If such an adversary could use the information by itself or in combination with information that such an adversary is likely to have to identify individuals, then that information has NOT been de-identified and should be treated as PII[9].

There will be breaches

According to the Canadian Bankers Association, there were 74.5 million credit cards and 28.1 credit card accounts in 2011. If we assume that 10% of Canadians still get paper statements mailed to them, that means about 7 ½ million statements are mailed out each month. If a credit card statement is delivered to the wrong address, or is opened by the wrong person, then it should be presumed to be a privacy breach.

The opportunities for this kind of error are relatively easy to identify. The credit card company could mis-address a statement, the credit card holder could fail to notify the bank of a change of address in time, or the Canada Post could deliver to the wrong address. Let us assume that these are rare events and that 99.99% of the time each statement is delivered correctly. That means that each month an average of 74,500 statements are delivered incorrectly, each of which is likely to be a privacy breach.

Successful de-identification requires passing a two-stage test, as described above. The first stage involves the removal of direct identifiers. Direct identifiers are fields that can uniquely identify individuals and includes names, government issued identifying numbers, e-mail, or mail addresses. The removal of direct identifiers can be addressed using a set of techniques generally referred to as masking[10]. These techniques include:

- Variable Suppression - Removing direct identifiers from the data set.
- Randomization - Replace direct identifiers with fake (random) values.
- Shuffling - Take the value from one record and switch it with a value from another records.
- Create Pseudonyms - Replace unique values such as a Social Insurance Number with a one way hash or a random pseudonym.

Other techniques used for masking, but that are ineffective, include character scrambling, character masking, field truncation, and encoding. Data masking is a necessary component of de-identification but it is insufficient by itself.

Another important consideration in determining whether a data set or record can be de-identified is to determine whether or not 'free-form' text fields or attachments are part of the data set. Free form fields contain notes,

[9] I sometimes refer to this as the 'stupid' test. If there is a breach related to an organization not doing something, will it make the organization look stupid? Examples would be sensitive files found in a dumpster, or the loss/theft of an unencrypted laptop. Publishing a file as de-identified, and having journalists use that data to re-identify individuals also fails the 'we look stupid' test. The more polite phrase is 'reputational damage'.

[10] Emam, K. E. (2013). The False Promise of Data Masking. In *Risky Business: Sharing Health Data While Protecting Privacy* (pp. 129-133). Trafford Publishing.

comments, or literally anything else that could be entered. In a health setting this could be physicians' notes about a patient. In a commercial setting this could be a comment field in a customer relationship management (CRM) database. Similarly, attachments might include voice recordings from a customer support call, or emails from a client. The inclusion of attachments or free form text fields presumptively means that the data set should not be regarded as de-identified unless there are preventative controls to ensure that identifiable data are not included in the attachments or free form fields.

Once a data set is masked a privacy professional is left with set of records that looks, to a layman, like it has been de-identified. In the naïve sense of removing direct identifiers this is true. However, as noted above, de-identification is not complete until an assertion can be made that the data cannot be used to re-identify a person. A data sharing agreement requiring the requestor to not attempt to re-identify individuals in the data-set is also necessary, but not sufficient. Putting the processes in place to assert de-identification is the second part of the two stage test for de-identification described above, and can be the more difficult one because it requires defining an 'acceptable' level of re-identification risk. Factors that will influence this definition include who will have access to the data set, who the likely adversaries are, and the general sensitivity of the data. The acceptable level of risk for a data set that will be released to the public is different than the risk that could be accepted for a data set released to a business partner that has contractually agreed not to attempt to re-identify the data.

The literature on re-identification identifies three types of re-identification risk. Each type of risk is associated with a particular type of adversary, where adversary is defined as the entity that is motivated to try and re-identify an individual or individuals in a de-identified dataset.

Marketer Risk: This is the case where the adversary is seeking solely to identify as many individuals as they can, typically by attempting to link records in the de-identified dataset to a database to which the adversary has access, such as a list of property owners in an area.

Journalist Risk: This is the case where the adversary is seeking to identify a single individual in the de-identified dataset. For the purposes of a journalist identifying a single individual will suffice to embarrass the data custodian. Managing journalist risk will also manage marketer risk.

Prosecutor Risk: This is the case where the adversary is attempting to identify a specific person, where they know that the specific person is in the database and where they know something about that specific person. Managing prosecutor risk also manages journalist and marketer risk. The adversary can know that a person is in the database if:

- The dataset contains the whole population, such as a provincial disease registry.
- The dataset contains a large sample or subset of the population, but the adversary has information that makes it reasonable to infer that the target individual is included in the sample.
- The target person self-reveals, or is publicly 'outed' as being in the dataset, such as when a celebrity is known to enter a health facility or emergency department.

The requirement for a data custodian facing a request for their data is to determine which of these risks apply to the request that they have received. From a best practice perspective, if prosecutor risk applies, that is the risk to manage, otherwise the data custodian should manage journalist risk.

A key safeguard in ensuring that de-identified data remains de-identified is by making the data requestor responsible. As a matter of due diligence, a data custodian should evaluate a data requestor to ensure that they can meet their responsibilities to keep de-identified data secure and to not attempt to re-identify the data. This evaluation could, at one end of the due diligence spectrum, be a requirement that the data requestor submit to a third party audit of their security and privacy policies and practices, with a full report to the data custodian. If the data requestor has, as part of its own privacy and security program, completed privacy and security assessments of its systems it may be that an attestation to that affect may be sufficient to satisfy a data custodian. Responses that would tend to disqualify a requestor from receiving de-identified data might be some of the following[11]:

- “The data is de-identified, so it doesn’t need to be protected” - This suggests that the requestor fails to understand re-identification risk.
- “What’s a privacy impact assessment?” - This suggests that the requestor does not have the capability to protect privacy.
- “We password protect our files, to it’s OK” - This suggests that the requestor fails to understand the complexity of security, and will have weak security safeguards.

A minimum reasonable set of questions for a data requestor might be the following:

- Has there been a privacy impact assessment (PIA) of the system or systems that will process the requested information in the last two years?
    - Have the risk mitigation recommendations of the PIA been addressed?
    - Please provide a copy of the PIA, or a summary of the results of the PIA and how the recommendations were addressed.
- Has there been a threat and risk assessment (TRA) of the system or systems that will process the requested information in the last two years?
    - Have the risk mitigation recommendations of the TRA been addressed?
    - Please provide a summary of the results of the TRA and how the recommendations were addressed.
- Has there been a security or privacy incident involving the system or systems that will process the requested information in the last two years? If so please provide the details and the remediation steps that have been undertaken.

If the data requestor is asking for data based on highly sensitive information and is capable of ensuring privacy and security, these questions will be relatively easy to respond to. The requestor's own security and privacy policies may limit the details that they are allowed to provide, which can be reassuring, but in any case they should be able to provide summaries and some form of attestation to demonstrate their capabilities of meeting their responsibilities.

[11] These remarks have been encountered by the author in his practice.

A data custodian should establish what their acceptable re-identification risk levels are for their data. Since we know that there will always be SOME risk of re-identification, the data custodian should set a threshold value for re-identification risk. According to El Emam, "When we assess re-identification risk for a data set, we assign a probability of successful re-identification to each record in that data set. For identity disclosure, the probability of re-identification means the probability of that record being assigned a correct identity."[12]

In other words, there should be policy guidance based on precedent or the regulatory environment that establish what 'reasonable' means for that custodian's data. The following table [13] identifies potential thresholds that could be chosen:

Threshold	Interpretation
τ	The highest allowable probability of correctly identifying a single record.
α	The proportion of records that have a high probability of re-identification that would be acceptable to the data custodian.
λ	The average proportion of records that can be correctly re-identified that would be acceptable to the data custodian.

The mathematics for applying these rules are out of scope of this article, and readers would be well advised to consult with a subject matter expert to determine the metrics that would be applicable for them. It may be useful to think of de-identification in the same way that privacy professionals think of cryptography. While there may be the occasional privacy officer with a background that enables them to understand the mathematics of cryptography, this is not an expected qualification for the job. What is expected, however, is that a privacy officer will know when cryptography should be applied to data at rest or data in motion, and will not hesitate to consult with an expert in the field where required. The same logic, and reliance on technical expertise, should be used to establish risk metrics and thresholds for re-identification. Note that IT or database systems expertise does not normally include expertise in de-identification or in assessing re-identification risk.

Your privacy policies should set out the goals of your de-identification efforts, preferably in a technology neutral manner. As with any privacy policy there will be assertions to the affect that personally identifiable information will only be collected, used, or disclosed with consent for identified purposes or where permitted or required by law. If it is foreseeable that there will internal requests for record level data for purposes other than those identified, then there should be policy statements to the affect that de-identified data will be used for other purposes, and that individuals will not be identified when data is used for such purposes. To ensure transparency, this policy statement should be available to data subjects either on request or, preferably, as part of a publicly available privacy policy. Similar assertions should be made in respect of disclosures of de-identified data to external requesters. Because of the sensitivity of external disclosures, the policy should be as specific as possible

[12] (El Emam, 2013) p. 177

[13] Table 16.3 from (El Emam, 2013), p. 182

with respect to disclosures. For example, a policy could say that de-identified customer purchase data is provided to market research organizations to better determine customer requirements. It would be prudent to include a statement to the effect that when de-identified data is disclosed, the recipient will be required not to re-identify individuals or sell the data to other recipients.

Data requestors may have access to other data that could be used incidentally or deliberately to re-identify individuals. To avoid this embarrassment, de-identified data sets should still be classified as highly sensitive, or confidential. This should mean that security policies and processes apply to the de-identified data. This will in turn ensure that access to and usage of the de-identified data will be monitored and logged by the data requestor.

Segregation of duties is an internal control used to prevent fraud and error. A typical example is that a person requesting an expenditure can not be the person approving that expenditure. In that same vein the person or entity that requests access to personally identifiable information should not be the person or entity that determines the risk of re-identification. Responsibility for re-identification risk assessment should be assigned to a person, department, or third party that is accountable for managing organizational risk. That ensures that the assessment will balance the benefits from releasing potentially identifiable data with the associated risks to privacy. Needless to say, this responsibility should be matched with an appropriate level of expertise.

At the end of the day all data custodians will be faced with requests for access to their data. In order to ensure that they manage the privacy risk associated with these requests, custodians need to have the policies and processes in place to demonstrate that their privacy objectives are being met. These elements can be summarized as follows:

1) Policy

    - Policy statements that establish acceptable levels of risks, using data based metrics
    - Policy statements that establish accountability for approving data requests, ensuring segregation of duties between data requestors and data approvers.

2) Processes

    - Processes to manage data requests (internal and external)
    - Processes to approve (or reject) data requestors
    - Processes to de-identify and release data sets
    - Awareness and training to ensure that personnel are aware of, and can follow, the policies and processes.

By ensuring that line staff are aware of the real risks of re-identification they can be trusted to be the first and best line of defense against unauthorized use or disclosure of personally identifiable information. By implementing policies and processes that are based on data based validation of re-identification risk, a privacy officer can manage organizational risk while still meeting business requirements for data analyses from multiple stakeholders.

KRIS KLEIN'S
PRIVACY SCAN
WITH SHAUN BROWN AND MURRAY LONG

PrivacyScan is published by Law Office of Kris Klein, a professional corporation

To Order

By Phone:	**613.225.2906**
By e-mail:	privacyscan@krisklein.com

Please provide subscriber name, phone number and e-mail address as well as billing name and billing address.

Annual Subscription:	$900.00 plus GST
Four-month trial subscription:	$300.00 plus GST
Annual subscription for not-for-profit (non-governmental) groups	$600.00 plus GST

Please inquire about discounted group and corporate rates.

PrivacyScan is for the intended use of the subscriber only and is designed to provide up to date information on issues related to privacy legislation in the private sector in Canada. Every reasonable effort is made to ascertain the accuracy and reliability of information presented. However, any description and analysis reflects solely the interpretation and opinions of The Law Office of Kris Klein, a prof. corp. who makes no claims as to the absolute reliability and accuracy of any information presented herein, and accepts no liability or responsibility for any errors or omissions. Subscribers are encouraged to seek qualified legal advice on points of law or matters of interpretation.