

What We Talk About When We Talk About Personal Data

This is the fifth note of [S&R Data+](#), a series focused on distinguishing between personal and non-personal data. This note is divided in two: in part 1, we discuss the conceptual paradigm of personal data; in part 2, we discuss certain important categories, along with recent trends and future implications in connection with such data categories.

Part 1: The Conceptual Paradigm of Personal Data

Introduction

The terms ‘information’ and ‘data’ are both used in the context of privacy and data protection laws. For instance, India presently regulates the use of data under the Information Technology Act, 2000, as amended by the Information Technology (Amendment) Act, 2008 (collectively, the “**IT Act**”), along with its allied rules. More specifically, its existing data protection framework (the “**DP Framework**”) stems from Section 43A of the IT Act, along with the Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011 (“**SPDI**”).

The DP Framework

With respect to the DP Framework itself, Explanation (iii) of Section 43A of the IT Act, read with Rule 3 of SPDI, defines ‘sensitive personal data or information’ (emphasis added). This may include, among other things, financial and biometric information. However, Sections 2(1)(v) and 2(1)(o) of the IT Act, respectively, define ‘information’ and ‘data’ differently.

On the other hand, Rule 2(1)(i) of SPDI separately defines ‘personal information’ – as relating to “a natural person, which, either directly or indirectly, in combination with other information available or likely to be available with a body corporate, is capable of identifying such person.” Although somewhat similar, India’s current draft of the [Digital Personal Data Protection Bill, 2022](#) (“**DPDP**”) (released in November last year for public comments) defines ‘personal data’ in simpler but broader language – as “any data about an individual who is identifiable by or in relation to such data.” In turn, DPDP defines ‘data’ in a manner largely consistent with its equivalent under the IT Act – albeit more streamlined than in the latter – adding components such as suitability of communication and interpretation, respectively, along with outcomes from ‘automated processing’.

Other jurisdictions

Nevertheless, the use of either term – *i.e.*, ‘data’ or ‘information’ – may not signify a distinction *inter se*. Other jurisdictions also appear to *not* make this distinction. For instance, while the [Personal Data Protection Act 2012 of Singapore](#), as amended (“**PDPAS**”), uses terms such as personal and derived personal data, the [Privacy Act 1988 of Australia](#) (“**PAA**”) defines ‘personal information’ along similar lines. Certain other laws, such as those

of states like [Massachusetts](#) in the United States (the U.S. lacks a consolidated and/or sector-agnostic data privacy law), limit the ambit of personal information to certain critical data elements only, and further, related to residents alone. On the other hand, as of this year, the recently amended [California Consumer Privacy Act of 2018](#) (“**CCPA**”) gives consumers new rights – such as that of opting out of the sale/sharing of their personal information altogether, and/or to limit the use and disclosure of sensitive personal information.

Items under personal data

Personal data may include any kind of information about an individual person. Thus, it could involve objective information, such as the person’s name, mobile phone number, locational data (using functions on their cellular device, for instance), an internet protocol (“**IP**”) address, a cookie ID, or details about a hereditary condition. In addition, a person’s economic information (e.g., details about a bank loan or a household bill) may be considered personal data too – even though corporate borrowings or a firm’s consumption/expenditure records may not be similarly covered.

Generally speaking, information *only* about a business and/or corporate entity may not constitute personal data because this information would not identify a specific human individual. However, information about a person may be intertwined with relevant data about their business or company – to the extent that such data may constitute personal information too. For instance, where the business is owned and/or managed by a sole proprietor, the distinction between business information and personal data may overlap.

Further, information that is regulated separately (*i.e.*, under a different legal regime) may be considered ‘personal,’ and thus remain eligible for the same level of protection. Further, even if not recognized as personal data under an applicable law – say, under Australia’s PAA – but by dint of being explicitly acknowledged as such under a different legislation, such information may nevertheless be granted similar protections. For example, meta-data involving subscriber and account details under the Australian Telecommunications (Interceptions and Access) Act 1979 may be viewed as personal information under the PAA.

The meaning of personal data may also involve subjective information, including opinions, estimates, or assessments. Such subjective information may be especially relevant in sectors such as insurance (e.g., assessments made about an individual’s life expectancy) and banking (e.g., evaluations about the reliability or creditworthiness of a borrower). In this regard, although previous iterations of DPDP – such as Clause 3(18) of the [Personal Data Protection Bill, 2019](#) (“**PDP 19**”) and Clause 3(19) of the [Personal Data Protection Bill, 2018](#) (“**PDP 18**”) – had defined ‘financial data’ separately, DPDP does *not*. However, information or opinions about an individual from sundry activities, such as assessments about their retail or other preferences based on online purchases or web browsing history, may qualify as personal data. In this regard, even anonymized data, like aggregated statistics, may be used to enrich existing profiles of individuals, thus creating new data protection issues.

While certain jurisdictions, like Singapore, explicitly clarify that the underlying information may not necessarily be true or proven in order to constitute personal data, (see Section 2(1) of PDPAS), this is implicit in other national legislations as well. For instance, under DPDP, an individual has the right to correct her personal data (see Clause 13(1)), including when it is inaccurate or misleading. While a ‘data principal’ under DPDP is the individual in respect of whom the personal data relates (Clause 2(6)), a ‘data fiduciary’ is any person who alone or in conjunction with others determines the purpose and means of processing personal data (Clause 2(5)). Such terms in DPDP have the same meaning as those related to ‘data subject’ and ‘controller’, respectively, within the [General Data Protection Regulation](#) (“**GDPR**”) of the European Union (“**EU**”).

In sum, personal data includes information about people regardless of their position or capacity – *i.e.*, irrespective of whether they are consumers, patients, employees, customers, clients, etc. However, for information to be regarded as personal data, it must be about a specifically identified (or identifiable) individual. Information is ‘about’ an individual when there is a connection between the information and such individual. This is ultimately a question of fact, and hence, will depend on the context and the circumstances of each case. For example, information will be ‘about’ someone where the person is a subject matter of the information or opinion. Information will also be ‘about’ someone where it reveals or conveys something about them.

Accordingly, some information may not be personal data when considered on its own; however, when combined with other information, it may become so. Information can, therefore, be dynamic, and the character of data may change over time. For instance, certain kinds of de-identified, encrypted, or pseudonymized personal data – if susceptible to re-identification with respect to a specific individual in conjunction with other information – may continue to retain its status as such, and will thus be deemed to constitute personal data. However, when such data is rendered anonymous in a way that the individual concerned is no longer identifiable, it may cease to remain ‘personal’. For data to be truly anonymized, however, the corresponding procedure must be irreversible.

New challenges with identifiability

Developments in data science have revolutionized the idea of identifiability. Today, viewing data merely through a binary lens of identifiable or non-identifiable is not useful because of the ways in which data exists in the real world. For instance, whether dynamic IP addresses constitute data about a specific individual may depend on whether the person processing such data has access to additional information that could lead the latter to identify such individual. Further, the degree of identifiability of an IP address may be contextual in the sense that multiple people could be using the same device.

Another challenge associated with identifiability arises from the relative failure of de-identification techniques in respect of their ability to appropriately *conceal*. Current research suggests that in certain situations, it may be possible to identify specific individuals even from apparently anonymized datasets. Further, whether indirect identification is possible in a certain situation may depend on the viability of resources and options, as well as the nature of data available to a data fiduciary with regard to being able to integrate such options with the original information. On the other hand, resource availability may be subject to costs and the state of technology. Thus, even where an individual is not directly identifiable, data about them may need to be treated as personal, given the presence of other factors.

Anonymization

Anonymization requires the use of mathematical/technical methods to distort data adequately and irreversibly for the purpose of ensuring that (re)identification is *not* possible. In this aspect, anonymization is distinct from plain vanilla de-identification techniques that merely involve the masking of identifiers from datasets (although such methods *do* tend to make identification more difficult and/or costly). However, given the pace and nature of technological change, it is difficult to identify precise standards related to anonymization. After all, despite the removal of identifiers, de-identified data involves a higher risk of re-identification. Accordingly, several jurisdictions continue to treat de-identified data as ‘personal’.

Part 2: Important Categories of Personal Data - Trends and Implications

In this part, we discuss some important categories of personal information, including with respect to present trends and implications associated with the collection and use of such information.

Each of genetic, biometric, and health data was separately defined in earlier formulations of DPDP – such as in Clauses 3(20), 3(8), and 3(22), respectively, of PDP 18; and in Clauses 3(19), 3(7), and 3(21), respectively, of PDP 19. However, only biometric and health data have been cursorily referenced under DPDP – and that too with respect to provisions on ‘deemed consent’ alone (see Clause 8).

Nevertheless, as per Paragraph 2.2 of the July 2020 [Strategy Overview](#) of the National Digital Health Mission, health data can be classified into the category of (i) personal health data – including data with personally identifiable information (“**PII**”) of various stakeholders, such as healthcare professionals; and (ii) non-personal health data – which includes aggregated and anonymized health data, where all PII has been removed.

Further, under DPDP, a data principal will be deemed to have consented to the processing of her personal data if such processing is necessary for responding to a medical emergency involving a threat to her (or someone else’s) life or health (Clause 8(4)), or for taking measures to provide medical treatment or health services to an individual during a threat to public health (Clause 8(5)). Such provisions appear to resemble Recitals 53 and 54 of the EU’s GDPR – which, in turn, deal with the processing of sensitive data in the public health and social sectors.

However, while Recital 54 of GDPR indicates that the processing of data concerning health for reasons of public interest may *not* result in processing for other purposes by third parties such as employers (or insurance and banking companies) – under DPDP, if an individual shares their biometric data with an employer – say, for the purpose of marking attendance at the workplace – they will be deemed to have consented to the processing of such data, albeit for the purpose of attendance verification only (see Clause 8(7) of DPDP and its illustration).

Biometric data

Although the term ‘biometric data’ has been left undefined in DPDP, such information may include biological properties, physiological characteristics, living traits, or repeatable actions where such features and/or actions are both unique and measurable, even if the patterns used to ‘measure’ them involve a certain degree of probability. Typical examples of biometric data are fingerprints or retinal patterns, and may include less obvious features such as digitized versions of a handwritten signature or keystrokes.

A special feature of biometric data stems from the fact that it may serve as an identifier: *i.e.*, on account of its unique link to a specific individual, biometric data may be used to specifically identify that person. Thus, for instance, Clause 3(36)(vi) of PDP 19 and Clause 3(35)(vii) of PDP 18 had defined ‘sensitive personal data’ to include that which may reveal, be related to, or constitute, biometric data. However, pursuant to extraction and/or manipulation, if such data is subsequently digitized, the provisions of DPDP may come into play, including with respect to consent (see Clause 7).

Bio-surveillance

‘Bio-surveillance’ via the internet stems from both increased accessibility to data, as well as analytic tools provided by digital infrastructures of social media. In particular, control over mass markets is instrumentalized through mobile and web-based terminals, which remain equipped with a variety of sensors. Accordingly, vast

numbers of individuals may come in contact with such sensor technologies, including with respect to the 'measurement' of their individual characteristics. On account of stabler and faster networks, people now tend to be permanently online, using hand-held and/or connected devices. Accordingly, with the development of application software for smartphones, bio-surveillance has increased significantly. Further, such surveillance is participatory in nature – for example, on social networking or retail sites, which provide comment or rating functions. This leads to increased algorithmic analyses, where such online media platforms emerge as key sources of personal data.

Additional implications

On the one hand, modern devices, sensors, and networks create large volumes and various sources of data, using new technologies and techniques. On the other hand, the cost of storing such data has significantly reduced. As a result, the world has witnessed a growing demand for the re-use of this data, including for commercial purposes. Since the proper use of such datasets may generate substantial public benefit as well, anonymizing personal data may allow for the harnessing of these benefits without compromising the privacy, or risking the identifiability, of specific individuals. Nevertheless, even anonymized data may have an impact on particular persons: *e.g.*, through targeted marketing initiatives or segment-driven advertising, based on customized user profiles.

Moreover, it may be possible to extract sensitive information about specific individuals from social network graphs despite the application of techniques such as pseudonymization (which aims to disguise identities). For instance, a social network platform may operate under the assumption that the particular de-identification technique which it deems fit to use on customer information is robust enough to prevent (re)identification, and pursuant to such assumption, it may proceed to sell such data to a consumer goods company for the latter's marketing and advertising purposes. However, since the relationships between different individuals in the original dataset are unique, those relationships themselves can be used as identifiers.

Thus, pseudonymization is not a substitute for anonymization: the former merely reduces the ability to link a dataset with the original identity of a data principal, and therefore, at best, remains a useful (but fallible) security measure. On the other hand, anonymization techniques can provide privacy guarantees only when their application is engineered appropriately. Therefore, when a data fiduciary does not delete the original (and identifiable) information at the event-level before handing over part of this dataset (even after removing or masking identifiable information), the resulting dataset may nevertheless be considered personal information. Only if the data fiduciary aggregates the information to a level where the individual events are no longer identifiable, may the resulting dataset be considered anonymous.

On the other hand, a pseudonymized dataset can be combined with another in such a way that one or more individuals can be identified. Thus, pseudonymized search engine query strings, especially if coupled with other attributes (such as IP addresses or ancillary client configuration parameters), may remain susceptible to identification. Since pseudonymity allows for identifiability, it stays within the scope of data protection.

Key Takeaways

Various services that individual consumers require and rely upon on a daily basis through mobile and/or internet-linked devices invariably involve the use of app-based platforms which, in turn, collect and use various types of data, such as the user's financial information, their real-time location, as well as information concerning previous transactions. In an increasingly data-driven economy, data flows lie at the core of business processes for companies of all sizes and sectors. Further, new digital technologies tend to unlock new business

opportunities, along with expanding prospects and consequences for the general public, companies, and governments.

Unsurprisingly, therefore, new businesses continue to build expansive databases comprising consumer preferences and behavior. The underlying data can be compressed, organized, manipulated, discovered, and (re-)interpreted. As a result, this informational inventory can be transformed into commercially useful knowledge. In addition, along with reduced costs related to both storage and processing, data collection has also become easier on account of the Internet-of-Things (“IoT”), artificial intelligence (“AI”), machine learning (“ML”), and related technologies. Cumulatively, such circumstances have facilitated the creation of a vast number of customized user profiles with sufficient details and nuance.

For instance, an e-commerce website may track past purchases and other online behavior, and then utilize algorithms to predict the preference-based likelihood of a potential customer buying specific items. Indeed, pooled datasets make it possible to detect trends and thereby enable accurate targeting. In the healthcare sector, as an additional example, service providers can predict diagnoses and suggest treatments by collecting and analyzing large datasets of an individual’s medical records and previous hospital visits. Alternatively, an individual’s locational data could be used for monitoring traffic and improving driving conditions on the road.

While DPDP, when enacted, may replace the existing DP Framework in India, its provisions on notice, consent, deemed consent, and breach may be especially relevant in the future.

The Way Forward

The aggregation and use of customer information may involve both personal and non-personal data. Accordingly, the associated regulatory paradigm will need to consider the difference between the two. Nevertheless, unlike PDP 19 – where, for instance, the central government had the power to direct data fiduciaries to provide it with non-personal and anonymized personal data for targeted public service delivery – DPDP seeks to regulate only digitized personal data. However, going forward, the new digital data regime will likely remain alert about how personal data can be converted into its non-personal equivalent, and the ramifications of such conversion.

*This insight has been authored by **Deborshi Barat** (Counsel); he can be reached at dbarat@snrlaw.in for any questions. This insight is intended only as a general discussion of issues and is not intended for any solicitation of work. It should not be regarded as legal advice and no legal or business decision should be based on its content.*

© 2023 S&R Associates