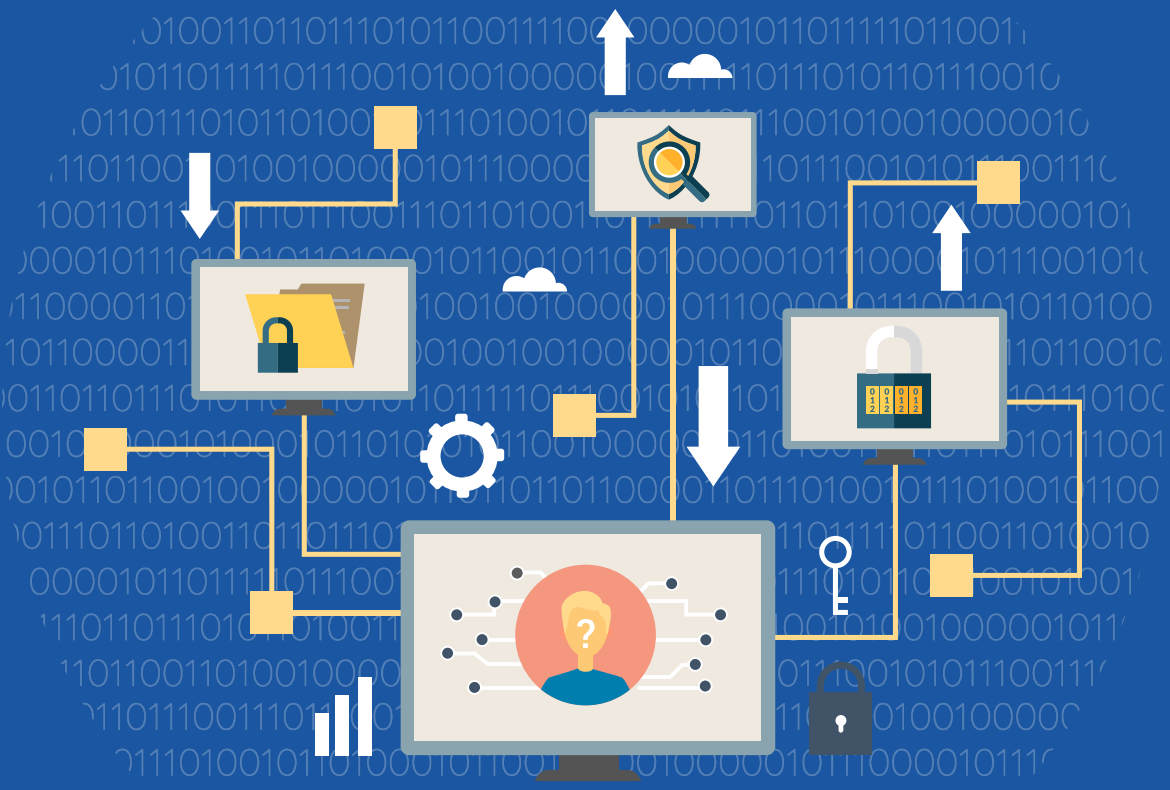
aepd agencia española protección datos

EDPS



10

MISUNDERSTANDINGS RELATED TO

ANONYMISATION

Anonymisation is the process of rendering personal data anonymous.

According to the European Union's data protection laws, in particular the **General Data Protection Regulation (GDPR)**[1], anonymous data is "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". **Datasets**[2] which include personal data may contain direct and indirect identifiers, which allow an individual to be identified or become identifiable. A **direct identifier** is specific information that references to an individual, such as name or an identification number. An **indirect identifier** (also called quasi-identifier) is any piece of information (e.g. a geographical position in a certain moment or an opinion about a certain topic) that could be used, either individually or in combination with other quasi-identifiers, by someone that has knowledge about that individual with the purpose of re-identifying an individual in the dataset[3] [4]. The **re-identification likelihood** is the probability in a given dataset of re-identifying an individual, by turning anonymised data back into personal data through the use of data matching or similar techniques. The **utility of a dataset** is a measure of how useful that information is for the intended purpose (e.g. a research study on a specific disease).

Throughout the years, there have been several **examples of incomplete or wrongfully conducted anonymisation processes** that resulted in the re-identification of individuals. In 2006, a movie-streaming service, for instance, published a dataset containing 10 million movie rankings made by 500,000 customers claiming that it was anonymous, but it was later found that it would only take a little bit of knowledge about the subscriber for an adversary to be able to identify that subscriber's record in the dataset[5]. Another example of deficient anonymisation: in 2013, the New York City Taxi and Limousine Commission published a data sheet with more than 173 million individual taxi trips containing the pickup and drop-off location, times and supposedly anonymised licence numbers. The dataset had not been correctly anonymised, and it was possible to identify the original licence numbers and even the individual drivers of those taxis[6].

Anonymous data play an important role in the context of research in the fields of medicine, demographics, marketing, economy, statistics and many others. However, this interest coincided with the spread of related misunderstandings. The objective of this document is to raise public awareness about some misunderstandings about anonymisation, and to motivate its readers to check assertions about the technology, rather than accepting them without verification.

This document lists ten of these misunderstandings, explains the facts and provides references for further reading.

---

1 http://data.europa.eu/eli/reg/2016/679/2016-05-04.

2 A dataset is a structured collection of data. A table where each column represents a particular variable and each row corresponds to a different record is an example of a dataset.

3 Barth-Jones, D. (2012). The 're-identification' of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now. Then and Now (July 2012). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397.

4 Khaled El Emam and Bradley Malin, "Appendix B: Concepts and Methods for De-identifying Clinical Trial Data," Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk (Washington D.C.: National Academies Press, 2015), http://www.ncbi.nlm.nih.gov/books/NBK285994.

5 Narayanan, A., & Shmatikov, V. (2006). How to break anonymity of the Netflix prize dataset. arXiv preprint cs/0610105. https://arxiv.org/abs/cs/0610105.

6 Pandurangan, V. (2014). On taxis and rainbows: Lessons from NYC's improperly anonymized taxi logs. Medium. Accessed November, 30, 2015. https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1.

MISUNDERSTANDING 1.

"Pseudonymisation is the same as anonymisation"

Fact: Pseudonymisation is not the same as anonymisation.

The GDPR defines 'pseudonymisation' as 'the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person'. This means that the use of 'additional information' can lead to the identification of the individuals, which is why pseudonymous personal data is still personal data.

Anonymous data, on the other hand, cannot be associated to specific individuals. Once data is truly anonymous and individuals are no longer identifiable, the data will not fall within the scope of the GDPR.

MISUNDERSTANDING 2.

"Encryption is anonymisation"

Fact: Encryption is not an anonymisation technique, but it can be a powerful pseudonymisation tool.

The encryption process uses secret keys to transform the information in a way that reduces the risk of misuse, while keeping confidentiality for a given period of time. Because the original information needs to be accessible, the transformations applied by encryption algorithms are designed to be reversible, in what is known as decryption. The secret keys used for decryption are the aforementioned 'additional information' (see Misunderstanding 1), which can make the personal data readable and, consequently, the identification possible.

The secret keys used for decryption are the aforementioned 'additional information' (see Misunderstanding 1), which can make the personal data readable and, consequently, the identification possible.

In theory, it could be considered that deleting the encryption key of encrypted data would render it anonymous, but this is not the case. One cannot assume that encrypted data cannot be decrypted because the decryption key is said to be "erased" or "unknown". There are many factors affecting the confidentiality of encrypted data, especially in the long term. Among these factors are the strength of the encryption algorithm and of the key, information leaks, implementation issues, amount of encrypted data, or technological advances (e.g. quantum computing[7]).

7 TechDispatch #2/2020: Quantum Computing and Cryptography, 7 August 2020, European Data Protection Supervisor https://edps.europa.eu/data-protection/our-work/publications/techdispatch/techdispatch-22020-quantum-computing-and_en

MISUNDERSTANDING 3.

"Anonymisation of data is always possible"

Fact: It is not always possible to lower the re-identification risk below a previously defined threshold whilst retaining a useful dataset for a specific processing.

Anonymisation is a process that tries to find the right balance between reducing the re-identification risk and keeping the utility of a dataset for the envisaged purpose(s). However, depending on the context or the nature of the data, the re-identification risks cannot sufficiently mitigated. This could be the situation when the total number of possible individuals ('universe of subjects') is too small (e.g. an anonymous dataset containing only the 705 members of the European Parliament), when the categories of data are so different among individuals that it is possible to single these individuals out (e.g. device fingerprint of the systems that accessed a certain website) or when the case of datasets include a high number of demographic attributes[8] or location data[9].

MISUNDERSTANDING 4.

"Anonymisation is forever"

Fact: There is a risk that some anonymisation processes could be reverted in the future. Circumstances might change over time and new technical developments and the availability of additional information might compromise previous anonymisation processes.

The computing resources and new technologies (or new ways to apply existing technologies) available to an attacker that could try to re-identify an anonymous dataset change overtime. Nowadays, cloud computing provides affordable computing capability to levels and prices that were unthinkable years ago. In the future, quantum computers might also alter what is nowadays considered "reasonable means" [10].

Also, the disclosure of additional data over the years (e.g. in a personal data breach) can make it possible to link previously anonymous data to identified individuals. The release of many decades old records containing highly sensitive data (e.g. criminal records) could still have a severely detrimental effect on an individual or relatives[11] .

8 Rocher, L., Hendrickx, J. M., & De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1), 1-9, https://doi.org/10.1038/s41467-019-10933-3

9 Xu, F., Tu, Z., Li, Y., Zhang, P., Fu, X., & Jin, D. (2017, April). Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data. In Proceedings of the 26th international conference on world wide web (pp. 1241-1250), https://dl.acm.org/doi/abs/10.1145/3038912.3052620

10 EDPS TechDispatch - Quantum computing and cryptography. Issue 2, 2020, https://data.europa.eu/doi/10.2804/36404

11 Graham, C. (2012). Anonymisation: managing data protection risk code of practice. Information Commissioner's Office. https://ico.org.uk/media/1061/anonymisation-code.pdf .

MISUNDERSTANDING 5.

"Anonymisation always reduces the probability of re-identification of a dataset to zero"

Fact: The anonymisation process and the way it is implemented will have a direct influence on the likelihood of re-identification risks.

A robust anonymisation process aims to reduce the re-identification risk below a certain threshold. Such threshold will depend on several factors such as the existing mitigation controls (none in the context of public disclosure), the impact on individuals' privacy in the event of re-identification, the motives and the capacity of an attacker to re-identify the data[12].

Although a 100% anonymisation is the most desirable goal from a personal data protection perspective, in some cases it is not possible and a residual risk of re-identification must be considered.

MISUNDERSTANDING 6.

"Anonymisation is a binary concept that cannot be measured"

Fact: It is possible to analyse and measure the degree of anonymization.

The expression "anonymous data" cannot not be perceived as if datasets could simply be labelled as anonymous or not. The records in any dataset have a probability of being re-identified based on how possible it is to single them out. Any robust anonymisation process will assess the re-identification risk, which should be managed and controlled over the time[13].

Except for specific cases where data is highly generalised (e.g. a dataset counting the number of visitors of a website per country in a year), the re-identification risk is never zero.

12 External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use (2016) https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data_en-0.pdf

13 Step 4: Measure the data risk. De-identification Guidelines for Structured Data, Information and Privacy Commissioner of Ontario June 2016. https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf

MISUNDERSTANDING 7.

"Anonymisation can be fully automated"

Fact: Automated tools can be used during the anonymisation process, however, given the importance of the context in the overall process assessment, human expert intervention is needed.

On the contrary, it requires an analysis of the original dataset, its intended purposes, the techniques to apply and the re-identification risk of the resulting data[14].

The identification and deletion of direct identifiers (also known as 'masking'), while being an important part of the anonymisation process, must always be followed by a cautious analysis for other sources of (indirect) identification[15] (generally through quasi-identifiers). While direct identifiers are somewhat trivial to find, indirect identifiers, on the other side, are not always obvious, and the failure to detect them can result in the reversion of the process (i.e. re-identification), with consequences for the privacy of individuals.

Automation could be key for some steps of the anonymisation process, such as the removal of direct identifiers or the consistent application of a generalisation procedure over a variable.[16] On the contrary, it seems unlikely that a fully automatised process might identify quasi-identifiers in different contexts or decide how to maximise data utility by applying specific techniques to specific variables.

14 Recommendation section (5.2) of Article 29 Data Protection Working Party. (2014). Opinion 05/2014 on Anonymisation Techniques. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

15 Guess Who? 5 examples why removing names fails as anonymization, https://www.syntho.ai/5-examples-why-removing-names-fails-as-anonymization

16 See for examples e.g. F. Diaz, N. Mamede, J. Baptista (2016), Automated Anonymization of Text Documents, https://www.hlt.inesc-id.pt/~fdias/mscthesis/automated_text_anonymization.pdf

MISUNDERSTANDING 8.

"Anonymisation makes the data useless"

Fact: A proper anonymisation process keeps the data functional for a given purpose.

The purpose of anonymisation is to prevent individuals in a dataset from being identified. The anonymisation techniques will always restrict the ways in which the resulting dataset can be used. For example, grouping dates of birth into year intervals will reduce the re-identification risk while at the same time reducing the dataset utility in some cases. This does not mean that anonymous data will become useless, but rather that its utility will depend on the purpose and the acceptable re-identification risk.

On the other hand, personal data cannot be permanently stored beyond its original purpose, waiting for a chance where it might become useful for other purposes. The solution for some controllers might be anonymisation, where personal data can be detached and discarded from the dataset, while the remaining dataset still retains a useful meaning. An example could be the anonymisation of access logs of a website, by keeping only the access date and accessed page, but not the information on who accessed it.

The "data minimisation" principle requires the controller to determine if it is necessary to process personal data in order to fulfil a particular purpose, or if that purpose can also be achieved with anonymous data.

In certain cases, this might lead to the conclusion that rendering the data anonymous will not fit the intended purpose. In such cases, the controller will have to choose between processing personal data (and use e.g. pseudonymisation) and

apply the GDPR, or not to process the data at all.

MISUNDERSTANDING 9.

"Following an anonymisation process that others used successfully will lead our organisation to equivalent results"

Fact: Anonymisation processes need to be tailored to the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons.

Anonymisation cannot be applied akin to following a recipe, because the context (nature, scope, context and purposes of the processing of the data) are likely different from one circumstance to another, and from one organisation to another. An anonymisation process might have a re-identification risk below a certain threshold when the data is only made available to a limited number of recipients, whereas the re-identification risk will not be able to meet that threshold when the data is made available to the general public.

Different datasets might be available in different contexts. These could be cross-referenced with the anonymous data affecting the re-identification risk. For example, in Sweden, details of taxpayers' personal data are publicly available, while in Spain they are not. Therefore, even if datasets including information of Spanish and Swedish citizens would be anonymised following the same procedure, the re-identification risks could be different.

MISUNDERSTANDING 10.

"There is no risk and no interest in finding out to whom this data refers to"

Fact: Personal data has a value in itself, for the individuals themselves and for third parties. Re-identification of an individual could have a serious impact for his rights and freedoms.

Attacks against anonymisation can be either deliberate attempts at re-identification, unintended attempts at re-identification, data breaches or releasing data to the public[17]. The likelihood of someone trying to re-identify an individual only touches upon the first type. The possibility of someone re-identifying at least one person in a dataset, be it out of curiosity, by chance or driven by an actual interest (e.g. scientific research, journalism or criminal activity) cannot be disregarded[18].

It can be difficult to accurately assess the impact of re-identification on a person's private life, because it will always depend on the context and on the information that is correlated. For example, the re-identification of a data subject in the context of the seemingly harmless context of his or her movie preferences might lead to inferring about that person's political leanings or sexual orientation[19]. Such particularly sensitive data are however accorded special protection under the GDPR.

17 Khaled El Emam and Luk Arbuckle, Anonymizing Health Data (p. 29-33).

18 Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, Bradley Malin, "A Systematic Review of Re-Identification Attacks on Health Data", 11 December 2011.

19 Narayanan, Arvind; Shmatikov, Vitaly. "Robust De-anonymization of Large Sparse Datasets" (PDF). Retrieved 2 March 2021. https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf .