Canadian Assessment Framework

Big Data Assessment for Canadian
Private Sector Organizations Project

28 February 2017

CONTENTS

INTRODUCTION

Organizations need to engage in advanced analytics to be successful today. To do so, they need to be able to determine that their big data analytics are legal, fair and just. The purpose of this document is to introduce an assessment framework that helps an organization not only determine whether its big data activities achieve the ethical goals of legal, fair and just but also demonstrate how that determination was reached.

Big data, "data sets so large, lacking in structure, and changeable from one moment to another that traditional methods of data analysis no longer apply",[1] are critical to innovation, and innovation is critical to a healthy, sustainable economy. The Canadian Commissioner of Competition, Innovation and Infrastructure recently remarked that: "[s]trong competition drives innovation, which in turn drives productivity, efficiency and economic growth" and that "the collection, analysis and use of data is increasingly becoming an important source of competitive advantage, driving innovation and product improvement."[2] For "consumers, innovation brings more choices and higher quality products and services in a dynamic marketplace."[3] As the OPC has observed, "[n]ew technologies enable businesses and governments to collect and analyze exponentially greater quantities of information using complex computer algorithms, leading to advances in areas ranging from the tailored treatment of diseases to the optimization of traffic flows.[4]

[1] OPC (Office of the Privacy Commissioner of Canada) (2016), "Consent and Privacy, A discussion paper exploring potential enhancements to consent under the Personal Information Protection and Electronic Documents Act" ("Consent Discussion Paper"), OPC, Gatineau, p. 6, *https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2016/consent_201605/*.
[2] Pecman, J. (2016), "Competition, Innovation and Infrastructure", Competition Bureau, Government of Canada, *http://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/eng/04092.html*.
[3] *Id.*
[4] OPC (2016), *2015-2016 Annual Report to Parliament on the Personal Information Protection and Documents Act and the Privacy Act, Time to Modernize 20th Century Tools*, OPC, Gatineau, p. 1, *https://www.priv.gc.ca/media/4160/ar_201516_eng.pdf*.

The *Personal Information Protection and Electronic Documents Act* (PIPEDA) is a pragmatic, flexible framework for a healthy, sustainable economy derived from big data analytics.[5] PIPEDA is technology and sector neutral. By recognizing both the right of privacy of individuals with respect to their personal information and the need of organizations to collect, use or disclose personal information for purposes that a reasonable person would consider appropriate in the circumstances, PIPEDA balances the rights of individuals and the interests of organizations.[6] PIPEDA requires accountability, specifying that organizations must be responsible for the personal information under their control.[7] However, big data analytics can create accountability challenges, and an assessment to determine whether big data activities are legal, fair and just may be part of an organization's privacy management program.[8]

Generally, under PIPEDA, organizations must obtain an individual's consent, whether express or implied, in order to collect, use or disclose an individual's personal information.[9] PIPEDA's consent requirements also establish a practical framework for advanced data analytics by contemplating circumstances where organizations must process personal information in connection with the provision of a product or service.[10] Specifically, under PIPEDA, organizations can require an individual to consent – within the terms and conditions of the product or service – to data analytics provided that (i) the organization complies with the specified transparency and data minimization requirements (both of which are consistent with other PIPEDA requirements)[11], and (ii) the collection, use and disclosure in question is required to fulfill "legitimate purposes" (Legitimate Purposes Based Consent). Although the phrase "legitimate purposes" is not defined, it is informed by Section 5(3) of PIPEDA, which describes

[5] While this Introduction and the Five Key Values cite PIPEDA, the private sector and health sector privacy laws are equally applicable to the assessment process set forth herein.

[6] PIPEDA's purpose is to establish rules to govern the collection, use and disclosure of information in an era in which technology increasingly facilitates the circulation and exchange of information. Section 3 of PIPEDA. PIPEDA's purpose provision, which attempts to reconcile two competing interests, the individual's right to privacy and the commercial need for access to personal information, should be interpreted with flexibility, common sense and pragmatism. *Englander v. Telus Communications Inc.,* 2004 FCA 387 at ¶ 46, *http://www.canlii.org/en/ca/fca/doc/2004/2004fca387/2004fca387.html*.

[7] Principle 4.1 in Schedule 1 to PIPEDA.

[8] Privacy management programs help "promote trust and confidence on the part of consumers, and thereby enhance competitive and reputational advantages for organizations." Office of the Information and Privacy Commissioner of Alberta, Office of the Privacy Commissioner of Canada, and Office of the Information and Privacy Commissioner for British Columbia (2012), *Getting Accountability Right with a Privacy Management Program* ("Joint Accountability Guidance"), p. 1, *https://www.priv.gc.ca/media/2102/gl_acc_201204_e.pdf*.

[9] Principles 4.3.1 and 4.3.6 in Schedule 1 to PIPEDA.

[10] Principle 4.3.3 in Schedule 1 to PIPEDA provides: "An organization shall not, as a condition of the supply of a product or service, require an individual to consent to the collection, use or disclosure of information beyond that required to fulfill the explicitly specified, and legitimate purposes."

[11] The phrase "explicitly specified" is informed by PIPEDA Principle 4.2 Identifying Purposes and Principle 4.8 Openness, and the data minimization wording ("beyond that required") is informed by and consistent with Principle 4 Limiting Collection and Section 5(3) Appropriate Purposes.

appropriate purposes as those that are "appropriate in the circumstances" as considered by a reasonable person,[12] and by the balancing of interests within the Purpose Section of PIPEDA.[13]

Organizations in all sectors have been engaging in data analytics for decades in order to conduct research and development, prevent fraud, secure systems, and operate and provide products and services. In order to bolster their ability to maintain that their particular types of data analytics are reasonable, legitimate and appropriate in a given set of circumstances, organizations may conduct an assessment process that considers the ethical goals of legal, fair and just when assessing the impact of big data analytics. The Information Accountability Foundation (IAF) has developed such an assessment process. The IAF's assessment process supplements the fair information practice principles and relies on five values: (i) Beneficial, (ii) Progressive, (iii) Sustainable, (iv) Respectful and (v) Fair.[14] The consideration of these five values enhances an organization's privacy management program and its compliance with its accountability obligations under PIPEDA. Ultimately, by providing a framework for establishing that the purpose and nature of a big data analytics activity are reasonable, legitimate and appropriate in a given set of circumstances, the assessment process helps an organization, as part of its privacy management program, determine whether big data activities are legal, fair and just and demonstrate how that determination was reached. When determining whether big data activities achieve the ethical goals of legal, fair and just, the individual's rights are paramount to the interests of the organization.

THE FIVE KEY VALUES

To understand these five values, it is important to appreciate that big data analytics may not be equally impactful on the individuals to whom the data pertains. Big data analytics usually can be separated into two phases: "thinking with data" and "acting with data". Generally, "thinking with data" is where new insights, which go beyond experience and intuition and come instead from correlations among data sets, are discovered. "Acting with data," generally, is where these insights are put into effect and where individuals may be affected as these insights are employed in an individually unique manner. The obligations of legal, fair and just apply to both the "thinking" and "acting" with data phases. While the "acting with data" phase often is individually impactful,[15] the "thinking with data" phase may not be as individually impactful if aggregate data is used (the risks related to false insights usually are the primary concern in the "thinking with data" phase).[16] It often is necessary to distinguish between "thinking with data"

[12] Section 5(3) of PIPEDA.
[13] Section 3 of PIPEDA.
[14] The Information Accountability Foundation (2014), "A Unified Ethical Frame for Big Data Analysis", The Information Accountability Foundation, *http://informationaccountability.org/wp-content/uploads/IAF-Unified-Ethical-Frame.pdf*.
[15] When organizations "act with data," individuals have the ability either to opt in or to opt out.
[16] An organization has the ability to "think with data within the reasonable expectations of the individual, and when the organization does so, the organization can assume that the individual has consented." *See* Principle 4.3.5 in Schedule 1 to PIPEDA.

and "acting with data" when considering the five key values.

(i) Beneficial

The purpose of assessment is achieving the benefits that come with data-driven activities while mitigating the possible risks. Essentially, the beneficial value requires a proportionality test.[17] Both the "thinking with data" and "acting with data" phases require an organization to define the benefits that will be created by the analytics and should identify the parties that gain tangible value from the effort. The act of big data analytics may create risks for some individuals. Those risks must be counter-balanced by the benefits created for all individuals and/or society as a whole. Benefits to the organization cannot equal or outweigh those of the individual. Indeed, the organization will need to consider whether it will be necessary to set aside its own interests after "thinking with data" with aggregate data. This balancing concept is consistent with PIPEDA's stated purpose "to establish, in an era in which technology increasingly facilitates the circulation and exchange of information, rules to govern the collection, use and disclosure of personal information in a manner that recognizes the right of privacy of individuals with respect to their personal information and the need of organizations to collect, use or disclose personal information for purposes that a reasonable person would consider appropriate in the circumstances."[18]

To define benefits, one must have an understanding of why the data is being collected, used or disclosed. While big data does not always begin with a hypothesis, it usually begins with a sense of purpose about the type of problem to be solved.[19] Data scientists, along with others in an organization, should be able to define the usefulness or merit that comes from solving the problem, so it might be evaluated appropriately. The risks[20] also should be clearly defined so that they may be evaluated as well. If the benefits that will be created are limited, uncertain, or if the parties that benefit are not the ones at risk from the processing, those circumstances should be taken into consideration, and appropriate mitigation for the risk should be developed before the analysis begins.[21]

(ii) Progressive

Since bringing large and diverse data sets together and looking for hidden insights or

[17] The proportionality test was first set out in *R. v. Oakes* [1986] 1 S.C.R. 103.
[18] Section 3 of PIPEDA.
[19] OPC (2011), *Expectations: A Guide for Submitting Privacy Impact Assessments to the Office of the Privacy Commissioner of Canada* ("OPC Expectations"), OPC, Gatineau, p. 4, *https://www.priv.gc.ca/en/privacy-topics/privacy-impact-assessments/gd_exp_201103/*. The "federal entities undertaking PIAs for particularly intrusive or privacy-invasive initiatives or technologies are expected to first demonstrate that the activity or program is necessary to achieve a specific or legitimate purpose."
[20] An articulation of possible big data risks is set forth in the Canadian Assessment Framework: Questionnaire in Part IV *supra.*
[21] OPC Expectations (The assessment should help determine whether the initiative raises privacy risks; measures, describes and quantifies these risks; and proposes solutions to eliminate privacy risks or mitigate them to an acceptable level.)

correlations may create some risks for some individuals, the value from big data analytics should be materially better than not using big data analytics. If the anticipated improvements can be achieved in a less data-intensive manner, then less intensive processing should be considered.[22] Precision is not required. One might not know the level of improvement in the "thinking with data" phase. Yet, by the time one is proposing to move to the "acting with data" phase, the organization should be better equipped to measure the level of improvement. This application of new learnings to create materially better results is what drives innovation.

Progressive must be assessed in the context in which the processing takes place. There are examples of big data being used to reduce congestion, manage disaster relief and improve medical outcomes where the level of improvement would not have been possible without big data analytics. However, there are other examples where organizations may analyze data and achieve only marginal improvements but only use big data analytics because big data is new and interesting. If there are other methods that will accomplish the same objectives, organizations should consider pursuing those other methods rather than using big data analytics to produce the same or lesser results with greater risks.[23]

(iii) Sustainable

Sustainable covers two issues. The first is understanding how long an insight might be effective, while the second relates to whether the data used for the insight might be available when acting with data.

Algorithms for data analytics have an effective half-life – a period in which they effectively predict future behavior. Some are very long; others are relatively short. Big data analysts should understand this concept and articulate their best understanding of how long an insight might endure once it is reflected in application. Big data insights, when placed into production, should provide value that is sustainable over a reasonable time frame. Considerations that affect the longevity of big data analytics include whether the source data will be available for a period of time in the future, whether the data can be kept current, and whether the discovery may need to be changed or refined to keep up with evolving trends and individual expectations. These considerations are consistent with PIPEDA's Accuracy and Collection Limitation Principles.[24]

There are situations where data, particularly de-identified data, might be available for the

[22] *Id.* (The OPC expects federal entities to demonstrate that the intrusion on privacy is proportional to the benefit to be derived.) The concept of proportionality comes into play when conducting assessments on all of the values, but it particularly comes into play on the progressive value. Proportionality in this context is a process minimization consideration, and therefore it is not a balancing test.

[23] *Id.* (The OPC expects federal entities to demonstrate that no other less privacy intrusive alternative would achieve the same purpose.)

[24] Principle 4.4 in Schedule 1 to PIPEDA ("Information shall be collected by fair and lawful means."); Principle 4.6.1 ("The extent to which personal information shall be accurate, complete, and up-to-date will depend upon the use of the personal information, taking into account the rights of the individual. Information shall be sufficiently accurate, complete and up-to-date to minimize the possibility that inappropriate information may be used to make a decision about the individual.")

"thinking with data" phase but would not be available in the "acting with data" phase because of legal or contractual restrictions.[25] These restrictions affect sustainability.

(iv) Respectful

Respectful relates directly to the context in which the data originated and whether that context is reasonable and to the contractual or notice related restrictions on how the data might be applied. As the OPC stated in its recent discussion paper on consent and privacy, "the principle of "respect for context" bears some conceptual resemblance to the idea of "consistent use" employed in the federal Privacy Act in which a use or disclosure that is consistent with the purpose for which the information was originally collected may not require the individual's consent. The key to employing either concept is the way in which the original "context" or original "use" is defined, since this will determine how broad a range of other uses can be considered "respectful" or "consistent."[26] The analogy to the idea of consistent use in the federal Privacy Act does not suggest that consent, either express, implied or Legitimate Purposes Based, is not needed. Rather the analogy illuminates that the key to respect for context is the way the original "context" is defined since that definition determines how broad a range of other uses can be considered respectful.

Big data analytics may affect many parties in many different ways. Those parties include individuals to whom the data relates, organizations from whom the data originates, organizations that aggregate the data and those that might regulate the data. All of these parties have interests in the data that must be taken into consideration and respected. The assessment framework can help organizations understand what is respectful. When determining whether big data activities are respectful, the individual's rights are paramount to the interests of the organization.

Organizations using big data analytics should understand and respect the interests of all the parties involved in, or affected by, the analytics, and that in certain circumstances the rights of the individual have priority. Anything less would be disrespectful.

(v) Fairness

While "respectful" speaks to the conditions related to, and the processing of, the data, "fair" relates to the insights and applications that are a product of big data. The impacts of that processing must be fair and just.

Canadian law prohibits discriminatory practices based on race, national or ethnic origin, colour, religion, age, and sex.[27] Yet, big data analytics may predict those characteristics without actually looking for fields labeled race, national or ethnic origin, colour, religion, age, or sex. The same can be said about genotypes, particularly those related to physical characteristics. Inferring

[25] *Id.* at Principle 4.4 ("information shall be collected by fair and lawful means.")
[26] Consent Discussion Paper, p. 17.
[27] Canadian Human Rights Act, R.S.C. 1985, c. H-6. Similar provincial laws exist.

characteristics and using them to make decisions based on prohibited grounds is not just. Big data analytics, while meeting the needs of the organization that is conducting or sponsoring the processing, must be fair to both the individuals to whom the data pertains and to whom it will be applied.[28]

The analysis of fairness needs to protect against unseemly or risky actions but also to enhance beneficial opportunities. There are risks related to being too reticent with data. Human rights speak to shared benefits of technology and broader opportunities related to employment, health and safety. Pre-empting such opportunities is also a fairness issue.[29] Indeed, a benefit of value to the organization may also lead to a value to the public. If organizations do not do the analytics, then society will not benefit, but if the analytics are done, then the assessment framework needs to evaluate whether a benefit actually flows to the individual.

In considering the value of being fair, organizations should take steps to balance individual rights in a manner that gives weight to the interests of other parties but recognizes that not all interests have the same weight and that the rights of those individuals who will be impacted by the analysis have priority.[30] Results should not be gamed to favor the data users.

HOW THIS ASSESSMENT FRAMEWORK MAY BE USED

The purpose of this assessment framework is to assist organizations to leverage the potential of big data that pertains to individuals in a manner that is consistent with Canadian law while protecting individuals from the risks of both using and not using data. This assessment framework raises additional considerations that may not be covered in a typical privacy impact assessment (PIA). It assists organizations in looking at the rights and interests impacted by data collection, use and disclosure in data-driven activities.

This assessment process is broader in scope than the typical PIA process. For example, all data are considered in this assessment process and not just personal information. Therefore, all aspects of this assessment process include data in the aggregate, non-identifiable form that are outside the scope of PIPEDA. However, to the extent an assessment process can be used to consider and appropriately mitigate the impact of a personal information practice, such a process may supplement (or be woven into) the organization's PIA process. In this regard, this assessment process may enhance an organization's privacy management program and compliance with its accountability obligations under PIPEDA.

---

[28] Section 3 of PIPEDA. See text at note 18 *supra*.

[29] Principle 4.4 in Schedule 1 to PIPEDA ("Information shall be collected by fair and lawful means").

[30] *See* OPC Expectations (The OPC asks government departments to answer the following four questions, which are based on the proportionality test set forth in *R. v. Oakes*:
Is the measure demonstrably necessary to meet a specific need?
Is it likely to be effective in meeting that need?
Is the loss of privacy proportional to the need?
Is there a less privacy-invasive way of achieving the same end?)

This assessment framework does not replace PIAs; it should be used in conjunction with PIAs; it is not a complete PIA. Organizations may incorporate this assessment framework in whole or in part into their own unique processes and programs and may use a triage process to determine the questions that are appropriate to ask considering their own circumstances and the level of assessment necessary. For example, if the activity in question is only minimally changed from the past, no assessment might be necessary. If data is being used in a manner that is crystal clear from privacy notices and context, then a PIA might be all that is necessary.

This assessment framework may be used as big data activities reach key milestones or decision points. Some level of assessment may be appropriate at each phase of a big data activity. Big data analytics may include phases when the activity is first conceived, then approved for programming, put into operation and eventually reviewed. Questions need not be repeated in later phases if underlying conditions have not changed. If there have been changes to the activity that impact answers, then the questions may need to be repeated. If questions between the two assessments are duplicated, then they may not need to be repeated, and organizations may need to determine where in their processes it makes sense for the questions to be asked.

Benefits and risks, and their likelihood of materializing, may be assessed based on an organization's approach to risk. As new data analysis and new applications of insights can change over time, the process of assessing benefits and risks may need to be repeated. Regardless of when the assessment is conducted, the results of each assessment may be presented to decision makers for a determination on whether to proceed with an activity.

The assessment framework identifies key issues that decision makers in organizations may consider. No score is generated that makes decisions for users. Rather, if decision makers take into account what they learn from the assessment process, then decisions may be made in a manner that gives weight to the interests of other parties but recognizes that the rights of those individuals who will be impacted by the data analysis have priority.

The sample questions help evaluate whether based on the assessment the activity is reasonable, appropriate and legitimate. Use of the assessment framework helps determine whether the decisions reached on the appropriateness of an activity were well reasoned and demonstrate how that determination was reached. Organizations may wish to disclose that they use this assessment framework or one based on it.

While developed for big data activities, this assessment framework and the "thinking and acting with data" distinction may be used to assess any activities within an organization where data is collected, used and disclosed in a manner that may not have been anticipated by the individuals to whom the data pertains. By providing a framework for determining whether the purpose and nature of data analytics are reasonable, appropriate and legitimate in a given set of circumstances, the assessment process helps an organization, as part of its privacy management program, determine whether its data analytics are legal, fair and just and demonstrate how that determination was reached.

Canadian Assessment Framework: Questionnaire

The assessment framework questionnaire follows a Privacy Impact Assessment (PIA) format, but it does not constitute a complete PIA. Other elements are necessary (e.g., security safeguards and records management) for this questionnaire to constitute a complete PIA process. The questionnaire starts with the common elements of the PIA process and is followed by elements specific to Big Data activities. Each section contains Factors for Consideration and sample Questions. It is not expected that this framework will be the assessment process. Rather, this framework is proposed as a structure for organizations to use in developing their own assessment processes. This questionnaire is overly detailed so that decision makers will have options in developing their own assessment processes, and this assessment process may require involvement by stakeholders from different parts of the organization. It is expected that organizations with existing mature PIA processes will only use parts of this questionnaire as applicable. The extent to which this questionnaire is incorporated into existing PIA processes will depend on the prior experience of the organization, previous knowledge of the stakeholders, the nature of the activity, and the type of data involved. While all questions may not be incorporated into an organization's PIA process, it is expected that an organization will be able to demonstrate that all of the five key values were considered as part of the PIA assessment process.

Contents

INTRODUCTORY ELEMENTS

CHARACTERIZING THE ACTIVITY

The team working on the activity should consist of people with knowledge to answer the questions in this questionnaire, and the person knowledgeable about each question used should answer it.

ACCOUNTABILITY: Identify the individuals who are responsible and accountable for the activity.

Factors for Consideration:

Activity team can include:

- Data capture/acquisition
- Data preparation/ management
- Oversight for restrictions (i.e., legal or contractual)
- Application of the analysis/insights

Questions:

Who has the ultimate decision making authority?

Who needs to be involved in making the decision?

PURPOSE: Understand the purpose and intended outcomes of the activity.

Factors for Consideration:

Examples of business needs related to the insight or the problem/question that needs to be solved include:

- Risk management
- Solution and product capability
- Distribution network
- Brand enhancement
- Marketing: traditional direct mail, email, telemarketing, digital advertising
- Service improvement
- Utilization of an existing service/product
- Education of client base
- Product development
- Market development
- Organizational effectiveness
- Process improvement
- Cost savings
- Derive deep insights
- Create products

Questions:
What is the business need that prompted this activity?
Is this activity an expansion of a previous activity? If yes and a previous assessment was done, that assessment should be attached to provide continuity.
Does the purpose of the activity fit within a larger theme of work that is currently being contemplated or undertaken?

DATA: Understand the nature of the data.

Factors for Consideration:
Data linking is combining data associated with an identifiable individual in one data base with data associated with an identifiable individual in one or more other data bases. Linkable data is data about or related to an individual for which there is a possibility of logical association with other data about the individual. Examples of sensitive data: ▪ Medical records ▪ Biological/physiological features ▪ Health information ▪ Income and tax records ▪ Sexual orientation ▪ Race or ethnic origin ▪ Religious beliefs ▪ Political affiliation or opinions Unstructured data is information that is not organized in a pre-defined manner.
Questions:
Is the data identifiable personal information?
Has there been data linking of an identifiable individual's data?
Is the data reasonably linkable to a particular individual?

Is the data anonymous?
Are there data elements that are the product of a probability based process, such as a score?
Has the data been aggregated such that it is no longer identifiable personal information?
Has the data been combined with other data so that the data is re-identifiable?
Is the data sensitive?
Is the source data structured or unstructured?
SOURCES: Understand the sources of data to be used in the activity.
Factors for Consideration:
Examples of data sources: ▪ Individuals, collected directly by your organization ▪ Third parties, collected from individuals and provided to your organization ▪ Created through statistical analysis or calculations (identify whether this data is created by your organization or a third party) Examples of data origins: ▪ Provided by the individual ▪ Scraped from the web ▪ Obtained from public sources ▪ Provided by third-party aggregator ▪ Observed in some other fashion ▪ Derived from data (i.e., transformation/manipulation) ▪ Inferred from analytics ▪ Provided by vendor ▪ Obtained from social media Source reliability/data accuracy: Data accuracy may be directly related to how data was sourced. Data directly observed may be more precise than data inferred from an algorithm. Data directly observed may be more accurate than data volunteered by individuals.
Questions:
What are all the sources of the data, internal and external?

What actual data elements are found in the data?
How was the data from each source originated?
How reliable is the source of the data for its purpose?
Who has custody or control of the data source?
What are the governance arrangements among the data source controllers?

PREPARATION: Understand the pre-processing that will be done before the analysis.

Factors for Consideration:

Organizations may have standard processes to manage data preparation. If so, this section may not be necessary once the standard process has been reviewed. It is appropriate to rely on standard processes, and it is not expected that this process will re-evaluate standard data preparation processes if they exist within the organization.

Examples of data protection:

- Data linking
- De-identification*
- Anonymization
- Obfuscation
- Encryption
- Removal of personal information

*The Information and Privacy Commissioner of Ontario (IPC) has issued guidance on de-identification: IPC (2016), "De-Identification Guidelines for Structured Data", https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf.; IPC (2014),"De-Identification Protocols: Essential for Protecting Privacy https://www.ipc.on.ca/wp-content/uploads/Resources/pbd-de-identifcation_essential.pdf.; IPC (2013), "Looking Forward: De-Identification Developments – New Tools, New Challenges", https://www.ipc.on.ca/wp-content/uploads/2013/05/pbd-de-identification_developments.pdf.; and IPC (2011), "Dispelling the Myths Surrounding De-Identification: Anonymization Remains a Strong Tool for Protecting Privacy", https://www.ipc.on.ca/wp-content/uploads/2016/11/anonymization.pdf. Organizations may have existing processes for de-identification relying on this guidance, and it is not expected that this process will not re-evaluate those existing processes if they exist within the organization. If de-identification processes do not already exist, organizations are encouraged to refer to the cited guidance.

Questions:
Are there unique data protections that should apply to this activity?
Has there been a risk analysis of the chosen method of data protection? If yes, what is the risk of re-identification?

CONTRACTUAL AND LEGAL CONDITIONS: All processing and applications should be within the context of the conditions associated with the data.

Factors for Consideration:

As a general obligation, uses of data should be transparent. Organizations should make or have means of making information about their use of data accessible to individuals for the purposes of consent. The answers to these questions relate to transparency.

Obligations associated with the data include:

- Applicable laws/regulations
- Policies
- Contracts
- Industry obligations, including codes of conduct
- Authorization
- Privacy Policy
- Privacy Statement
- Internal Policies
- Disclosures (other than the Privacy Policy and Privacy Statement)

Examples of restrictions that may apply to publicly available data:

- Website agreements
- Terms and conditions, terms of use or website policy (e.g., scraping or commercialization prohibited by website, sensor or derived data)
- Limitations on data uses from public sources

Questions:
What obligations apply to the data?
Have all obligations associated with the data been identified?
If there are obligations, how are these obligations being satisfied?
Does consent or other legal authority exist to use the data?

Is the activity covered by identified purposes or notices? If yes, provide the applicable wording.
Have obligations to or from third parties been established?
Are there prohibitions against re-identification that apply?
Are there any restrictions on data (opt outs, withdrawals of consent, age) that would affect the use of the data? If yes, list them.
If the data has been collected by and received from others, does that party have authority and can the authority of that party be relied upon?
If the data is being shared with a third party, are there obligations that would preclude or limit that sharing? • If there are limits on the sharing of the data, will the third party agree to the transfer subject to those limitations.
If the source of the data is public, are there any restrictions that would apply to the use of the data?
<u>ACCURACY:</u> Evaluate the accuracy of the consolidated data.
Questions:
What steps are being taken to determine the accuracy of the data?
Is the data appropriate for the purposes of the activity?

The foregoing are just the Introductory elements of a PIA. Other elements of a PIA, such as security safeguards and records management, should be considered in addition to the Big Data Elements of a PIA.

BIG DATA ELEMENTS

CHARACTERIZING THE ACTIVITY

The team working on the activity should consist of people with knowledge to answer the questions in this questionnaire, and the person knowledgeable about each question used should answer it.

PURPOSE: Understand the purpose and intended outcomes of the activity.

Factors for Consideration:

If the initial purpose is generating insights, this questionnaire may need to be repeated when the insight is applied to solve a particular question/problem or if the insight generates a particular question/problem to be solved.

(Note: Data flow mapping may be a technique that can help answer these questions.)

Questions:

Is the purpose of the activity "thinking with data" (i.e., generating insights)? If yes, what insight is the activity expected to generate?

Is the purpose of the activity "acting with data" (i.e., solving a particular question/problem)? If yes, what is the particular question/problem this activity is trying to solve?

How does the purpose of the activity fit within the current business model or strategy?

How does the purpose of the activity fit within the values of the organization?

How does the purpose of the activity fit within the values of society?

INSIGHTS: Understand what insights are expected from the analysis.

Factors for Consideration:

The questions below link to "thinking with data" (i.e., generating insights). If the original activity being assessed did not contemplate "acting with data", all assessment questions from the beginning will need to be asked again when the insights are applied or acted on.

Insights reflect a particular point in time and can naturally change over time. The actual application of an insight may begin the process of change. Many of the questions below are designed to understand how durable those insights might be.

A demonstration can be useful in helping to understand the insights. Insights can be a report. Insights used beyond effectiveness can have negative impact. Application of insights may lead to changed behavior, and this changed behavior may impact predictiveness.

Insights may have unintended impacts. Consider additional analysis for unintended correlations. Consider how the answer relates to the scope of the activity. It will be necessary to go back to the beginning if "acting with data".

Secondary use means the use in research of identifiable personal information originally collected for a purpose other than the current research purpose.* All secondary uses constituting research involving humans will be governed by the Tri-Council Policy Statement.

*Canadian Institute of Health Research, National Sciences and Engineering Council of Canada and Humanities Research Council of Canada (2010), *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans*, p. 62, http://www.pre.ethics.gc.ca/eng/archives/tcps2-eptc2-2010/Default/. ("Tri-Council Policy Statement")

Questions:

How were the potential insights derived?

Will the activity expand on insights from a previous activity? If yes, could the activity be a secondary use?

How might the potential insights be used?

Are uses of the potential insights internal or external?

How long might the potential insights endure?

Could the potential insights become less useful or valuable over time?

Are the potential insights repeatable and for how long?

Could the application of the potential insights impact behaviour in a manner that could reduce the predictive value of the insights over time?

Are the potential insights reliable enough for the purposes of the activity?
Is there a less data-intensive way to obtain the potential insights?
Is it foreseeable that the potential insights might seem inappropriate or discriminatory or might be considered offensive causing distress or humiliation?
<u>PREPARATION:</u> Understand the pre-processing that will be done before the analysis.
Factors for Consideration:
Steps in preparation of data for analysis are: ▪ ETL (i.e., extract, transform, load) ▪ Data standardisation ▪ Data hygiene ▪ Data integration (i.e., consolidation) ▪ De-identification "Nature of the data" refers to quantitative data consisting of numbers (i.e., quantities) which represent counts or measurements and to qualitative data consisting of nonnumeric characteristics (such as gender or race)
Questions:
What work will be done to put the data in a consistent format?
How will errors and redundancy in the data be identified and dealt with?
Will time impact the data set? If yes, how will the impact be dealt with?
How will the data sources be consolidated for analysis? Has the data been evaluated before consolidation?
Will further synthesising of the data, deriving different data elements from various source elements, be necessary?
Will the nature of the data change when thinking with data?
When preparation is complete, will the identifiability of the data change?

ACCURACY: Evaluate the accuracy of the consolidated data.

Factors for Consideration:

Inferences from data analytics may not be treated with the same confidence as information collected directly from individuals.

Data Transformation: Data from diverse sources in diverse formats must be put into a form where the data can be analysed. This process can impact the accuracy of the data itself and the insights that might come from big data.

Questions:

Does the age of the data affect its accuracy?

Is the predictive level of inferred data taken into consideration?

Are there additional steps that could be taken to protect individuals without impacting the accuracy of the data?

Has consolidation/transformation impacted the data in such a way that accuracy is affected?

What is the potential impact of inaccurate data?

IMPACTED PARTIES: Identify all the impacted parties.

Factors for Consideration:

Relevant impacted parties include:

- Individuals/data subjects (including consumers, customers, prospects and employees)
- Organizations (including businesses and non-governmental organizations)
- Political entities/government/law enforcement
- Society/public-at-large/community
- Regulators
- Others

Questions:

Who are all the relevant impacted parties?

ASSESSMENT: Determine Activity Impact

IMPACTED PARTIES: Identify all the concerns of the impacted parties.

Factors for Consideration:

Key concerns regarding the impacted parties include:

- Data use out of context
- Contract, legal, privacy or security obligations
- Data sensitivity
- Revenue/business needs
- Harm to either individuals or to the organization or to society or to all three
- Reticence risk

Other concerns include:

- Cultural differences
- Commonly held societal values
- Compatibility with organizational values
- Compatibility with the use of sensitive information.

Questions:

What are the key concerns parties may have arising from this activity, if any?

Are there any other concerns with respect to impacted parties that should be taken into account?

OUTPUTS: Understand what is expected from applying the insights.

Factors for Consideration:

The questions below link to “acting with data” (i.e., solving a particular question or problem).

Outputs reflect status at a particular point in time. Status can naturally change over time. The actual application of an insight (“acting with data”, i.e., solving a particular question or problem) may begin the process of change. Many of the questions below are designed to understand expected outputs.

Output (i.e., application of insight) can be a report.

Results should be verified. Correlations can be spurious, and causation should be confirmed.

Questions:
How will the insight be applied?
Is it expected the application of the insight will have an impact on individuals, including employees and customers? If yes, what is the impact?
Is it foreseeable that application of the insight might be considered offensive causing distress or humiliation or might seem inappropriate or discriminatory?

BENEFITS AND IMPACTS

Factors for Consideration:

There may be more than one benefit for an impacted party.

Examples of impacts to organizations include:

- Improved profitability
- Enhanced employee satisfaction, engagement and productivity
- Enhanced customer relationship
- Undamaged brand/reputation
- Enhanced brand/reputation
- Increased market share
- Prevention of cyber-crime and fraud
- New/improved/innovative products/services
- Improved customer service

Examples of impacts to individuals include:

- More objective outcomes
- Safer interactions
- Better product selection
- Better access to new products and services
- Significant discounts
- Better product utilization
- Improved service
- Improved ease of use
- Engaged consumers/customers/employees
- More convenience
- Appropriately linked to other choices, etc.
- Anticipating or meeting of a need
- Exercise of self-determination
- Public sector access
- Anonymous transportation

Examples of impacts to society include:

- Better health care
- Improved education
- Positive impact on climate change
- More accessible/usable technology
- Protection of reasonable expectation of privacy, including anonymity
- Protection of freedom of religion, thought and speech
- Protection of prohibition against discrimination on basis of race, national or ethnic origin, colour, religion, age, sex, sexual orientation, marital status, or disability

Consider evaluating expected (versus actual) benefits and impacts.

There might be relevant subsets of the Impacted Parties where differential impact might occur.

Questions:

For each impacted party identified above, what are the benefits that are expected to come from the analysis and/or use of the insight?

Could the impact result in discrimination?

RISKS AND MITIGATION

Factors for Consideration:

When assessing risks to impacted parties, consider potential impacts of false positives or negatives.

In addition to the concerns identified above, examples of risks to individuals include:

- Physical harm
- Financial harm
- Health
- Reputational
- Embarrassment
- Shock or surprise
- Inappropriate discrimination
- Access to and misuse of data
- Crime, national security, and law enforcement
- Manipulation of needs (i.e., creation of a need where one previously did not exist)
- Incidental findings

In addition to the concerns identified above, examples of risks to organizations include:

- Negative media attention
- Negative regulatory impact
- Compliance
- Reputational
- Business continuity
- Financial loss
- Reticence risk
- Crime

In addition to the concerns identified above, examples of other risks that may impact individuals but also certain communities of segments of the society include:

- Discrimination
- Differential impact
- Exclusion
- Security breach

Risk mitigation includes outcomes.

- Evaluate expected outcomes.
- Consider whether false positive, false negative and security risks could occur.
- Application of data protection methods (e.g., data linking, de-identification, anonymization, obfuscation, encryption, removal of personal information)

Questions:

Does each concern identified above create a material risk for the impacted party?

For each impacted party identified above, are there any risks that could come from the analysis or the use of the insights?

For each identified risk, how likely is it to happen, and what is its severity?

How is each risk tested and mitigated?

Are there any residual risks after mitigation?

WEIGHING OF BENEFITS AND RISKS

Factors for Consideration:

After the Benefits and Impacts and Risks and Mitigations have been evaluated, there should be a weighing of the Benefits and Impacts against the Risks and Mitigations.

The Benefits and Impacts and the Risks and Mitigations should be rational and proportional to each other. The risks have to be proportional to the benefits.
Questions:
Are the mitigated risks sufficiently balanced by the benefits?

DECISION

OVERALL EVALUATION

Factors for Consideration:

The analysis in this section should be based on the answers in the previous sections. Factors not considered before but considered in this section should be noted. The consideration of all these sections is to establish that the purpose and nature of the activity is reasonable, appropriate and legitimate and, therefore, legal, fair and just.

Questions:

Based on the assessment, is the activity reasonable, appropriate and legitimate?