Anonymisation — a process living on borrowed time?

In the context of the Article 29 Working Party's Opinion on Anonymisation Techniques (WP216), Kate Brimsted, of Counsel at Reed Smith LLP (with Katalina Chin contributing) discusses why developments in Open Data and Big Data are driving an unprecedented need for reliable anonymisation techniques, whilst at the same time eroding their effectiveness

It seems to be a truth universally acknowledged that if some data sharing is good, then the sharing of so called 'Big Data' must be even better. Doing so can allow us to identify and act upon trends, plan smarter cities, reduce energy consumption, and enhance disease prevention and public health.

Incredibly fast computer processing speeds mean that data crunching can be accomplished on a scale never before achieved, and this continues to rise in line with Moore's Law (the observation that, over the history of computing hardware, the number of transistors in a dense integrated circuit doubles approximately every two years). Clearly, this process can benefit humanity collectively — but what about individuals' rights to be 'left alone', and to protect their privacy? Some of the information feeding the Big Data engine will be about people, or relate to them in some way. How can we achieve a balance between openness and privacy when it comes to information relating to living, identifiable people (i.e. 'personal data'), not forgetting that the scope of what constitutes 'personal data' is expanding?

A tantalisingly elegant solution presents itself: sever the link between the individuals and their data, and you can side step the Gordian complexity of EU data protection law. Thus we look to data custodians to cast the magic spell of anonymisation on the source data before sending them out into the world.

However, just as a certain unreliable coach turned back into a pumpkin, anonymised data are starting to exhibit troubling signs of becoming 're-identified' or at least 're-identifiable'. At the risk of labouring the fairy tale metaphor, we may not have the chimes of midnight to warn us, but it is clear that — due to the speed of technological advances in analytics — the data reversibility clock is ticking. Anonymisation — the solution and the price to pay to unlock the broader utility present in massive data sets — is getting more and more difficult to carry out with confidence.

It is reassuring, therefore, to hear from a body of EU data protection regulators that, despite the existence of some residual risk of identification, anonymisation techniques still have an important contribution to make.

The Opinion

Anonymisation was a welcome topic for the Article 29 Working Party to consider and produce guidance on. In Opinion 05/2014 on Anonymisation Techniques (WP216) (the 'Opinion') (copy available at www.pdpjournals.com/docs/88197), the Working Party has set out its analysis of the effectiveness and limits of common anonymisation techniques, and has provided recommendations to organisations that use anonymous data for various purposes.

The two most common anonymisation techniques, generalisation and randomisation, are addressed in the Opinion, and their effectiveness is assessed using three risk criteria, described below. However, the Working Party's efforts do not dispel the impression that this is a highly complex field, both legally and scientifically. In this respect, the UK regulator's Anonymisation Code of Practice, published in December 2012 (www.pdpjournals.com/docs/88198), appears to be a more pragmatic and user-friendly blueprint.

One of the Opinion's central themes is that, due to the fact that anonymisation and re-identification techniques are active areas of research, data controllers should pay attention to the state of technology and assess the use and adequacy of their processes in the light of this.

The Opinion does not identify any single technique that offers a sufficient guarantee of anonymity, and therefore recommends using a combination of the techniques that are available, keeping in mind the objectives and the context of the process.

The challenge of this subject, both for data controllers and regulators, is tacitly acknowledged by the Working Party with the statement that 'legal regulations…must therefore be formulated in a technologically neutral

(Continued on page 4)

(Continued from page 3)

manner and ideally take into account the changes in the developing potentials of information technology.'

The Working Party concludes that, in many cases, even an anonymised data set can represent residual risk to data subjects — and that can be true even if it is no longer possible to precisely retrieve the record of an individual, since it may still be possible to glean information about that individual with the help of other sources (whether public or not).

When are data 'identified' or 'identifiable'?

For nearly two decades, the Data Protection Directive 95/46/EC (the 'Directive') and national legislation that implements it across the EU Member States, have regulated the use of data that, either in and of themselves or together with 'means likely reasonably to be used' by a controller or any other person (Recital 26 of the Directive), allow an individual to be identified. However, over that time, the full value and potential uses of data have become more apparent and, as a result, the use of and need for anonymisation techniques has risen. As mentioned above, by anonymising data, organisations can remove or alter data categories that would otherwise allow a data subject to be identified, thereby taking any processing of those data outside of the scope of the Directive.

The Opinion begins by noting that there are specific references to the use of anonymisation in the Directive and the e-Privacy Directive (Directive 2002/58/EC), and that while there is no prescribed technique that should be used, the concept is defined in an outcome-focussed manner, requiring that the data 'be stripped of sufficient elements such that the data subject can no longer be identified'. When considering whether an individual may be identified, the Directive states that regard must be had to 'all' of the means 'likely reasonably' to be used to identify individuals by either the data controller or a third party. On this point, the Opinion advises that data controllers should consider the 'current state of technology', and that the risks need to be reassessed regularly. In addition, it acknowledges that there is a residual risk of re-identification inherent in any anonymisation process.

> ***"However, just as a certain unreliable coach turned back into a pumpkin, anonymised data are starting to exhibit troubling signs of becoming 're-identified' or at least 're-identifiable'"***

The Opinion also suggests that the 'likelihood' of re-identification does not depend on the intention of the discloser or recipient. That appears to imply that where there is a limited release of data between two contracting parties (for example), contractual restrictions imposed on the recipient would not be relevant to an assessment of 'likelihood of re-identification'. That sounds rather counter-intuitive.

Processing data to achieve anonymity

Due to the very broad scope of 'processing', even the act of applying anonymisation techniques to personal data (and thereby changing them from 'personal data' to mere 'data') is regulated under the Directive. This triggers the need to ensure that the organisation responsible meets the requirements of Article 6 and Article 7 of the Directive.

Providing that the processing of personal data has taken place in accordance with the applicable law, the further processing necessary in order to produce anonymous data will be considered to be compatible with the original purposes of processing, and therefore will not violate the Article 6(b) further processing principle. The Opinion also states that the legal basis for anonymisation can be found in any of the Article 7 grounds, including the data controller's legitimate interests.

Article 6 of the Directive lays down various requirements governing the quality of processed data. Amongst these, Article 6(1)(e) states that data should be 'kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed.' So, it can be said that the role of anonymisation in data protection compliance was identified from the outset of the Directive. To this end, the Opinion states that 'if [a] data controller wishes to retain such personal data once the purposes of the original or further processing have been achieved, anonymisation techniques should be used so as to irreversibly prevent identification.' The Working Party therefore states that further processing in order to anonymise data can be considered to be compatible with the original purpose of processing, but only insofar as the anonymisation process is such as to reliably produce anonymised information.

Developing an anonymisation process

Recalling Opinion 4/2007 (WP 136) on 'determining personal data', the Working Party advises that, when considering whether a person is 'identified or identifiable', the 'means likely reasonably to be used' test in the Directive is a criterion in assessing whether the anonymisation process is adequate to ensure the required outcome. In considering this question, the Opinion suggests some factors to take into consideration:

- data controllers should focus on what would be necessary in order to reverse the anonymisation technique, including the resources and technical knowledge necessary, the likelihood of this taking place and the severity of consequences

if it did take place. In this regard, the Opinion notes that the risk of identification may increase over time;

- the 'means likely reasonably to be used' criterion refers to the means to be used by the data controller or any other person. Therefore, if the data controller retains original, identifiable data at event-level, and provides a third party with a part of those data which have been treated to remove or mask identifiable data, that data set will still be classed as 'personal data' and therefore it will still be governed by the Directive. However, if the data controller were to delete the raw data and aggregate the original data to a level where individual events could no longer be identified, the resulting data set being released would be deemed anonymous. (Thanks to the UK House of Lords' decision in *Common Services Agency v Scottish Information Commission* ([2008] UKHL 47), which is also cited in the Information Commissioner's Office Anonymisation Code, the UK appears to take a more pragmatic approach, and would not require the deletion of the raw data in order to treat the data set being released as anonymous);
- third parties which obtain anonymous data may process them lawfully provided that they cannot directly or indirectly identify the original data subjects. To this end, third parties must take into account the possibility of identifying data subjects, at which point those personal data would fall within the ambit of the Directive.

Techniques

Several anonymisation techniques are examined in the Opinion, the most notable being randomisation and generalisation. (These are explained in some detail in the Annex to the Opinion, which offers a primer on anonymisation techniques).

Randomisation is explained as a technique that alters the veracity of the data to remove the strong link between the data and the individual, so that they can no longer be referred to a specific person. It preserves the singularity of the record, as it means each record derives from a single data subject. This may be done by shuffling the values in a table so that some are artificially linked to different data subjects. Generalisation, as the name suggests, involves generalising or pooling the attributes of data subjects, for example, by presenting combined data on a region, rather than a city. When considering which technique (or combination) is or are appropriate in the circumstances, and the guarantees that technique will achieve, the Opinion sets out three risks that must be considered:

Singling out — the possibility of isolating some or all records which identify an individual in the dataset. (This is why pseudonymisation does not equate to anonymisation — see below).

Linkability — the ability to link at least two records concerning the same subject or a group of data subjects.

Inference — the ability to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.

The analysis concludes that no single technique offers a guarantee of anonymity; data controllers should consider the limitations of the techniques before devising an anonymisation process. Using a combination of techniques may enhance the effectiveness of the process, but ultimately the assessment must be made on a case-by-case basis. Where the techniques proposed do not meet one of the criteria set out above, the Working Party instructs data controllers to evaluate the identification risks.

Pseudonymisation

In addition, the Opinion also addresses the technique of pseudonymisation, which typically involves replacing identifying fields with artificial identifiers. Whilst it is acknowledged that the technique is a useful security measure, the Working Party clarifies that it is not a method of anonymisation. Despite this point, it is clear that the technique is becoming increasingly useful; the text of the proposed General Data Protection Regulation that was voted on by the European Parliament in March 2014 contains specific provisions for its use in relation to profiling and the processing of personal data concerning health.

Concluding thoughts

Despite the slightly daunting 'primer' annexed to the Opinion, we should banish the notion that anonymisation is merely the stuff of statisticians and too esoteric to be an issue of importance. The drive for anonymisation as a gateway for the release of public and private sector data is on the rise. But this should be accompanied by a realistic appreciation of its limitations, which are also becoming more apparent — it is not merely research in de-identification which is a thriving and active area.

The Working Party advocates the use of anonymisation techniques and also advises data controllers to be vigilant and consider the changes in techniques and risks over time. This is understandable, but not an easy responsibility for data controllers to discharge. For one thing, it is not reasonable to expect 20:20 hindsight. Once a data set has been released into the public domain, it is not feasible to recover it. The data genie is out of the bottle.

Rather than dwell on the difficulties and bemoan the future fragility of anonymisation as a solution for protecting individuals when facilitating 'Open Data' initiatives, we could view the challenges as a sign that we need a new data protection paradigm. A new model could focus on risk and the effect on individuals of their data use, rather than struggling on with concepts of 'identity', which look increasingly irrelevant in the face of widespread, sophisticated profiling techniques. Perhaps the time has come for legislators to do something truly radical. Now, where to find that glass slipper?

Kate Brimsted
(with contribution by Katalina Chin)
Reed Smith LLP
kbrimsted@reedsmith.com