UTS CRICOS 00099F

Ian Oppermann
Industry Professor, Faculty of Engineering and IT
University of Technology Sydney, NSW
Ian.Oppermann@uts.edu.au
29/11/2020

RE: Review of the Privacy Act 1988 (Cth) Issues Paper

The University of Technology Sydney (UTS) Faculty of Engineering and IT (FEIT) appreciates the opportunity to respond to the Commonwealth on the review of Privacy Act 1988.

UTS FEIT is supportive of reforms in:

- Updating the definition of 'personal information' to capture technical data and other online identifiers - Recommendation 16(a)
- Strengthening existing notification requirements - Recommendation 16(b)
- Strengthening consent requirements and pro-consumer defaults - Recommendation 16(c) and Introducing a direct right of action to enforce privacy obligations under the Privacy Act - Recommendation 16(e).
- Consideration of the right to erasure – noting challenges around practicality achieving this
- Consideration of a binding privacy code for organisations that trade in personal information

Further, with regard to the intersection of privacy and data, UTS FEIT encourages:

- a concerted effort to distinguish between Personal Information (captured about a person, their location in time and space, their actions, their preferences, their relationship to people and objects) and Personally Identifiable Information (reasonable ability to identify an individual) in any future legislation which deals with people centric data,
- incentives to improve levels of education for all actors involved in data sharing and use (including data analysis) in the important areas of data governance, obtaining meaningful consent, assessment of data quality,
- developing guidance on appropriate use of analytical insights,
- mandating a minimum level of accreditation in data governance for actors using personal information for commercial purposes.

The Commonwealth is encouraged to note the work of the international standards community in data sharing and use, in particular, the new Working Group on Data Usage operating under JTC1's Subcommittee 32, Working Group 6. This standards effort is exploring several foundational areas of work related to privacy and data specifically: terminology and use cases; appropriate use of data and insights; and clarification of the use of terms Personal Information versus Personally Identifiable Information in JTC1 standards.

The Commonwealth is also encouraged to note the ongoing work of the Australian Computer Society in its efforts to develop privacy preserving data sharing frameworks.

Finally, the Commonwealth is encouraged to note the ongoing work of the Data61/CSIRO to extend the work of the ACS to develop a numerical measure of Personal Information Factor (PIF) in deidentified data sets.

1. BACKGROUND

Future Smart Services for homes, factories, cities, and governments rely on sharing of large volumes of often personal data between individuals and organisations, or between individuals and governments. The benefit is the ability to create locally optimised or individually personalised services based on personal preference, as well as an understanding of the wider network of users and providers.

Data from more sources used in more ways is an inevitable part of our future and yet, some fundamental issues have not yet been adequately addressed in Australia or elsewhere in the world to support the increased use of people centred data.

Most importantly, there is currently no way to unambiguously determine if the level of personal information in aggregated data or when increasing levels of personal information reaches the point of being (reasonably) personally identifiable. Context plays an important role in understanding if information is personal, or (reasonably) personably identifiable. Where and when someone is, what they are doing, who they are with, what objects they are connected to, what services they access are all ways of building a picture of who someone is. As more services become digital and connected, the ability to rapidly form this picture of who someone is, and gain knowledge of actions, preferences, and relationships becomes increasingly possible. When these data sets are reviewed by a person with their own context and knowledge of the world, the challenge of not inadvertently identifying an individual or gaining very personal information becomes even more difficult.

De-identification and aggregation are common approaches used to reduce the level of personal information in a dataset when linking or releasing. Different deidentification approaches and different levels of aggregation are used by organisations depending on a perceived value of an associated risk. The implications of this are profound when thinking of the use cases which come in and out of scope depending on the level of aggregation used.

2. CONSIDERATIONS OF PRIVACY WHEN USING DATA

2.1 Personal Information in People Centric Data

The terms Personal Information and Personally Identifiable Information are often used interchangeably as seen in different legislative frameworks around the world and even in different standards. Date of birth is often considered personal information (information about a person) but used alone, this single feature is not personally identifiable information unless it uniquely identifies an individual known to be in a dataset.

The question becomes, how many features can be linked before *Personal Information* becomes *Personally Identifiable Information* of a person known to be in a dataset? A recent paper published in Nature Communications[1] provides a means to estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset. This paper is one in a long series of works which show the small number of features which can be linked to identify a unique individual in a population. The focus on "heavily incomplete" datasets in the paper shows the limitations of the protection associated with creating uncertainty as to whether an individual is in a dataset.

[1] L. Rocher, J. M. Hendrickx and Y. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models", Nature Communications, July 2019. Available online https://www.nature.com/articles/s41467-019-10933-3

The concepts of *Personal Information* versus *Personally Identifiable Information* are not clearly differentiated in regulatory frameworks. Personal information is typically described so as to cover a very wide field and is described differently in different parts of the world. For example, in the state of NSW:

> *"… personal information means information or an opinion (including information or an opinion forming part of a database and whether or not recorded in a material form) about an individual whose identity is apparent or can reasonably be ascertained from the information or opinion".*

The legal tests for personal information generally relate to the situation where an individual identity can *"..reasonably be ascertained"*. The definition is very broad and in principle covers any information that relates to an identifiable, living individual for 30 years after their death.

2.2 Risk Frameworks for Data Analysis – the "Five Safes"

It is important to think of privacy and personal information at all phases of the data lifecycle not just at the point of collection or analysis including collection, transmission, storage, analysis, reuse of data or analysis and ultimately deletion of data.

The Five Safes Framework is a conceptual model to consider aspects of risk and governance at the analysis phase of the data lifecycle. Several organisations around the world including the Australian Bureau of Statistics use the Five Safes framework to help make decisions about effective use of data which contains personal information or is sensitive. The Five Safes model is however not a widely accepted model internationally.

The Five Safes framework is relatively easy to conceptualise when considering the extreme cases of 'extremely' Safe although it does not unambiguously define what this is. An extremely Safe environment may involve researchers who have had background checks, projects which have ethics approval and rigorous vetting of outputs from that data environment ("Outputs"). Best practice may be established for such frameworks, but none of these measures is possible to describe in unambiguous terms as they all involve judgement.

In September 2017, the Australian Computer Society (ACS) released a technical whitepaper which explored the challenges of data sharing and use[2]. The paper highlighted the fundamental challenge for the creation of smart services is addressing the question of whether a dataset contains personal information. The paper proposed a modified version of the "Five Safes" framework[3] for data sharing and use which attempts to quantify different thresholds for "Safe". In November 2018, the ACS released a second technical whitepaper on Privacy Preserving Frameworks[4] which evolved the concepts introduced in the first paper. A third whitepaper was released in 2019 which extended the framework further and produced a simple measure of Personal Information in a data set.

The adapted model described by the ACS whitepapers explores different, quantifiable levels of "Safe" for each of People, Projects, Setting, Data and Outputs as well as how these different "Safe" levels

[2] See ACS website, available online https://www.acs.org.au/content/dam/acs/acs-publications/ACS_Data-Sharing-Frameworks_FINAL_FA_SINGLE_LR.pdf

[3] T. Desai, F. Ritchie, R. Welpton, *"Five Safes: designing data access for research"*, October 2016, "http://www.nss.gov.au/nss/home.NSF/533222ebfd5ac03aca25711000044c9e/b691218a6fd3e55fca257af700076681/$FILE/The%20Five%20Safes%20Framework.%20ABS.pdf

[4] See ACS website, available online https://www.acs.org.au/content/dam/acs/acs-publications/Privacy%20in%20Data%20Sharing%20-%20final%20version.pdf

could interact in different situations. Figure 1 shows the dimensions of the adapted "Five Safes" framework taken from the 2018 ACS Technical whitepaper.

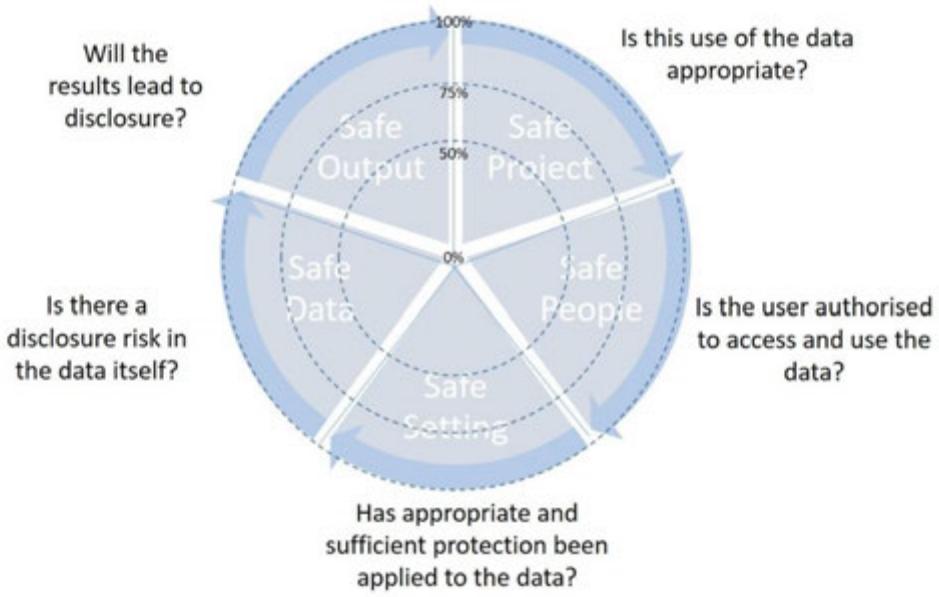

Figure 1. Modified Five Safes Framework

2.3 A Personal Information Factor – Measuring How "Safe" Data Is

The ACS Technical whitepapers explored a hypothetical parameter, the "Personal Information Factor" (PIF) which was a measure of the personal information in a linked, deidentified dataset or in the outputs of analysis.

A PIF of 1 means sufficient personal information exists to identify an individual: the total personal information (PI) is personally identifiable (PII). A value of 0 means there is no personal information. It is important to note that the PIF described is not a technique for anonymisation: rather, it is a heuristic measure of potential risk of reidentification.

The PIF for both data and outputs is described based on:

- A measure of the information content of the dataset used to conduct analysis or the output of the analysis (the simplest analysis may be sharing of data);
- The uniqueness of the most unique individual (group) in the dataset or output;
- Additional information required by the observer to be able to identify an individual from the data or outputs.

Figure 2 shows the context for evaluating the degree of personal information as part of assessing the PIF in a closed system. The data available in a closed environment is finite and a PIF can be described mathematically based on knowledge of uniqueness of combinations of features describing an individual.

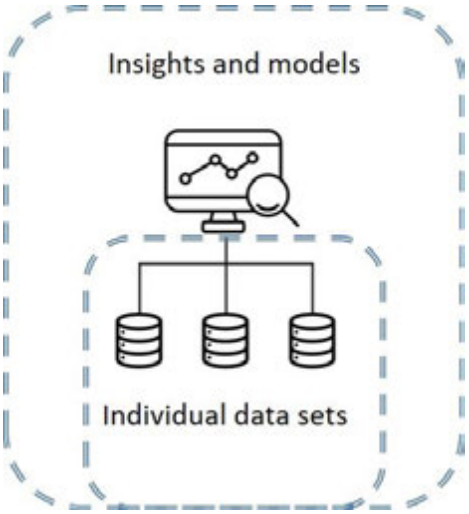

Figure 2. Closed system context for evaluating PIF

The degree of personal information contained in data may be very high (a unique identifier such as a social security number), moderate (such as surname), low (such as eye colour), or very low (such as month of birth). The linking of datasets may however not be sufficient to identify an individual, that is, the linked dataset does not contain sufficient personal information to be identifying (does not have PII).

It is expected that the degree of personal information (the PIF) in a linked dataset will generally increase as more, diverse datasets are linked. As conceptually shown in Figure 3, as more datasets containing PI are linked, a point may be reached where an individual is personally identifiable (a PIF of 1), or "reasonably" identifiable (a PIF within "epsilon" of 1). The dataset is then considered to have PII. The "epsilon" in this figure is an indication of the difference represented by the gap before the "reasonable" threshold is met.

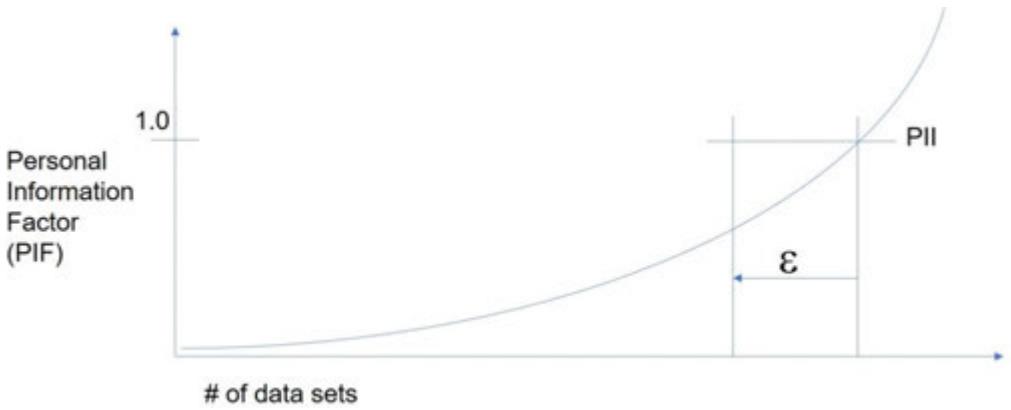

Figure 3. Conceptualisation of a Personal Information Factor (PIF) and the threshold point of PII

3 A SNAPSHOT OF JTC1 STANDARDS ACTIVITY

Through Standards Australia, Australia is an active participant in standards work of ISO, IEC and JTC1. A new Working Group was created in 2020 to explore Data Usage. This Working Group is led by Australia and supports other Working Groups including AI and Smart Cities. A snapshot of activities of the new Working Group is provided below.

3.1 Appropriate Use of Data Analytics

Whilst the restrictions to data use are often cited as concerns related to privacy, many of the concerns relate to sensitivities of data use or unintended consequences of use of insights including:

- sensitive subjects captured in data,
- concerns about data quality used for analysis (accuracy, resolution, scope, bias),
- confidence in outputs (accuracy, precision, consistency, explainability, bias),
- consequences of how outputs (insights or data driven decisions) will be used,
- concerns about whether human judgement will be applied before an insight becomes a decision,
- possible harms resulting from use or reuse of outputs (reversable harms, reversable with cost, or irreversible harms),
- results from analysis leading to negative surprises or embarrassment,
- the need for expert knowledge or context required to appropriately interpret results of analysis,
- concerns about accidental release of data or insights (outputs),
- concerns about data age (or data which has never been examined).

The dimensions of sensitivity can be addressed in turn and frameworks developed to address these concerns or risks.

Appropriate use of analytics outputs is ultimately a subjective matter and supported by appropriate governance frameworks. Such governance frameworks address controls in response to dimensions of risk and sensitivities.

Additional work on metadata focused on data collection, data provenance, data quality, data subject, and data use is considered a priority area for data usage and would support the governance frameworks for appropriate use of analytics outputs.

3.2 Terminology and Use Cases

Data use is relevant to many areas and data sharing and use has been described in different ways across existing groups and even in different standards. Standardised terminology and harmonised use cases are needed in the market to facilitate wider data usage and to unlock the value of data sharing, exchange and exploitation.

3.3 Data Quality

Data quality up to the point of usage is dependent on numerous parameters associated with the early stages of the data value chain. An example is given for illustration (Figure 6). The non-exhaustive list

of parameters could all be supported by appropriate metadata. The context in which a data is used determines which kinds of metadata are most valuable and the utility which is needed. Metadata on provenance, collection methods, data quality are fundamental to data usage. The scope of Metadata fields will be different depending on use cases. Figure 7 provides examples of different data collection (creation) scenarios depending on whether data was A) collected by a device, B) generated by a person, C) if the data is received by a device, or D) by a human.

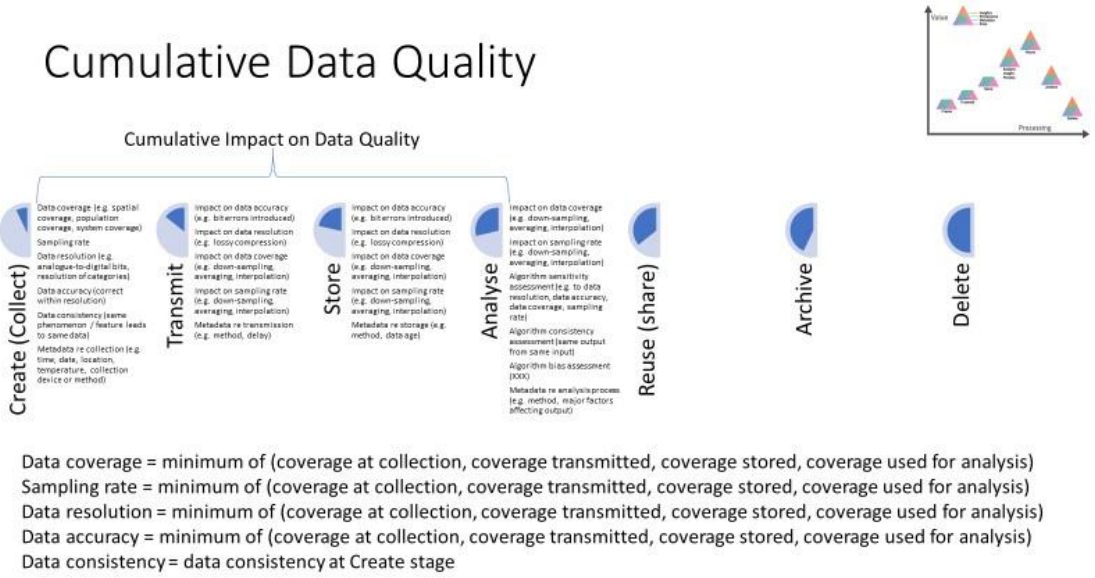

Figure 6. Example metadata field for data quality collected by a device and transmitted to a device

The context in which a data is used determines which kinds of metadata are most valuable and the utility which is needed. Metadata on provenance, collection methods, consent, and data quality are fundamental to data usage. The scope of Metadata fields will be different depending on use cases. Figure 7 provides examples of different data collection (creation) scenarios depending on whether data was A) collected by a device, B) generated by a person, C) if the data is received by a device, or D) by a human.

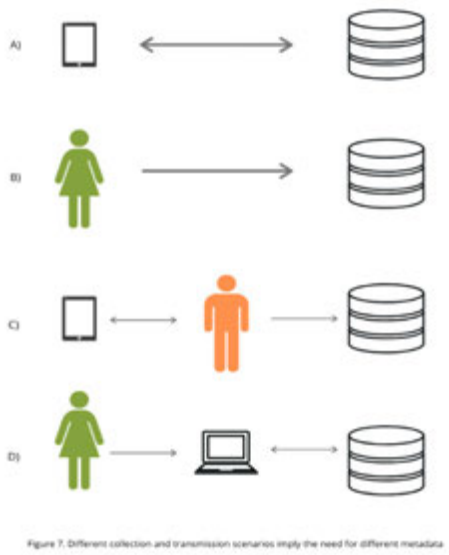

Figure 7. Different collection and transmit scenarios imply the need for different metadata

3.4 Metadata

If a data is to be shared amongst individuals, organizations, or applications, each of the users of data must understand the structure, meaning, categorization, etc. of the data in the same way. Without a common understanding of those characteristics of data sets, usage of the data is unlikely to produce the desired results.

Metadata is the mechanism for providing the required knowledge about data so that it can be shared (and used) appropriately. The nature of specific data sets and the nature of the uses to which the data sets are being applied determine the kind(s) of metadata that is required. Although a great many data sets are associated with at least one set of metadata, that metadata tends to be structural metadata. (For example, SQL databases and XML documents that are valid with respect to an XML schema or a DTD.)

Provision of metadata other than structural metadata is much less common in practice than it should be. In an apparent majority of situations, most non-structural metadata is found in organization culture, natural-language written documentation, comments in computer code, and other forms that are difficult to use automatically. Although this situation is sub-optimal, the bare minimum metadata required in order to create metadata utility is structural metadata. Much of the other kinds of metadata needed for appropriate usage of data can be shared between users of the data set, either in writing or orally.

While structural metadata can create utility for data usage, considerable value is added to data if other kinds of metadata are known and made available.