Anonymising personal data: where do we stand now?

Sophie Stalla-Bourdillon Senior Privacy Counsel with Immuta and Professor in Information Technology Law and Data Governance at the University of Southampton, explores the continuing confusion around whether personal data may be truly anonymised, and advises on how to do so using a risk-based approach

When does data that were once personal cease to be so? Is there such a thing as anonymous data within the meaning of the General Data Protection Regulation ('GDPR')?

Even if the GDPR largely echoes the previous Data Protection Directive (95/46/EC), the topic of anonymisation is still heavily debated in the European Union. As an illustration, the French Data Protection Authority — which since 2016 is vested with a power to certify anonymisation practices — is still in the process of building its own approach on the matter. Many practitioners also continue to advise their clients that they should work from the assumption that data can never be truly anonymised, and build compliance strategies accordingly. Why these hesitations? Is there confusion? If yes, where could it come from?

Defining anonymisation

Recital 26 of the GDPR, just like Recital 26 of the Data Protection Directive, excludes from its remit anonymous data. It specifies that, "[t]he principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable." The test found in the Recital to determine whether an individual is identifiable is based on "all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly." Reading Recital 26, it seems that the drafters of the GDPR agreed that the standard for determining whether data are anonymous should be based on a probabilistic assessment, or, in other words, on a tailored risk-based assessment.

This seems also to be the view of the UK Information Commissioner's Office ('ICO') which in 2012, reformulated the definition of anonymised data in its Code of Practice on anonymisation in the following way: "We use the term 'anonymised data' to refer to data that does not itself identify any individual and that is unlikely to allow any individual to be identified through its combination with other data." The Code of Practice offered a useful tool to determine the likelihood of re-identification: a motivated intruder test, which is intended to make controllers anticipate the future behaviour of third parties. The Code of Practice explicitly derived the main consequence of an approach based on risks: whether data are considered anonymous is explicitly dependent on who controls the data and why. Said otherwise, "[t]his means that anonymised data disclosed within a secure local environment, e.g., when disclosed to a particular research organisation, could remain anonymous even though if published, the likelihood of re-identification would mean that the anonymised data would become personal data."

The former Article 29 Working Party's Opinion

All this seems pretty clear and relatively aligned. However, in 2014 the Article 29 Working Party (now the European Data Protection Board) issued guidance on the topic in its opinion on anonymisation techniques ('the Opinion'), with the unfortunate effect of creating confusion. To be sure, the Working Party's task — to generate consensus between 28 Data Protection Authorities and amongst them, computer scientists, and lawyers — was not easy. However, this was reflected in the result, which has been heavily criticised.

One passage in the Opinion was particularly problematic. Reformulating the test for anonymisation, the Working Party wrote: "More precisely, the data must be processed in such a way that it can no longer be used to identify a natural person by using 'all the means likely reasonably to be used' by either the controller or a third party. An important factor is that the processing must be irreversible."

The Working Party also added that "it is critical to understand that when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data." Such statements are hardly compatible with a risk-based approach.

(Continued on page 4)

As with any political document, one should probably read in between the lines of the Opinion. The fact that it was heavily debated is confirmed by the insertion of contrasting statements such as "the Working Party has therefore already clarified that the 'means... reasonably to be used' test is suggested by the Directive as a criterion to be applied in order to assess whether the anonymisation process is sufficiently robust, i.e. whether identification has become 'reasonably' impossible."

Approaches to anonymisation

Based on the author's informal discussion with national DPAs (or Supervisory Authorities as they are now known), it seems that two approaches are possible: either an approach based on the Opinion, which requires assessing and mitigating three types of re-identification risks (singling out, linkability, and inference) or a broader approach based on risks.

Such a view seems reassuring. What remains to be determined is how to appropriately conduct a risk-based approach. This is where tensions emerge again. While the ICO's approach in its Code of Practice seems sensible, it is arguable whether it will always — or, at the very least, often — be properly implemented.

Notable cases

One case in point is the Queen Mary University case (*Queen Mary University of London v Information Commissioner*, August 2016). The background facts concern a freedom of information request for clinical trial patient data that had been collated by researchers from Queen Mary University of London ('QMUL') working on the PACE trial investigating treatments for chronic fatigue syndrome. This request was ultimately rejected by QMUL after the results of an internal review. That rejection was then appealed to the ICO, which ordered QMUL to disclose to the complainant the information.

After a lengthy and divided opinion, the First-Tier Tribunal ultimately agreed with the ICO. Why did the ICO order the disclosure? Because the data should have been considered anonymised, as per the application of the 'motivated intruder' test. Yet, there appeared to be a sharp division between experts from both camps as regards the way such a test should be applied in context. This is because the motivated intruder test relies upon two fundamental assumptions, irrespective of the degree of sensitiveness of the data: firstly, a motivated intruder does possess prior knowledge, and therefore professionals bound by confidentiality obligations should not be considered as motivated intruders; and secondly, a motivated intruder is reasonably competent. In addition, the data at hand were still individual-level data. QMUL was therefore of the view that the data could only be pseudonymised and not anonymised.

"Generally speaking, one major concern as regards the barnardisation technique is that it does not on its own mitigate all re-identification risks, which creates problems when the data are released in a barnardised state without additional controls."

The First-Tier Tribunal, however, refused to question the robustness of the anonymisation technique. Would the First-Tier Tribunal's approach have been the same under the GDPR? The question is worth asking as the GDPR expressly defines pseudonymisation in its Article 4, and the prevailing view among Supervisory Authorities seems to be that pseudonymised data should still be considered personal data, as hinted by GDPR Recital 26. This is not to say that the ICO would necessarily agree with this view, as it had identified two routes leading to anonymisation in 2012.

As a matter of principle, re-identification depends upon access to additional information. And if access is made impossible or very hard, there is an argument that data that have undergone pseudonymisation could be considered anonymised. However — and this is crucial — additional controls such as legal obligations, access control and training, will have to be put in place to make the case that access to additional information becomes a remote possibility.

Another case worth evoking is the 2008 case of *Common Services Agency versus Scottish Information Commissioner*, although it predates both the ICO Code of Practice and the GDPR. The Supreme Court in this case did not directly solve the issue whether the data at hand should be considered personal data once the technique of barnardisation had been applied on them, but doubts had been raised.

Barnardisation versus differential privacy, etc.

Generally speaking, one major concern as regards the barnardisation technique is that it does not on its own mitigate all re-identification risks, which creates problems when the data are released in a barnardised state without additional controls. Contrary to techniques such as differential privacy or even k-anonymity, banaradisation does not even provide formal (mathematical) guarantees against re-identification risks.

Differential privacy, for example, is a much stronger technique. It limits the effect that an individual record can have on the output. Further, it randomises the output around the true value, and ensures that whatever the possible randomized outputs are, they are almost as likely on versions of the database that differ from this

one by the addition or removal of any single row. This limits the ability to infer the content of the database from a differentially private analysis result or output. Even in the unlikely event that one knows every other row in the database, and has a differentially private result over the database, it is almost equally likely to have received this differentially private analysis from versions of the database that include or do not include the record in question.

Surprisingly, the Working Party seemed to reject that differential privacy could eventually lead to anonymisation in its Opinion. This is because it had started from a very restrictive assumption: as long as the initial raw data are not destroyed, anonymisation is impossible, whatever the hands in which the queries sit.

Final remarks

Anonymising data in accordance with a risk-based approach, and claiming that the data are rendered outside the scope of data protection law, requires skills and resources. The starting point should be a good technical grasp of a range of anonymisation techniques.

Differential privacy — a technique based on the injection of randomised noise — is one of the strongest and as such should always be considered, and as early as possible. This is the case, in particular, if the controller is interested in deriving aggregates (averages, sums, counts, minima, maxima) from its queries.

The second lesson is that even if techniques such as differential privacy are used, it is likely that in many instances a 'release and control' model will be preferable to a 'release and forget' model. This is true even in the UK, where section 172 of the Data Protection Act 2018 makes it an offence to 'knowingly or recklessly re-identify information that is de-identified personal data without the consent of the controller responsible for de-identifying the personal data.' This is because technological solutions evolve constantly and combining different types of controls ensures a higher degree of effectiveness and therefore prevention.

While legal controls are certainly important, process and system controls are at least as important. With process and system controls, it is possible to allocate roles within organisations, restrict data access accordingly, apply sanitisation techniques directly upon the data, as well as detect anomalies through monitoring and auditing. What is more, when process and system controls are combined together, controls become scalable and organisations are finally in a position to accelerate their workflows. Said otherwise, controls become innovation enablers.

Sophie Stalla-Bourdillon
Immuta
sstalla-bourdillon@immuta.com