White Paper

The Definition of Persona Data: Seeing the Complete Spectrum

Omer Tene, Senior Fellow
Christopher Wolf, Founder and Co-Chair
The Future of Privacy Forum

January 2013

The Future of Privacy Forum (FPF) is a Washington, DC based think tank that seeks to advance responsible data practices. The forum is led by Internet privacy experts Jules Polonetsky and Christopher Wolf and includes an advisory board comprised of leaders from industry, academia, law and advocacy groups.

The Definition of Personal Data: Seeing the Complete Spectrum

European law considers data to be "personal" if they relate to "an identified natural person or a natural person who can be identified, directly or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person…."[1] Under both existing and proposed European data privacy regimes, once data are defined to be "personal data" the full complement of legal requirements applies.[2]

The implication of this binary approach to privacy law is that once data are deemed sufficiently de-identified or anonymized, there are no privacy-related obligations that must be met. It is no surprise therefore that the bar has historically been set quite high for determining when data can be considered sufficiently de-identified as to exempt its controllers and processors from data protection requirements. The current binary approach, we argue in this paper, is neither practical nor as protective of individuals as it can and should be.

Whether or not data are "reasonably likely" to be related to an identifiable individual should depend not only on the application of technical measures of de-identification but also on the use of administrative safeguards and legal commitments undertaken by a controller (and imposed downstream on any third party transferees). Only when data are publicly released without appropriate administrative or legal strings and obligations should a solely technical-based de-identification standard apply. And in those cases, differential privacy provides a robust framework to satisfy technical de-identification requirements.

---

[1] General Data Protection Regulation ("GDPR"), Articles 4(1) and 4(2). References to the proposed GDPR are to the European Commission proposal of 25 January 2012: Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), COM(2012) 11 final. References to the Rapporteur's Report are to the Draft Report on the proposal for a regulation of the European Parliament and of the Council on the protection of individual with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (COM(2012)0011 – C7 0025/2012 – 2012/0011(COD)), Committee on Civil Liberties, Justice and Home Affairs, Rapporteur: Jan Philipp Albrecht.

[2] There are sound reasons to question the application of the data protection framework to data held by a party who has no capability of re-identification, simply because "the controller or any other natural or legal person" could re-identify those same data. Consider, for example, a data processor that stores strictly encrypted data without holding the key. Such a processor would be subject to data protection obligations without a practical ability to make sense of the data stored on its equipment. However, this language dates back to the 1995 Data Protection Directive and is therefore not addressed in this White Paper.

The proposed approach enables stronger protections for individuals' right to privacy because it envisions the use of a more comprehensive range of data protection measures. In addition, this approach adjusts practically to the reality that while contemporary mathematical capabilities may theoretically enable re-identification to occur, their impact can functionally and effectively be constrained by legal and organizational measures.

Pseudonymous data

The Rapporteur's proposal of a new concept of "pseudonymous data" is commendable.[3] Such a concept could ostensibly bridge between data that are subject to the full gamut of legal obligations and those that remain exempt from the reach of the law. Yet, the definition of the new term in the Rapporteur's Report is vague (*e.g.*, What does "specific to one given context" mean? Is online behavioral advertising a "specific context"? Does the term seek to distinguish between first and third party uses?); and more importantly, under the Rapporteur's Report controllers lack sufficient (if any) incentive to pseudonymize data.

Specifically, pseudonymization should excuse controllers from certain obligations under the GDPR, such as obtaining explicit data subject consent or providing rights of access and rectification. However, according to the Rapporteur's Report, Article 10 of the GDPR, which excuses a controller from "acquir[ing] additional information in order to identify the data subject for the sole purpose of complying with any provision of this Regulation," does not apply to pseudonymous data. This is the result of the Rapporteur's Report addition of the words "single out" to the Commission proposed text of Article 10. Pseudonymous data is defined as information that is not directly identifiable but "allows the singling out of a data subject." And under the Rapporteur's Report, Article 10 only applies to data that "do not permit the controller to identify or single out a natural person"; hence not to pseudonymous data.[4]

Moreover, the Rapporteur's Report revisions to the consent provisions imply that consent to the processing of pseudonymous data "may be given by automated means using a technical standard with general validity in the Union (…) without collecting identification data."[5] Yet the obligation to document and prove data subject consent under Article 7(1) of the GDPR renders the collection of consent from unauthenticated users theoretically appealing but realistically impractical. Indeed, while the Rapporteur's Report relies on novel theories of authentication without identification,[6] the deployment of such infrastructures has been slow to gain traction in practical business settings. In addition, the Rapporteur's Report's explanatory comments refer to a "Do Not Track" standard,

[3] The Rapporteur's Report; proposed Article 4(2a).
[4] This logical impasse is reinforced by the Rapporteur's Report revisions to recital 45, under which: "If the data processed by a controller do not permit the controller to identify or single out a natural person, the data controller should not be obliged to acquire additional information in order to identify the data subject for the sole purpose of complying with any provision of this Regulation" (adding the words "or single out").
[5] The Rapporteur's Report, proposed Article 7(2a).
[6] The Rapporteur's Report, proposed recital 52.

which has yet to be implemented as a technical standard much less agreed upon from a policy perspective.[7] The end result of these shortcomings is that few controllers are likely to pursue data pseudonymization.

The Limits of Technical De-Identification

As noted, under current European privacy law, data that are not identifiable to an individual are exempt from legal obligations. For many years, technical measures of anonymization or de-identification were thus perceived as a "silver bullet," allowing organizations to collect, analyze, retain and repurpose information while at the same time preserving individuals' privacy.

With the advent of "Big Data" collection, storage and analysis, the societal value ascribed to de-identified data has become immense, driving innovations in healthcare, energy conservation, fraud prevention, data security, and more.[8] Optimally, the legal framework would allow for use of data in ways that maximize such value while minimizing privacy risk. Yet over the past decade, it has become clear that in a world of big data, de-identification based solely on technical measures is increasingly strained by the existence of increasingly effective re-identification techniques.[9]

Today, ample evidence exists of re-identification of apparently de-identified data, through methods such as multiple queries on a single database or linking the data in one database with that in another, which is often publicly available online.[10] Moreover, measures commonly taken to de-identify personal data, such as scrubbing or aggregating certain data fields, degrade the utility of the data for future use.[11] The result is that strictly technical de-identification has become encumbered with a deadweight cost, reducing data utility without eliminating privacy risks.

Internal Use and Limited Sharing: Administrative and Legal Protections

The limitation of technology-based de-identification does not mean, however, that all data should henceforth come under the full remit of data protection laws, regardless of how remote the risk of re-identification. When using data internally or sharing them with third party service providers or business associates, organizations can put in place stringent administrative and legal safeguards *in addition* to technical de-identification, to greatly reduce the likelihood of re-identification. Moreover, organizations frequently have no interest in re-

[7] *See* discussion in Omer Tene & Jules Polonetsky, To Track or 'Do Not Track': Advancing Transparency and Individual Control in Online Behavioral Advertising, 13 MINN. J. L. SCI. & TECH. 281 (2012).
[8] Omer Tene & Jules Polonetsky, Privacy in the Age of Big Data: A Time for Big Decisions, 64 STAN. L. REV. ONLINE 63 (2012).
[9] Paul Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, 57 UCLA L. REV. 1701 (2010).
[10] *See, e.g.*, Arvind Narayanan & Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets, 2008 IEEE SYMPOSIUM ON SECURITY AND PRIVACY 111.
[11] In technical terms, existing methods for aggregation of fields, such as k-anonymity, "fail to compose," meaning that they do not survive multiple instantiations.

identifying specific individuals from de-identified datasets. Treating all such data as personal would create perverse incentives for organizations to forgo administrative and legal safeguards and retain as much personal data as they can.

Not all data are intended for public release or to be shared with unencumbered third parties. Different obligations should be imposed depending on whether the intended recipients of data are the public at large, contractually bound service providers or business associates, researchers subject to institutional review boards, or internal employees. The UK Information Commissioner's Office, in its recent code of practice on anonymization, suggests a similar approach: "It is important to draw a distinction between the publication of anonymised data to the world at large and limited access".[12]

When assessing whether data are considered personal and which data protection obligations should apply, respect should be had not only for technical de-identification but also for administrative and legal safeguards that make re-identification unlikely. Administrative safeguards include data security policies, access limits, employee training, data segregation guidelines, and data deletion practices that aim to stop confidential information from leaking. Legal controls include contractual terms that restrict how service providers or business associates may handle, use or share information, penalties for contractual breaches, and auditing rights to ensure compliance. By implementing administrative and legal safeguards, organizations provide important privacy protections independent of technical de-identification. In other words, the identifiability of data should be assessed on a spectrum as opposed to the existing binary scale. Data that are only identifiable at great technical or legal cost should remain subject to the legal framework yet trigger the application of only a subset of legal obligations.[13]

For example, it makes little sense to provide individuals with a right of access and rectification to data that are not readily identifiable, as this would require data controllers to proactively authenticate identities and re-identify data, increasing privacy risks. The explicit consent requirement in Article 4(8) of the GDPR, particularly when coupled with burden of proof obligations under Article 7(1), will incentivize organizations to increasingly rely on an architecture of authenticated identities. In the context of processing of pseudonymized or indirectly identifiable data, the requirement of explicit consent creates a perverse result that runs counter to other provisions of the GDPR, such as Article 10. The Rapporteur's Report, while recognizing the benefits of data pseudonymization, fails to provide sufficient incentive for organizations to deploy de-identification techniques as it exculpates controllers from neither explicit consent requirements nor access and rectification obligations.

---

[12] Information Commissioner's Office, ANONYMISATION: MANAGING DATA PROTECTION RISK CODE OF PRACTICE, November 2012, http://www.ico.gov.uk/news/latest_news/2012/~/media/documents/library/Data_Protection/Practical_application/anonymisation_code.ashx.

[13] Paul Schwartz & Dan Solove, The PII Problem: Privacy and a New Concept of Personally Identifiable Information, 86 NYU L. REV. 1814 (2011).

Given the stiff requirements for consent, a new legal basis should be added to the GDPR to authorize the processing of pseudonymized data without consent based on a controller's legitimate interests. This would incentivize organizations to implement pseudonymization and prevent the use of authenticated identities strictly in order to prove compliance with the consent provisions of the GDPR. Certain organizations implement business models that require authenticated identities in one context (*e.g.*, webmail services), while at the same time using anonymization or pseudonymization in other contexts (*e.g.*, targeting ads). Such diversity should be encouraged, not stifled, provided that in certain contexts individual accountability, which is based on identity authentication, is key, while in other contexts identification of specific individuals is neither required nor warranted.

Public Release: Differential Privacy

Just as data generation, analysis and re-identification methods have evolved over the past decade, so have privacy enhancing techniques. The existence and promise of such techniques should be recognized and their use should be incentivized in privacy laws.

One new approach that is relevant to de-identification is called "differential privacy."[14] Differential privacy provides a workable, formal definition of privacy-preserving data access, along with algorithms that can furnish mathematical proof of privacy preservation. Differential privacy avoids the ailments of de-identification by allowing data sharing in a way that maintains data quality while at the same time preserving individuals' privacy. It enables data controllers to share derivative data with third parties without subjecting any individual to more than a minimal risk of harm from the use of his or her data in computing the values to be released, even when those values are combined with other data that may be reasonably available. In other words, it allows data controllers and third parties to draw lessons and derive valuable conclusions from a data set, without being able to determine whether or not such conclusions are based on the personal data of any given individual. Hence, differential privacy emphasizes not whether an individual can be directly *associated* with a particular revealed value; but rather the extent to which any revealed value *depends* on an individual's data.

Differential privacy does not solve all privacy problems.[15] It does not provide protection *vis-à-vis* a data controller who is in possession of the data set

[14] Cynthia Dwork, Differential Privacy, in 33RD INTERNATIONAL COLLOQUIUM ON AUTOMATA, LANGUAGES AND PROGRAMMING (ICALP 2006), http://www.dbis.informatik.hu-berlin.de/fileadmin/lectures/SS2011/VL_Privacy/Differential_Privacy.pdf; Cynthia Dwork, Frank McSherry, Kobbi Nissim & Adam Smith, Calibrating noise to sensitivity in private data analysis, in: PROC. OF THE 3RD THEORY OF CRYPTOGRAPHY CONFERENCE 265 (2006). *Also see* Ed Felten, Protecting privacy by adding noise, TECH@FTC BLOG, June 21, 2012, https://techatftc.wordpress.com/2012/06/21/protecting-privacy-by-adding-noise.

[15] Other privacy enhancing technologies in this space include homomorphic encryption and crowd-blending. *See* Craig Gentry, A FULLY HOMOMORPHIC ENCRYPTION SCHEME, PhD Dissertation, Sept.

containing the raw data. And it could leak information where positive correlations exist between individuals whose data reside in a given data set. Yet when compared to existing de-identification methods, differential privacy provides greatly increased privacy and salvages the utility of data which would otherwise have been suppressed under de-identification. In addition, as research progresses, differential privacy science improves accuracy while keeping privacy loss fixed.

Conclusion

European law should avoid a rigid distinction between personal and non-personal data based on strictly technical criteria and divorced from the circumstances of data collection, retention, and use. Where data are maintained by a controller or shared with a restricted group of service providers or business associates, additional safeguards – administrative and legal – beyond technical de-identification can prevent re-identification. Where data are shared publicly or made freely available to transferees, technical safeguards such as differential privacy can be applied to provide adequate privacy. As currently structured, the GDPR should introduce the concept of pseudonomyzed data and give it credence by allowing the processing of such data without consent. This would prevent organizations from pursuing a perverse incentive to identify individuals strictly in order to comply with data protection law.

2009, http://cs.au.dk/~stm/local-cache/gentry-thesis.pdf; Johannes Gehrke, Michael Hay, Edward Lui & Rafael Pass, Crowd-Blending Privacy, 32ND INTERNATIONAL CRYPTOLOGY CONFERENCE (CRYPTO 2012), http://cs.colgate.edu/~mhay/pdfs/gehrke2012crowd.pdf. *Also see* Tom Simonite, How to Share Personal Data While Keeping Secrets Safe, MIT TECH. REV., AUG. 7, 2012, http://www.technologyreview.com/news/428733/how-to-share-personal-data-while-keeping-secrets.