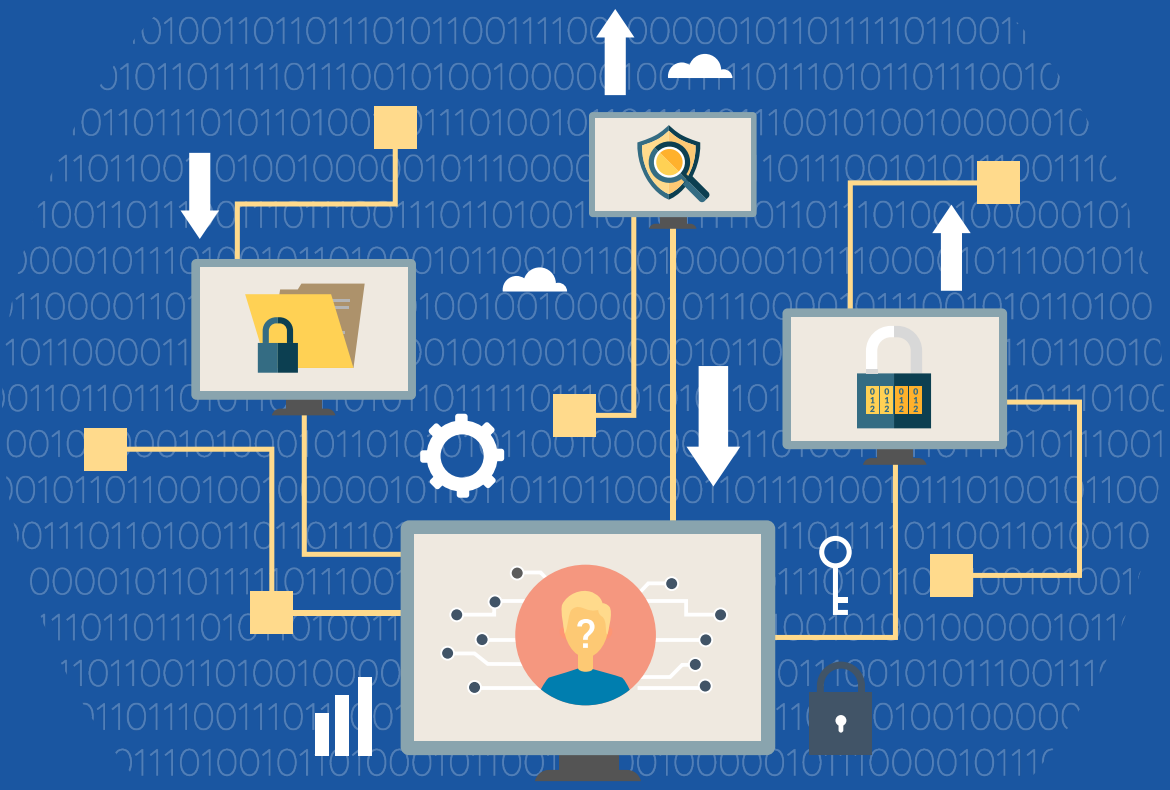
aepd agencia española protección datos

EDPS



10

MALENTENDIDOS RELACIONADOS CON LA ANONIMIZACIÓN

La **anonimización** es el proceso mediante el cual los datos personales se convierten en anónimos.

De conformidad con la legislación en materia de protección de datos de la Unión Europea, en concreto, el **Reglamento general de protección de datos (RGPD)**[1], los datos anónimos constituyen «aquella información que no hace referencia a personas naturales identificadas o identificables o a datos personales que se anonimizan de tal forma que dejan de ser identificables». Los **conjuntos de datos**[2] que incluyen datos personales pueden contener identificadores directos e indirectos, lo que permite que se identifique o que pueda identificarse a una persona física. Un **identificador directo** es la información específica que puede atribuirse a un individuo, como su nombre o un número de identificación. Un **identificador indirecto** (también denominado cuasi-identificador) es cualquier dato (por ejemplo, una situación geográfica en un momento determinado o una opinión sobre un tema en particular) que podría utilizarse, ya sea de forma individual o combinada con otros cuasi-identificadores, por alguien que posea conocimientos sobre ese individuo con el fin de reidentificarle en el conjunto de datos[3] [4]. La **probabilidad de reidentificación** es la probabilidad de que se reidentifique a un individuo en un conjunto de datos determinado mediante la conversión de datos anonimizados en datos personales a través del uso de la comparación de datos o de técnicas similares. La **utilidad de un conjunto de datos** es una unidad que mide la utilidad de esa información para un propósito determinado (por ejemplo, un estudio de investigación sobre una enfermedad específica).

A lo largo de los años, ha habido varios **ejemplos de procesos de anonimización que se han llevado a cabo de forma incompleta o errónea**, lo que supone la reidentificación de los individuos. Por ejemplo, en 2006 un servicio de visionado en streaming de películas publicó un conjunto de datos que contenía 10 millones de clasificaciones de películas realizadas por 500 000 clientes alegando que era anónimo, pero posteriormente se descubrió que era suficiente con saber unos pocos datos sobre el suscriptor para poder identificarlo en el registro de ese conjunto de datos[5]. Otro ejemplo de anonimización deficiente: en 2013, la Comisión de Taxis y Limusinas de la ciudad de Nueva York publicó una ficha de datos con más de 173 millones de viajes individuales en taxi que contenían la ubicación de recogida y destino, los horarios y los números de licencia supuestamente anonimizados. El conjunto de datos no se anonimizó de forma correcta, con lo cual, era posible identificar los números de licencia originales e, incluso, a los conductores de dichos taxis[6].

Los datos anónimos desempeñan un papel importante en el contexto de la investigación en áreas como la medicina, demografía, marketing, economía, estadística y muchas otras. Sin embargo, este interés ha supuesto la difusión de malentendidos al respecto. El objetivo del presente documento es **sensibilizar al público sobre algunos malentendidos relacionados con la anonimización** y motivar a sus lectores para que comprueben las afirmaciones sobre la tecnología, en lugar de aceptarlas sin una verificación.

El presente documento incluye una lista de diez de estos malentendidos, explica la realidad y ofrece referencias para una lectura pormenorizada.

---

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

2 Un conjunto de datos es una recopilación de datos estructurada. Un ejemplo de conjunto de datos es una tabla en la que cada columna representa una variable concreta y cada fila corresponde a un registro diferente.

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 (julio 2012). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397.

4 Khaled El Emam y 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. Recuperado el, 30 de noviembre de, 2015. https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1.

EQUIVOCO 1.

«La seudonimización es lo mismo que la anonimización»

Realidad: «La seudonimización no es lo mismo que la anonimiza»

El RGPD define «seudonimización» como «el tratamiento de datos personales de manera que no puedan atribuirse a un interesado sin utilizar información adicional, siempre que dicha información adicional figure por separado y esté sujeta a medidas técnicas y organizativas destinadas a garantizar que los datos personales no se atribuyan a una persona física identificada o identificable». Esto significa que el uso de «información adicional» puede suponer la identificación de los individuos; por ese motivo los datos personales seudonimizados también son datos personales.

Por el contrario, los datos anónimos, no pueden asociarse con un individuo en particular. Una vez que los datos son realmente anónimos y los individuos dejan de ser identificables, dejan de estar incluidos en el ámbito de aplicación del RGPD.

EQUIVOCO 2.

«El cifrado es anonimización»

Realidad: El cifrado no constituye una técnica de anonimización, pero puede ser una buena herramienta de seudonimización.

El proceso de cifrado utiliza claves secretas para transformar la información de tal forma que se reduzca el riesgo de uso indebido y, al mismo tiempo, se mantenga la confidencialidad durante un periodo de tiempo determinado. Dado que la información original debe ser accesible, las transformaciones aplicadas por los algoritmos de cifrado están diseñadas para ser reversibles, lo que se conoce como descifrado.

Las claves privadas que se utilizan para el descifrado son la «información adicional», mencionada anteriormente (véase Equívoco 1), lo que puede hacer que los datos sean legibles y, en última instancia, que la identificación sea posible.

En teoría, podría considerarse que la eliminación de la clave de cifrado de los datos cifrados los convertiría en anónimos, pero no es así. No se puede dar por hecho que los datos cifrados no puedan descifrarse porque se diga que la clave de descifrado se ha «borrado» o es «desconocida». Hay muchos factores que afectan a la confidencialidad de los datos cifrados, en particular a largo plazo. Entre estos factores se encuentran la solidez del algoritmo de cifrado y de la clave, las fugas de información, los problemas de implantación, la cantidad de datos cifrados o los avances tecnológicos (por ejemplo, la computación cuántica[7]).

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

EQUIVOCO 3.

«Los datos siempre pueden anonimizarse»

Realidad: No siempre es posible reducir el riesgo de reidentificación por debajo de un umbral definido de forma previa y mantener, al mismo tiempo, la utilidad de un conjunto de datos para un tratamiento específico.

La anonimización es un proceso que trata de encontrar el equilibrio adecuado entre la reducción del riesgo de reidentificación y el mantenimiento de la utilidad de un conjunto de datos para los fines previstos. Sin embargo, en función del contexto o la naturaleza de los datos, los riesgos de reidentificación podrían no mitigarse lo suficiente. Esta situación puede darse cuando el número total de posibles individuos («universo de sujetos») es demasiado reducido (por ejemplo, un conjunto de datos anónimos que contenga sólo los 705 miembros del Parlamento Europeo), cuando las categorías de datos son tan diferentes entre los individuos que es posible individualizarlos (por ejemplo, la huella digital del dispositivo de los sistemas que accedieron a un determinado sitio web) o cuando el caso de los conjuntos de datos incluye un elevado número de atributos demográficos[8]o datos de localización[9].

EQUIVOCO 4.

«La anonimización es permanente»

Realidad: Existe un riesgo de que ciertos procesos de anonimización puedan revertirse en el futuro. Las circunstancias pueden cambiar a lo largo del tiempo y los nuevos avances técnicos y la disponibilidad de información adicional pueden poner en peligro los procesos de anonimización previos.

Los recursos informáticos y las nuevas tecnologías (o nuevos usos de tecnologías ya existentes) disponibles para un atacante que pudiera intentar reidentificar un conjunto de datos anónimos van cambiando a lo largo del tiempo. Hoy en día, la computación en la nube proporciona una capacidad de computación asequible a niveles y precios que eran impensables hace años. En el futuro, los ordenadores cuánticos también podrían alterar lo que en la actualidad se consideran «medios aceptables»[10].

Además, la divulgación de datos adicionales a lo largo de los años (por ejemplo, en una filtración de datos personales) puede permitir que los datos que anteriormente eran anónimos se atribuyan a personas identificadas. La divulgación de registros de muchas décadas de antigüedad que contengan datos muy sensibles (por ejemplo, antecedentes penales) podría continuar teniendo un efecto bastante perjudicial para un individuo o sus familiares[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. En Proceedings of the 26th international conference on world wide web (páginas 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 .

EQUIVOCO 5.

«La anonimización siempre reduce la probabilidad de reidentificación de un conjunto de datos a cero»

Realidad: El proceso de anonimización y la forma en que se aplique tendrán una influencia directa en la probabilidad de riesgos de reidentificación.

Un proceso de anonimización sólido tiene como objetivo la reducción del riesgo de reidentificación por debajo de un determinado umbral. Dicho umbral dependerá de varios factores, como los controles de mitigación existentes (ninguno en el contexto de la divulgación pública), la repercusión en la privacidad de los individuos en caso de reidentificación, los motivos y la capacidad de un atacante para reidentificar los datos[12].

Aunque una anonimización del 100% es el objetivo más deseable desde el punto de vista de la protección de los datos personales, en algunos casos no es posible y debe contemplarse un riesgo residual de reidentificación.

EQUIVOCO 6.

«La anonimización es un concepto binario que no puede medirse»

Realidad: El grado de anonimización puede analizarse y medirse.

La expresión «datos anónimos» no puede entenderse como si los conjuntos de datos pudieran etiquetarse como anónimos o no. Existe una probabilidad de que los registros de cualquier conjunto de datos se reidentifiquen en función de la posibilidad de individualizarlos. Cualquier proceso sólido de anonimización evaluará el riesgo de reidentificación, que debe gestionarse y controlarse a lo largo del tiempo[13].

Excepto en casos específicos en los que los datos estén muy generalizados (por ejemplo, un conjunto de datos que cuente el número de visitantes de un sitio web por país en un año), el riesgo de reidentificación nunca puede considerarse nulo.

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

EQUIVOCO 7.

«La anonimización puede automatizarse totalmente»

Realidad: Es posible utilizar herramientas automáticas durante el proceso de anonimización, pero, dada la importancia del contexto en la evaluación de dicho proceso, la intervención del experto humano es necesaria.

Al contrario, requiere un análisis del conjunto de datos original, sus fines previstos, las técnicas que deben aplicarse y el riesgo de reidentificación de los datos resultantes[14].

Pese a que la identificación y eliminación de los identificadores directos (también conocida como «enmascaramiento») constituye una parte importante del proceso de anonimización, debe ir siempre acompañada de un análisis cauteloso que busque otras fuentes de identificación (indirecta) [15] (en general, a través de cuasi-identificadores). Mientras que encontrar los identificadores directos es algo trivial, los identificadores indirectos, en cambio, no siempre son obvios, y el hecho de no detectarlos puede dar lugar a la reversión del proceso (es decir, la reidentificación), lo que tiene consecuencias para la privacidad de los individuos.

La automatización podría ser clave en algunos pasos del proceso de anonimización, como la eliminación de identificadores directos o la aplicación coherente de un procedimiento de generalización sobre una variable[16] . Por el contrario, parece poco probable que un proceso totalmente automatizado pueda identificar cuasi-identificadores en diferentes contextos o decidir cómo maximizar la utilidad de los datos aplicando técnicas específicas a variables concretas.

EQUIVOCO 8.

«La anonimización inutiliza los datos»

Realidad: Un proceso de anonimización adecuado mantiene la funcionalidad de los datos para un fin determinado.

El objetivo de la anonimización es evitar que se identifique a los individuos de un conjunto de datos. Las técnicas de anonimización siempre restringirán las formas en que se puede utilizar el conjunto de datos resultante. Por ejemplo, agrupar las fechas de nacimiento en intervalos de un año reducirá el riesgo de reidentificación y, al mismo tiempo, la utilidad del conjunto de datos en algunos casos. Esto no significa que los datos anónimos sean inútiles, sino que su utilidad dependerá de la finalidad y del riesgo de reidentificación que se acepte.

Por otra parte, los datos personales no pueden almacenarse de forma permanente más de lo que estipule su finalidad original, a la espera de una oportunidad en la que puedan resultar útiles para otros fines. La solución para algunos responsables del tratamiento podría ser la anonimización, en la que los datos personales pueden independizarse y desecharse del conjunto de datos, mientras que el conjunto de datos restante continúa conservando un significado útil. Un ejemplo podría ser la

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 Véase, por ejemplo, F. Diaz, N. Mamede, J. Baptista (2016), Automated Anonymization of Text Documents, https://www.hlt.inesc-id.pt/~fdias/mscthesis/automated_text_anonymization.pdf

anonimización de los registros de acceso a un sitio web, si solo se conserva la fecha de acceso y la página a la que se ha accedido, pero no la información sobre quién ha accedido.

El principio de «minimización de los datos» exige que el responsable del tratamiento determine si es necesario tratar los datos personales para cumplir un objetivo concreto, o si ese objetivo puede alcanzarse también con datos anónimos.

En algunos casos, esto puede conducir a la conclusión de que la anonimización de los datos no se ajusta a la finalidad prevista. En estos casos, el responsable del tratamiento tendrá que decidir entre tratar los datos personales (y utilizar, por ejemplo, la seudonimización) y aplicar el RGPD, o no tratar los datos de ninguna forma.

EQUIVOCO 9.

«Seguir un proceso de anonimización que otros utilizaron con éxito hará que nuestra organización obtenga resultados equivalentes»

Realidad: Los procesos de anonimización deben adaptarse a la naturaleza, el alcance, el contexto y los fines del tratamiento, así como a los riesgos de diversa probabilidad y gravedad para los derechos y libertades de las personas físicas.

La anonimización no puede aplicarse como si se siguiera una receta, porque el contexto (naturaleza, alcance, contexto y fines del tratamiento de los datos) probablemente difiera de una circunstancia a otra y de una organización a otra. Un proceso de anonimización puede tener un riesgo de reidentificación por debajo de un determinado umbral cuando los datos sólo se ponen a disposición de un número limitado de destinatarios, mientras que el riesgo de reidentificación no podrá alcanzar ese umbral cuando los datos se pongan a disposición del público en general.

Puede haber diferentes conjuntos de datos disponibles en diferentes contextos. Estos podrían cruzarse con los datos anónimos, lo que afectaría al riesgo de reidentificación. Por ejemplo, en Suecia, la información relativa a los datos personales de los contribuyentes está disponibles de forma pública, mientras que en España no lo están. Por tanto, aunque los conjuntos de datos que incluyen información de ciudadanos españoles y suecos se anonimizaran siguiendo el mismo procedimiento, los riesgos de reidentificación podrían ser diferentes.

EQUIVOCO 10.

«No existe un riesgo ni interés alguno en saber a quién se atribuyen estos datos»

Realidad: Los datos personales tienen un valor en sí mismos, para los propios individuos y para terceros. La reidentificación de un individuo podría tener una repercusión grave en lo relativo a sus derechos y libertades.

Los ataques contra la anonimización pueden materializarse en forma de intentos deliberados de reidentificación, intentos involuntarios de reidentificación, brechas de seguridad o divulgación de datos al público[17]. La probabilidad de que alguien intente reidentificar a un individuo solo se refiere al primer tipo. No se puede descartar la posibilidad de que alguien reidentifique al menos a una persona en un conjunto de datos, ya sea por curiosidad, por casualidad o por un interés real (por ejemplo, investigación científica, periodismo o actividad delictiva)[18].

Puede ser difícil evaluar con precisión el impacto de la reidentificación en la vida privada de una persona, porque siempre dependerá del contexto y de la información que se correlacione. Por ejemplo, la reidentificación de un interesado en el contexto aparentemente inofensivo de sus preferencias cinematográficas podría llevar a inferir sobre las inclinaciones políticas o la orientación sexual de esa persona[19].

Sin embargo, estos datos especialmente sensibles gozan de una protección especial en virtud del RGPD.

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 de diciembre de 2011. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0028071&type=printable

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