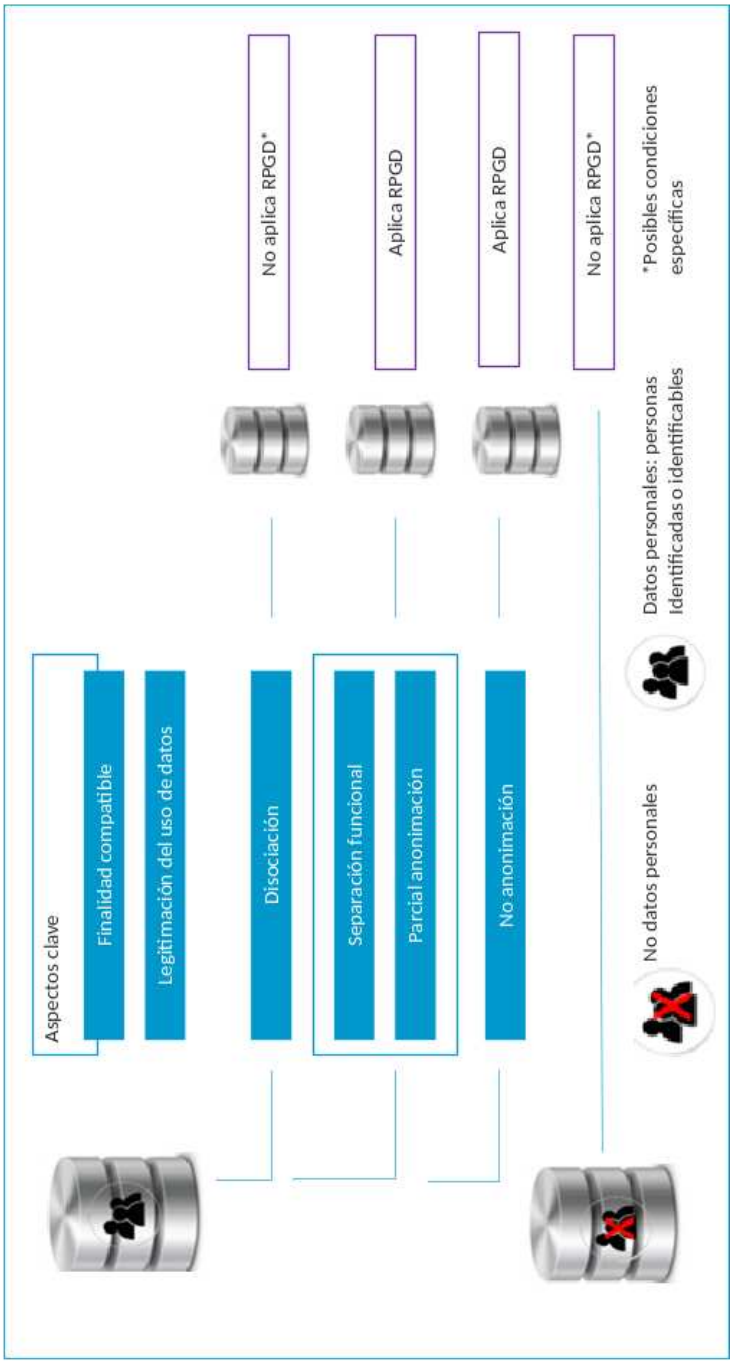
Gobierno del Dato y Toma de Decisiones

Tema 10. La disociación de datos personales y técnicas de anonimización

Índice

Esquema



10.1. Introducción y objetivos

En este tema vamos a introducir algunas técnicas de anonimización de datos y conoceremos los riesgos de dichas técnicas y, en general, de los procesos de anonimización, frente a la reidentificación de los individuos.

Este tema nos introduce en el mundo de la anonimización de datos, como respuesta a las necesidades del tratamiento de grandes volúmenes de datos, y en cómo es en el ámbito científico y estadístico o del *marketing* y de las implicaciones en materia de protección de datos.

Como herramienta fundamental, hablaremos de la disociación de datos, proceso que facilita la obtención de conjuntos de datos totalmente anonimizados (datos disociados, a partir de datos personales que no permiten la reidentificación de los individuos y sobre los que no es aplicable la normativa de protección de datos).

Hablaremos de técnicas de anonimización, de los riesgos asociados de reidentificación sobre un conjunto de datos, parcialmente anonimizados. También introduciremos las recomendaciones del **grupo de trabajo del artículo 29** sobre los conjuntos de datos disociados y las técnicas de anonimización.

10.2. Definiciones

- **Procedimiento de disociación:** «todo tratamiento de datos personales de modo que la información que se obtenga no pueda asociarse a persona identificada o identificable» (art. 3.f, LOPD 15/1999, de 13 de diciembre).
- **Seudonimización:** «el tratamiento de datos personales de manera tal que ya 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» (art. 4, RGPD).
- **Anonimización:** «proceso por el cual deja de ser posible establecer por medios razonables el nexo entre un dato y el sujeto al que se refiere. Es aplicable también a la muestra biológica» (art. 3.c, Ley 14/2007, de 3 de julio).
- **Dato agregado:** son datos tratados de manera que los valores representan acumulados o variables estadísticas como medias o han sido agrupados en rangos de datos.
- **Microdatos:** datos a nivel de registro que no han sido objeto de tratamiento de agregación.
- **Identificadores indirectos:** variables que combinadas pueden llegar a permitir la identificación directa, pero no son consideradas identificadores directos.
- **Seudónimos:** son identificadores personales que identifican a una persona, pero diferentes de los que habitualmente utiliza esa persona.
- **Cuasiidentificadores:** un conjunto de identificadores indirectos que combinados pueden identificar individuos, pero no lo hace tomados de manera independiente.
- **Supresión:** técnica que permite obtener un conjunto k-anonimato eliminando una

celda (valor) o un registro.

- **Deidentificación:** proceso de eliminación de identificadores, cuasiidentificadores, de manera que se dificulta la identificación de los individuos a menos que se cuenten con información adicional.

- **Riesgo de singularización:** la posibilidad de extraer de un conjunto de datos algunos registros (o todos los registros) que identifican a una persona.

- **Riesgo de vinculabilidad:** la capacidad de vincular como mínimo dos registros de un único interesado o de un grupo de interesados, ya sea en la misma base de datos o en dos bases de datos distintas. Si el atacante puede determinar (por ejemplo, mediante un análisis de correlación) que dos registros están asignados al mismo grupo de personas, pero no puede singularizar a las personas en este grupo, entonces la técnica es resistente a la singularización, pero no a la vinculabilidad.

- **Riesgo de inferencia:** la posibilidad de deducir con una probabilidad significativa el valor de un atributo a partir de los valores de un conjunto de otros atributos.

10.3. La disociación y anonimización de datos

El RGPD, en su considerando 26, indica que no es aplicable sobre datos anónimos.

> «Los principios de protección de datos no deben aplicarse a la información anónima, es decir información que no guarda relación con una persona física identificada o identificable, ni a los datos convertidos en anónimos de forma que el interesado no sea identificable, o deje de serlo. En consecuencia, el presente Reglamento no afecta al tratamiento de dicha información anónima, inclusive con fines estadísticos o de investigación» (considerando 26, RGPD).

La disociación de datos es una herramienta útil para abordar los procesos de *analytics* sin las limitaciones que incorpora la normativa en protección de datos personales, siempre y cuando se controlen razonablemente los potenciales riesgos de reidentificación.

Deberemos considerar que, en el caso de que dicho riesgo se materialice, volveremos a estar sujetos a todas las obligaciones derivadas de la normativa.

La justificación de esto nos la da propia normativa en la definición de su objeto de aplicación, es decir, si no tratamos datos personales.

El término ***disociación*** es un concepto definido en la Ley Orgánica de Protección de Datos 15/1999 como:

«Todo tratamiento de datos personales de modo que la información que se obtenga no pueda asociarse a persona identificada o identificable» (art. 3.f, LOPD, de 13 de diciembre)».

Esta definición se reitera en el Real Decreto 1720/2007: «todo tratamiento de datos personales que permita la obtención de datos disociados» (art. 5.p, RD 1720/2007, de 21 de diciembre), siendo un

dato disociado «aquél que no permite la identificación de un afectado o interesado» (art. 5.e, RD 1720/2007, de 21 de diciembre).

Por tanto, si atendemos a la aplicación de la LOPD, en su artículo 2: «la presente Ley Orgánica será de aplicación a los datos de carácter personal registrados en soporte físico, que los haga susceptibles de tratamiento y a toda modalidad de uso posterior de estos datos por los sectores público y privado» (art. 2.1, LOPD, de 13 de diciembre).

Tendremos como resultado que, **aplicando procesos de disociación, al no contar con datos que identifiquen o faciliten la identificación de los individuos, no estarían sometidos al amparo de las normas de protección de datos** y, por consiguiente, ese conjunto de datos podría tratarse sin las limitaciones que de la normativa en tratamientos de datos personales se deriva: estaríamos tratando datos disociados.

En opinión del WP29, en su «Opinion 05/2014 on Anonymisation Techniques», opinión adoptada el 10 de abril de 2014, resalta el hecho de que diferentes estudios versan sobre la dificultad de la creación de verdaderos conjuntos de datos anonimizados a partir de un conjunto de datos personales extensos, mientras se intenta mantener un conjunto de datos suficientes para el objeto de la tarea o procesamiento.

Los procesos de anonimización pueden restar información que sea necesaria para el procesamiento. Además, un conjunto de datos considerado anónimo, combinado con otros, puede originar que uno o varios individuos sean identificados.

Además, existe un factor de riesgo inherente a la anonimización: es cada vez más difícil de lograr auténticos conjuntos disociados a partir de conjuntos de datos personales por el avance de la tecnología informática moderna y la disponibilidad ubicua de la información. Hay autores que cuestionan esta opinión, considerando que los riesgos son realmente bajos.

Este factor de riesgo ha de considerarse en la evaluación de la conveniencia de cualquier técnica de anonimización (incluyendo los posibles usos de los datos que son anonimizados a través de estas técnicas de anonimización) y el potencial impacto sobre los afectados y la probabilidad de ocurrencia debe ser evaluada.

Un proceso de disociación para alcanzar una anonimización completa requerirá que se evite cualquier posibilidad razonable de establecer un vínculo con los datos de otras fuentes con el fin de reidentificar a los individuos, pero en la práctica hay unas zonas grises muy significativas, donde un responsable de tratamiento puede creer que un conjunto de datos está disociado (anonimizado), pero un tercero motivado será todavía capaz de identificar al menos algunos de los individuos de la información dada a conocer.

Gestionar y revisar periódicamente el riesgo de reidentificación, incluyendo la identificación de los riesgos residuales, es un elemento importante de cualquier sólido enfoque en esta área.

Es de reseñar que, cuando un responsable del tratamiento no elimina los datos originales (identificables) a nivel de evento y el responsable del tratamiento maneja parte de este conjunto de datos (por ejemplo, después de la eliminación o enmascaramiento de los datos de identificación), el conjunto de datos resultante todavía será de datos personales.

Para el responsable de tratamiento existe la posibilidad de proceder a la reidentificación mediante medios razonables.

Mientras que, si un tercero procesa un conjunto de datos tratados con un proceso de disociación adecuado (contará con datos anónimos y facilitados por el encargado de tratamiento original), pueden hacerlo legalmente sin necesidad de tomar en cuenta las necesidades de protección de datos, ya que no podrá (directa o indirectamente) identificar los titulares de los datos en el conjunto de datos original, que no está a su disposición.

Anonimización parcial

No siempre será posible la disociación completa o anonimización debido a la naturaleza del procesamiento (por ejemplo, donde puede haber una necesidad de volver a identificar a los sujetos de datos o la necesidad de utilizar los datos más granulares que, como efecto secundario, puede permitir la identificación indirecta).

En estos casos, una ayuda será la de un proceso de anonimización parcial, en el que **el conjunto resultante no puede tener la consideración de anonimizado,** pero sí se habrá dificultado la identificación de los individuos.

Para ello nos asisten diversas técnicas que pueden ser aplicadas (incluyendo seudoanonimización, *keycoding*, *keyed hashing*, la eliminación de identificadores y valores atípicos, sustitución de identificadores únicos, introducción de ruido y otros). La selección de la técnica más adecuada deberá poner el foco en que el resultado sea lo más difícil posible de proceder a una identificación.

Además, al no contar con información totalmente anonimizada deberemos aplicar medidas adicionales de seguridad entre las que, en el caso de España, estarán las correspondientes al nivel de seguridad aplicable a los datos personales del fichero origen.

En este sentido, el WP29 recomienda la aplicación de medidas adicionales como, por ejemplo:

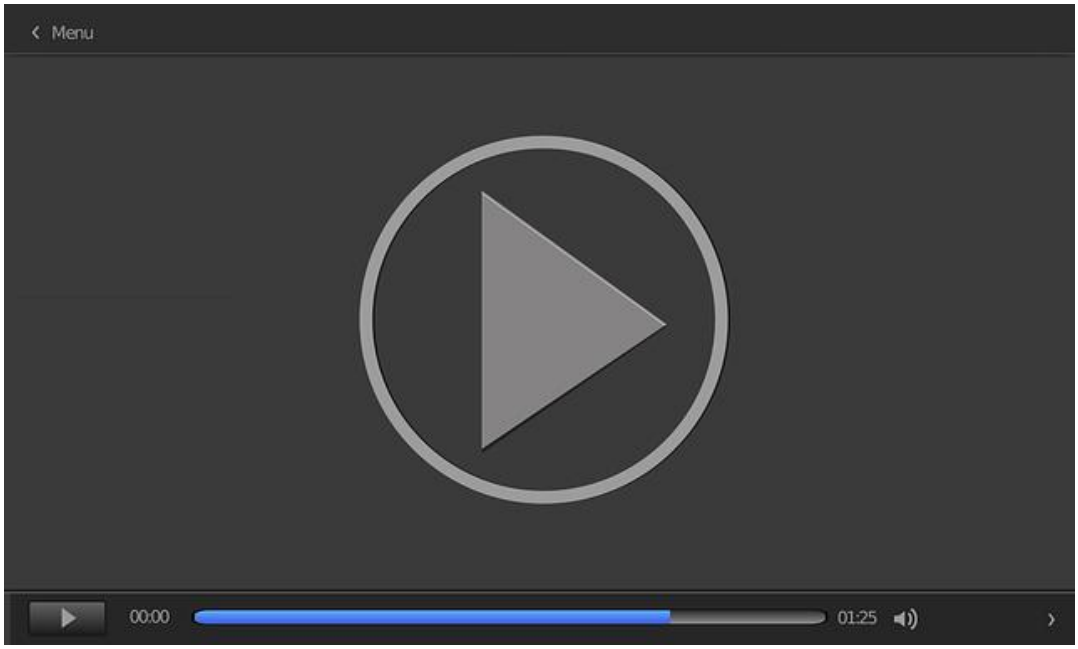
- La adopción de medidas adicionales de seguridad específicas (como el cifrado).
- En caso de seudoanonimización, asegurarse de que los datos que permite la vinculación de la información a un objeto de datos (las claves) que los identifica han sido codificados o encriptados y almacenados por separado.
- La incorporación de un tercero de confianza en situaciones en las que una serie de organizaciones quieran anonimizar los datos personales que tienen y se quieran utilizar en un proyecto en colaboración.
- Restringir el acceso a los datos personales aplicando el principio de la **necesidad de saber,** equilibrando cuidadosamente los beneficios de una mayor difusión contra los riesgos de la divulgación inadvertida de datos personales a personas no autorizadas. Esto puede incluir, por ejemplo, permitir acceso de solo lectura en locales controlados o la aplicación de medidas para permitir solo su acceso en entornos seguros y comunidades cerradas.

También es importante establecer obligaciones legalmente exigibles de confidencialidad a los beneficiarios de los datos, incluyendo la prohibición de la publicación de información. Siempre hay que considerar que la publicación de información tiene un alto riesgo por diferentes razones:

- Errores de divulgación involuntaria.
- Procesos de disociación erróneos que facilitan la identificación de los individuos.
- La anonimización considerada como absolutamente irreversible que potencie la publicación de datos ante la falsa sensación de seguridad.

Ideas clave

Para acabar este apartado, accede al vídeo *Tratamiento de datos de menores.*



Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=f5ff9972-9396-4d49-95a0-abd800c24758

10.4. Técnicas de anonimización

Existen diferentes técnicas de anonimización y en los últimos años, desde la comunidad científica, espoleados por los riesgos asociados de reidentificación vinculados a las técnicas existentes, se han ido mejorando o introduciendo nuevas técnicas.

Por lo general, el proceso de anonimización de un conjunto de datos va a requerir la aplicación conjunta de estas técnicas con objeto de reducir las debilidades que cada una de ellas presenta frente al riesgo de reidentificación.

Elegir la técnica o técnicas más adecuadas para aplicar a cada caso va a depender de aspectos relacionados con la sensibilidad de la información, la necesidad de mantener una alta correspondencia con los datos originales para el fin que se les vaya a dar y el riesgo de reidentificación comprometido.

Clasificación de las técnicas de anonimización

Las técnicas de anonimización se pueden clasificar en función de su naturaleza, el efecto que tienen sobre los datos o en función de su aplicación.

- **Clasificación en función de su naturaleza:** existen dos familias de técnicas principales:
    - Aleatorización: modifican la veracidad que tienen los datos reduciendo el vínculo de los datos y las personas reduciendo probabilidad de inferencia entre los datos.
    - Generalización: su aplicación generaliza un valor en un atributo o columna o diluye los atributos (por ejemplo, reduce el código postal a los dos primeros dígitos). Reducen la significación de un individuo a partir de un valor de atributo singular.

Ideas clave

- Clasificación en función del efecto que tienen sobre los datos:
  - Técnicas de **reducción de atributos:** se reducen valores en las tablas de datos eliminando aquellos que facilitan la reidentificación:
    - Supresión de identificadores directos.
    - Agregación: reducción del nivel de detalle de la información.
    - Muestreo: selección de datos de entre una amplia muestra.
  - Técnicas de **modificación de los datos**: los datos son alterados para reducir la posibilidad de reidentificación:
    - Generalización.
    - Adición de ruido.
    - Asignación al azar (aleatorización de los valores de los datos).
    - Permutación o intercambio de datos.
    - Privacidad diferencial.
    - Supresión de datos.
    - Seudoanonimización.
  - Métodos **basados en restricción:** se introducen restricciones en el conjunto de datos para eliminar atributos que facilitan la reidentificación:
    - Supresión de celdas.
    - Cambio del esquema de clasificación.

- **Clasificación en función de su aplicación** a microdatos o macrodatos, así tenemos:
  - Técnicas para reducir los riesgos de identificación en microdatos.
  - Técnicas para reducir los riesgos de identificación en macrodatos.

Técnicas para reducir los riesgos de identificación en microdatos

Las aproximaciones habituales para reducir los riesgos de reidentificación en el ámbito de la estadística y en el tratamiento de microdatos son:

- **Técnicas que implican la reducción de atributos:** la reducción de datos contenidos en la muestra es más frecuentemente utilizada que la modificación de los datos y generalmente esto consiste en:
  - **Supresión de identificadores directos:** eliminar del conjunto de datos los identificadores que de manera directa identifican a los individuos. Por ejemplo, eliminación de nombres y apellidos, DNI y número de seguridad social.
  - **Agregación y reducción del nivel de detalle de la información:** una aproximación muy común es una forma de redondeo de datos o agrupación de los mismos en torno a grandes categorías, llevando el conjunto a un riesgo aceptable de reidentificación. Para ello se reduce el número de registros que presentan atributos con una única combinación y que hace que el riesgo de reidentificación sea elevado. Por ejemplo:

- La fecha de nacimiento se reduce al año.
- El código postal se deja en los dos primeros dígitos.Entre estas técnicas destacan: las técnicas de agregación o k-anonimato, diversidad *l* y proximidad *t*.

- **Muestreo:** si el conjunto de datos tiene más registros de los necesarios, una selección muestral de registros nos puede ayudar a limitar los riesgos. Hace años era una práctica muy habitual por las limitaciones en procesamiento. La amplia superación de estas limitaciones en la era *big data* ha llevado a que esta técnica sea menos utilizada.

- **Técnicas que implican la modificación de los datos:** la modificación de datos es una aproximación más radical, pero útil a la hora de reducir la capacidad informativa de los datos. Algunas de estas técnicas son:

  - **Generalización:** consiste en repetir el mismo valor en todas las filas del conjunto de datos. Para ello es frecuente escoger el atributo menos singular o específico y con sentido semántico. En la práctica, la generalización de un valor en un atributo se aproxima a la eliminación del atributo de la tabla. Ejemplo:

Edad	Sexo	Puntuación
12	M	5
12	M	7
12	H	4
12	H	8
13	M	5

Tabla 1. Ejemplo de generalización I. Fuente: elaboración propia.

Generalizamos el dato del sexo:

Edad	Sexo	Puntuación
12	Z	5
12	Z	7
12	Z	4
12	Z	8
13	Z	5

Tabla 2. Ejemplo de generalización II. Fuente: elaboración propia.

Entre las técnicas de generalización destacan las técnicas de agregación o k-anonimato, diversidad *l* y proximidad *t*, que veremos más detenidamente en el siguiente capítulo.

- **Adición de ruido a los datos:** es una técnica de aleatorización que suele ser útil cuando los atributos son altamente sensibles y pueden causar un impacto significativo sobre los individuos Es un método que se aplica a la información numérica, consistente en la adición de ruido aleatorio a todos los valores de la columna (variable) que debe ser protegida.Y se hace de manera que perturbe lo menos posible la propia variable, de modo que la suma de los valores aplicados sea cero o tienda a cero. Es decir, en una fila suma y en otra resta manteniendo el valor de la distribución sin alteración. Para ello la varianza del valor adicionada debe tender a cero. Frente a una alta varianza que implicaría que la perturbación incluida es significativa.Supongamos que un estudio de prevalencia del cáncer en zonas geográficas requiere de alta fidelidad del dato de **localidad** de residencia del individuo o del **código postal,** por lo que es preciso mantener las cuatro cifras del código postal. Una solución sería cambiar de forma aleatoria las cinco cifras de estos datos dentro de un aceptable rango. Normalmente, los paquetes estadísticos facilitan este tipo de transformaciones.Otro ejemplo, en un estudio de ingresos y gastos personales se dispone de los siguientes datos:

Edad	Género	Código postal	Ingresos	Gastos/mes
24	F	28001	20 000	1100
34	M	28001	32 000	1800
45	F	28004	68 000	2800
32	F	28056	31 500	2000
28	F	28044	22 000	1300
73	M	28024	28 000	1200

Tabla 3. Ejemplo de adición de ruido I. Fuente: elaboración propia.

Si seleccionamos añadir **ruido** de modo que la suma de los valores introducidos sea cero y la variancia sea solo de 1, la tabla quedará como sigue:

Edad	Género	Código postal	Ingresos	Gastos/mes
24	F	28001	19 828	1100
34	M	28001	32 960	1800
45	F	28004	66 951	2800
32	F	28056	28 957	2000
28	F	28044	23 862	1300
73	M	28024	28 942	1200

Tabla 4. Ejemplo de adición de ruido II. Fuente: elaboración propia.

Hemos introducido ruido sobre la variable «ingresos», de manera que hemos restado información que facilite la reidentificación de los individuos.

Ideas clave

La adición de ruido también es una técnica muy utilizada en las comunicaciones, de manera que estas se distorsionan y el observador accede a la comunicación con ruido añadido, lo que dificulta su análisis. La parte receptora reinvierte el algoritmo generador de ruido accediendo a la información.

Es una técnica que ofrece buenos resultados, pero normalmente debe aplicarse junto a otras como la eliminación de cuasiidentificadores y de atributos obvios; de otro modo nos vemos obligados a incorporar una distorsión importante para alcanzar los objetivos de anonimización deseados.

Muchos de los riesgos en la utilización de la técnica sobre la reidentificación vienen por la no adición de ruido suficiente o porque este no es consistente, orientando al adversario hacia el algoritmo de adición de ruido. La clave es que el adversario piense que se trata de valores ciertos sin serlo.

- **Asignación al azar: aleatorización de los valores de los datos :** esta técnica se utiliza con profusión en los entornos de pruebas de desarrollo de *software*. Por ejemplo, identidades de personas (de prueba) para los que el nombre y los apellidos se obtienen aleatoriamente a partir de un conjunto de nombre y apellidos reales.

- Los nombres se asignan con la misma distribución de la realidad, sin incluir aquellos extremos de frecuencia menor (los menos comunes). También se puede aplicar a números de identificación: DNI, número de la seguridad social. Existen programas que ayudan a realizar estas funciones generando también los códigos de control habituales de estos identificadores.

- **Permutación intercambio de datos:** es una técnica que puede ser considerada como de adición de ruido. Para su aplicación se identifican parejas de registros correspondientes a individuos con ciertas similitudes y se intercambian los identificadores quedando vinculados a distintos interesados. El conjunto final de datos está alterado frente al conjunto inicial, pero para la realización de análisis de datos ofrecen los mismos resultados. Tiene un uso extendido para la elaboración de estadísticas de datos agregados, pero es poco útil en análisis de regresiones multivariables.

    - El objetivo, en definitiva, es el de introducir cierta confusión ante un posible ataque y habitualmente los registros afectados dentro de los conjuntos de datos son del orden del 1 al 10 %.

Edad	Género	Código postal	Ingresos	Gastos/mes
24	F	28001	20 000	1100
34	M	28001	32 000	1800
45	F	28004	68 000	2800
32	F	28056	31 500	2000
28	F	28044	22 000	1300
73	M	28024	28 000	1200

Tabla 5. Ejemplo de permutación I. Fuente: elaboración propia.

La aplicación del método en este ejemplo nos permitiría intercambiar los valores de los datos de edad 28 y 24.

Edad	Género	Código postal	Ingresos	Gastos/mes
24	F	28044	19 828	1100
34	M	28001	32 960	1800
45	F	28004	66 951	2800
32	F	28056	28 957	2000
28	F	28001	23 862	1300
73	M	28024	28 942	1200

Tabla 6. Ejemplo de permutación II. Fuente: elaboración propia.

En este ejemplo intercambiamos las variables de código postal, ingresos y gastos, de manera que, en el caso de realizar un análisis por grupos de edad, el resultado no se ve alterado, pero sí hemos generado una limitación para la identificación del individuo, pues la edad no se correspondería ni el código postal.

Vemos que la permutación no altera los datos estadísticos de la columna sobre la que se realiza. En este caso mantiene la media aritmética pero también mantiene la moda, la mediana y la desviación típica. Por eso es una técnica fácil de aplicar que no distorsiona los análisis y sí introduce una dificultad de reidentificar al individuo al que corresponden los datos. Ahora bien, se debe analizar su aplicación para que sea más efectiva. Veamos otro ejemplo.

En una empresa se está realizando un estudio de ingresos por categoría profesional, sexo, edad y estudios.

Año	Sexo	Cargo	Ingresos
1998	M	Ingeniero	40 000
1997	M	Médico	45 000
2005	M	Operador	15 000
1965	M	Gerente	70 000
1957	M	Director ejecutivo	200 000
		Media	72 000

Tabla 7. Ejemplo de permutación III. Fuente: elaboración propia.

Con el objeto de anonimizar los datos se aplica una permutación a los datos salariales con la pretensión de que no se pueda asociar salarios a las personas.

Año	Sexo	Cargo	Ingresos
1998	M	Ingeniero	40 000
1997	M	Médico	45 000
2005	M	Operador	200 000
1965	M	Gerente	70 000
1957	M	Director ejecutivo	15 000
		Media	72 000

Tabla 8. Ejemplo de permutación IV. Fuente: elaboración propia.

Ahora, en el ejemplo propuesto vemos que la permutación al mantener el cargo tiene un efecto limitado, pues no es difícil inferir que el director ejecutivo y operador de las categorías existentes serán los que ocupen los extremos salariales de la muestra y, por consiguiente, es fácil determinar cuánto gana cada uno. En este caso, la permutación no es válida. En general, nos ofrece malos resultados cuando existe una fuerte correlación entre dos atributos (categoría laboral y salario).

- **Privacidad diferencial:** es una técnica que genera la alteración de los datos y pertenece a la familia de técnicas de aleatorización. Consiste en la adición de ruido selectivo a la salida de los datos, generando vistas anonimizadas del conjunto de los datos.Diferente en este sentido a las técnicas anteriores que se aplican directamente sobre el conjunto de datos base para su anonimización. En este caso, no se produce alteración de los datos, solo se muestran alterados.Tienen amplio predicamento en bancos de datos que facilitan acceso a terceros. La técnica facilita que, mediante la asignación de ruido, la misma consulta presente diferentes datos para distintos operadores.La debilidad que presenta es la potencial revelación del algoritmo de adición de ruido mediante el análisis de un volumen significativo de búsquedas. Esto se puede mitigar limitando las consultas posibles. El resto de las debilidades que presenta viene más de la deficiente aplicación de la técnica que por la naturaleza de la misma. Por ejemplo: poco ruido.

- **Técnicas que implican la supresión de datos:** cuando las técnicas de reducción o modificación no ofrecen un resultado óptimo sobre el riesgo residual de reidentificación, se pueden aplicar técnicas de supresión: eliminar los registros que ofrecen un mayor riesgo en la reidentificación de los individuos. Obviamente, esto puede introducir distorsión en los estudios, pues la supresión no es aleatoria al eliminar la fila completa del registro problemático.Calificaciones de matemáticas de los alumnos de la Comunidad de Madrid. Rendimiento académico de los niños:

Edad	Sexo	Puntuación
12	M	5
12	M	7
12	H	4
12	H	8
13	M	5

Tabla 9. Ejemplo de supresión I. Fuente: elaboración propia.

Se suprime el registro que frente al resto tiene mayor riesgo de reidentificación:

Edad	Sexo	Puntuación
12	M	5
12	M	7
12	H	4
12	H	8

Tabla 10. Ejemplo de supresión II. Fuente: elaboración propia.

- **Técnicas de pseudoanonimización:** es una técnica de desidentificación de un conjunto de datos, pero facilita la reidentificación de los mismos. El seudónimo permite vincular todos los datos relacionados, sin facilitar la identificación del individuo de una forma directa. Es una técnica muy extendida, existiendo aplicaciones comerciales que facilitan su aplicación.

Esta técnica reduce la vinculabilidad que podemos tener entre los datos y la identidad del individuo, pero no la elimina e incluso en algunos casos se mantiene vía un proceso reversible.

Ideas clave

Los seudónimos pueden ser de dos tipos, reversibles o irreversibles, y se suelen obtener de diferentes maneras, por ejemplo:

- La pseudoanonimización reversible, que se puede obtener por:
  - Vía codificación del identificador: con una técnica de codificación/encriptación reversible.
  - Estableciendo tablas de correspondencia entre identificador seudónimo: en este caso, por ejemplo, se asigna a cada identificador un seudónimo (números, caracteres), manteniendo en un conjunto separado de datos la correspondencia entre el seudónimo y los identificadores reemplazados.
- La pseudoanonimización irreversible, que se puede obtener mediante función *hash:* es una función que siempre devuelve una cadena de igual tamaño, independientemente del valor que se pase a la función. Es irreversible y vulnerable a la hora de poder obtener el identificador que seudoanimiza. Por ejemplo, si en un sistema hemos aplicado la función *hash* sobre el DNI.

La seudonimización en el ámbito científico o empresarial habitualmente es abordada con diferentes aproximaciones que limitan el riesgo de reidentificación:

- Los seudónimos y los identificadores son mantenidos dentro de la organización, limitándose el acceso a las claves criptográficas o las tablas de correspondencia para facilitar el anonimato de los datos.
- La correspondencia de seudónimos e identificadores son facilitados para su custodia a un tercero, ofreciendo en caso de necesidad la posibilidad de reidentificar el seudónimo. Esto es habitual en la investigación sanitaria.

Existe una especificación técnica para la seudonimización recogida en la ISO/TS 25237 (Health informatics. Pseudonynisation) y desarrollada para el ámbito sanitario, pero que es extensible a otros ámbitos de la estadística o investigación.

Técnicas para reducir los riesgos de identificación en macrodatos

Las técnicas más extendidas en este ámbito son:

- Métodos basados en **restricciones.** Los más comunes son los siguientes:

- **Supresión de celdas:** consiste directamente en la eliminación de las filas que contienen celdas que son problemáticas. En su estudio:

Año	Sexo	Cargo	Ingresos
1998	M	Ingeniero	40 000
1997	M	Médico	45 000
2005	M	Desempleado	5 000
1965	M	Gerente	70 000
1957	M	Director ejecutivo	200 000

Tabla 11. Supresión de celdas. Fuente: elaboración propia.

- **Cambio del esquema de clasificación:** se puede realizar modificando para cada categoría los puntos de corte (es decir, los extremos de la categoría) o mediante la agregación de celdas. Una aproximación frecuente es la denominada codificación de arriba a abajo. Por ejemplo, en un conjunto de datos podemos agrupar los extremos de las edades, agrupando todas las edades en torno a un valor o bajo un determinado valor. El objetivo es eliminar los grupos menores.

Veamos un ejemplo: se está procediendo a realizar un estudio en el ámbito sanitario; junto a la información requerida del estudio se incluye el género, la edad y la patología, los datos que se tienen.

sexo	edad	Enf Tiroides	sexo	edad	Enf Tiroides
M	11	Fibromialgia	M	44	Cáncer de colón
M	12	Lupus	M	44	Lupus
M	12	Neumonia	M	44	Lupus
H	12	Lupus	H	44	Cáncer de colón
H	12	Enf Tiroides	H	44	Lupus
M	13	Cáncer de colón	M	45	Enf Tiroides
M	13	Fibromialgia	M	45	Cáncer de colón
H	14	Cáncer de colón	H	45	Enf Tiroides
H	14	Enf Tiroides	M	46	Lupus
H	14	Enf Tiroides	H	46	Cáncer de colón
M	15	Hipertensión	M	47	Enf Tiroides
M	15	Lupus	H	47	Lupus
M	15	Neumonia	H	52	Cáncer de colón
M	17	Enf Tiroides	H	52	Hipertensión
H	17	Cáncer de colón	M	53	Enf Tiroides
M	18	Hipertensión	H	53	Enf Tiroides
H	18	Enf Tiroides	H	53	Hipertensión
M	19	Cáncer de colón	H	53	Hipertensión
M	19	Enf Tiroides	M	54	Cáncer de colón
H	19	Lupus	M	54	Enf Tiroides
H	19	Enf Tiroides	M	54	Enf Tiroides
H	23	Enf Tiroides	M	54	Enf Tiroides
H	23	Enf Tiroides	M	55	Enf Tiroides
H	24	Enf Tiroides	M	55	Enf Tiroides
H	24	Cáncer de colón	H	55	Lupus
H	24	Enf Tiroides	H	55	Hipertensión
H	25	Enf Tiroides	M	56	Cáncer de colón
M	26	Cáncer de colón	M	57	Enf Tiroides
M	27	Fibromialgia	H	59	Lupus
M	27	Lupus	H	91	Hipertensión
M	27	Cáncer de colón	H	99	Hipertensión
M	27	Lupus	M	39	Cáncer de colón
M	28	Lupus	M	39	Enf Tiroides
M	28	Enf Tiroides	M	39	Cáncer de colón
M	28	Enf Tiroides	M	39	Lupus
H	30	Enf Tiroides	M	39	Enf Tiroides
M	31	Fibromialgia	M	39	Cáncer de colón
M	32	Fibromialgia	H	39	Enf Tiroides
M	32	Fibromialgia	H	39	Cáncer de colón
M	32	Fibromialgia	H	39	Lupus
M	32	Fibromialgia	H	39	Enf Tiroides
M	32	Lupus	H	39	Cáncer de colón
M	32	Lupus	H	39	Lupus
M	33	Fibromialgia	M	40	Lupus
M	33	Fibromialgia	M	40	Enf Tiroides
M	33	Enf Tiroides	M	40	Cáncer de colón
M	33	Cáncer de colón	M	42	Enf Tiroides
M	33	Lupus	H	42	Lupus
M	33	Enf Tiroides	M	43	Cáncer de colón
M	33	Enf Tiroides	M	43	Lupus
M	33	Enf Tiroides	M	43	Enf Tiroides
M	33	Enf Tiroides	H	43	Cáncer de colón
M	33	Enf Tiroides	H	43	Enf Tiroides
M	33	Cáncer de colón	H	37	Lupus
M	33	Enf Tiroides	H	37	Enf Tiroides
M	33	Cáncer de colón	H	37	Cáncer de colón
M	33	Lupus	H	37	Lupus
M	33	Enf Tiroides	H	37	Enf Tiroides
M	33	Enf Tiroides	M	38	Cáncer de colón
M	33	Cáncer de colón	H	38	Lupus
M	33	Cáncer de colón	M	39	Enf Tiroides
M	34	Neumonia	M	36	Cáncer de colón
M	34	Hipertensión	H	36	Enf Tiroides
M	34	Lupus	M	37	Lupus
M	35	Enf Tiroides	M	37	Enf Tiroides
M	35	Cáncer de colón	H	37	Enf Tiroides
M	35	Fibromialgia	H	37	Cáncer de colón
M	35	Enf Tiroides			

Tabla 12. Ejemplo conjunto de datos I.

Con el objeto de identificar el riesgo de reidentificación, se realiza una clasificación de los registros, para lo cual se establece unas categorías por intervalos de fechas. El resultado referente al número de registros que obtenemos para cada categoría de

edad para hombres y mujeres nos muestra que hay una categoría para la que solamente tendríamos un registro. Lo que generaría que para este registro único tengamos un riesgo de reidentificación más elevado que si en la misma categoría coincidieran números registros.

	Menos de 25 años	De 26 a 35 años	De 36 a 45 años	Más de 45 años	Total
Mujeres	12	40	23	11	86
Hombres	15	1	21	12	49
Total	27	41	44	21	135

Tabla 13. Ejemplo cambio en la clasificación I. Fuente: elaboración propia.

Para reducir esto podemos modificar el intervalo de las categorías, por ejemplo:

	Menos de 20 años	De 21 a 35 años	De 36 a 45 años	Más de 45 años	Total
Mujeres	12	40	23	11	86
Hombres	9	7	21	12	49
Total	21	47	44	21	135

Tabla 14. Ejemplo cambio en la clasificación II. Fuente: elaboración propia.

Esta nueva categorización lleva a que la menor categoría de datos sea de 7 (7 registros). Al considerar que es un número insuficiente podemos eliminar una categoría uniéndola a otra:

	Hasta 35 años	De 36 a 45 años	Más de 45 años	Total
Mujeres	52	23	11	86
Hombres	16	21	12	49
Total	68	44	21	135

Tabla 15. Ejemplo cambio en la clasificación III. Fuente: elaboración propia.

Es evidente que la elección de los intervalos más adecuados o incluso la agrupación de los mismos es un proceso laborioso que siempre tendrá que tener en consideración la finalidad del estudio que se realiza. Existen herramientas que ayudan a realizar este proceso. Pero en cualquier caso tendrá un componente de análisis previo muy importante.

En nuestro ejemplo y tras aplicar sobre los datos el análisis realizado obtenemos:

sexo	edad	Enf Tiroides	sexo	edad	Enf Tiroides
M	hasta 35 años	Fibromialgia	M	de 36 a 45	Cáncer de colón
M	hasta 35 años	Lupus	M	de 36 a 45	Lupus
M	hasta 35 años	Neumonia	M	de 36 a 45	Lupus
H	hasta 35 años	Lupus	H	de 36 a 45	Cáncer de colón
H	hasta 35 años	Enf Tiroides	H	de 36 a 45	Lupus
M	hasta 35 años	Cáncer de colón	M	de 36 a 45	Enf Tiroides
M	hasta 35 años	Fibromialgia	M	de 36 a 45	Cáncer de colón
H	hasta 35 años	Cáncer de colón	H	de 36 a 45	Enf Tiroides
H	hasta 35 años	Enf Tiroides	M	más de 45	Lupus
H	hasta 35 años	Enf Tiroides	H	más de 45	Cáncer de colón
M	hasta 35 años	Hipertensión	M	más de 45	Enf Tiroides
M	hasta 35 años	Lupus	H	más de 45	Lupus
M	hasta 35 años	Neumonia	H	más de 45	Cáncer de colón
M	hasta 35 años	Enf Tiroides	H	más de 45	Hipertensión
H	hasta 35 años	Cáncer de colón	M	más de 45	Enf Tiroides
M	hasta 35 años	Hipertensión	H	más de 45	Enf Tiroides
H	hasta 35 años	Enf Tiroides	H	más de 45	Hipertensión
M	hasta 35 años	Cáncer de colón	H	más de 45	Hipertensión
M	hasta 35 años	Enf Tiroides	M	más de 45	Cáncer de colón
H	hasta 35 años	Lupus	M	más de 45	Enf Tiroides
H	hasta 35 años	Enf Tiroides	M	más de 45	Enf Tiroides
H	hasta 35 años	Enf Tiroides	M	más de 45	Enf Tiroides
H	hasta 35 años	Enf Tiroides	M	más de 45	Enf Tiroides
H	hasta 35 años	Enf Tiroides	M	más de 45	Enf Tiroides
H	hasta 35 años	Cáncer de colón	H	más de 45	Lupus
H	hasta 35 años	Enf Tiroides	H	más de 45	Hipertensión
H	hasta 35 años	Enf Tiroides	M	más de 45	Cáncer de colón
M	hasta 35 años	Cáncer de colón	M	más de 45	Enf Tiroides
M	hasta 35 años	Fibromialgia	H	más de 45	Lupus
M	hasta 35 años	Lupus	H	más de 45	Hipertensión
M	hasta 35 años	Cáncer de colón	H	más de 45	Hipertensión
M	hasta 35 años	Lupus	M	de 36 a 45	Cáncer de colón
M	hasta 35 años	Lupus	M	de 36 a 45	Enf Tiroides
M	hasta 35 años	Enf Tiroides	M	de 36 a 45	Cáncer de colón
M	hasta 35 años	Enf Tiroides	M	de 36 a 45	Lupus
H	hasta 35 años	Enf Tiroides	M	de 36 a 45	Enf Tiroides
M	hasta 35 años	Fibromialgia	M	de 36 a 45	Cáncer de colón
M	hasta 35 años	Fibromialgia	H	de 36 a 45	Enf Tiroides
M	hasta 35 años	Fibromialgia	H	de 36 a 45	Cáncer de colón
M	hasta 35 años	Fibromialgia	H	de 36 a 45	Lupus
M	hasta 35 años	Fibromialgia	H	de 36 a 45	Enf Tiroides
M	hasta 35 años	Lupus	H	de 36 a 45	Cáncer de colón
M	hasta 35 años	Lupus	H	de 36 a 45	Lupus
M	hasta 35 años	Fibromialgia	M	de 36 a 45	Lupus
M	hasta 35 años	Fibromialgia	M	de 36 a 45	Enf Tiroides
M	hasta 35 años	Enf Tiroides	M	de 36 a 45	Cáncer de colón
M	hasta 35 años	Cáncer de colón	M	de 36 a 45	Enf Tiroides
M	hasta 35 años	Lupus	H	de 36 a 45	Lupus
M	hasta 35 años	Enf Tiroides	M	de 36 a 45	Cáncer de colón
M	hasta 35 años	Enf Tiroides	M	de 36 a 45	Lupus
M	hasta 35 años	Enf Tiroides	M	de 36 a 45	Enf Tiroides
M	hasta 35 años	Enf Tiroides	H	de 36 a 45	Cáncer de colón
M	hasta 35 años	Enf Tiroides	H	de 36 a 45	Enf Tiroides
M	hasta 35 años	Cáncer de colón	H	de 36 a 45	Lupus
M	hasta 35 años	Enf Tiroides	H	de 36 a 45	Enf Tiroides
M	hasta 35 años	Cáncer de colón	H	de 36 a 45	Cáncer de colón
M	hasta 35 años	Lupus	H	de 36 a 45	Lupus
M	hasta 35 años	Enf Tiroides	H	de 36 a 45	Enf Tiroides
M	hasta 35 años	Enf Tiroides	M	de 36 a 45	Cáncer de colón
M	hasta 35 años	Cáncer de colón	H	de 36 a 45	Lupus
M	hasta 35 años	Cáncer de colón	M	de 36 a 45	Enf Tiroides
M	hasta 35 años	Neumonia	M	de 36 a 45	Cáncer de colón
M	hasta 35 años	Hipertensión	H	de 36 a 45	Enf Tiroides
M	hasta 35 años	Lupus	M	de 36 a 45	Lupus
M	hasta 35 años	Enf Tiroides	M	de 36 a 45	Enf Tiroides
M	hasta 35 años	Cáncer de colón	H	de 36 a 45	Enf Tiroides
M	hasta 35 años	Fibromialgia	H	de 36 a 45	Cáncer de colón
M	hasta 35 años	Enf Tiroides			

Tabla 16. Ejemplo conjunto de datos II.

- **Heurística:** a veces se toman decisiones con el objeto de reducir el riesgo, a partir de la experiencia y cierto sentido lógico.

- **Métodos basados en permutaciones:** coincidente con la aproximación de intercambio de atributos entre registros similares, pero en este caso actuamos sobre

los macrodatos.

10.5. K-anonimato y sus variantes

Técnica de anonimización de datos desarrollada por Latanya Sweeney (profesora de gobierno y tecnología en la Universidad de Harvard y directora del Data Privacy Lab) en un artículo publicado en el 2002: *k-Anonymity: a model for protecting privacy.* Sweeney tiene patentado varios procedimientos y programas para generar conjuntos k-anónimos.

Un conjunto de datos es k-anónimo frente al conjunto de datos representado si la información contenida para cada individuo no puede distinguirse al menos de otros k-1 individuos cuya información también aparece en el conjunto de datos. Será 4-anónimo si para cada atributo del individuo que está en la muestra hay al menos otros 3 individuos que tienen el mismo atributo.

Pongamos un ejemplo: k = 3.

Seudoidentificador	Edad	Sexo	Diagnóstico
1	0 a 10	M	Lupus
2	20 a 35	F	Lupus
3	0 a 10	M	Lupus
4	51 a 65	F	Lupus
5	20 a 35	M	Lupus
6	51 a 65	F	Lupus
7	0 a 10	M	Lupus
8	20 a 35	F	Lupus
9	51 a 65	F	Lupus
10	0 a 10	F	Lupus
11	20 a 35	M	Lupus
12	51 a 65	M	Lupus
13	0 a 10	M	Lupus
14	0 a 10	F	Lupus
15	20 a 35	M	Lupus
16	51 a 65	M	Lupus
17	0 a 10	F	Lupus
18	51 a 65	M	Lupus
19	0 a 10	F	Lupus
20	20 a 35	F	Lupus
21	20 a 35	F	Lupus
22	0 a 10	M	Lupus

Tabla 17. Ejemplo k-anonimato I. Fuente: elaboración propia.

En el ejemplo hemos marcado la clase de coincidencia de edad, sexo y diagnóstico, y vemos que como mínimo cada clase contiene 3 individuos, por lo que la muestra podemos decir que es 3-anónimo.

Ordenemos el conjunto de datos para que se vea más fácilmente:

Seudoidentificador	Edad	Sexo	Diagnóstico
10	0 a 10	F	Lupus
14	0 a 10	F	Lupus
17	0 a 10	F	Lupus
19	0 a 10	F	Lupus
1	0 a 10	M	Lupus
3	0 a 10	M	Lupus
7	0 a 10	M	Lupus
13	0 a 10	M	Lupus
22	0 a 10	M	Lupus
2	20 a 35	F	Lupus
8	20 a 35	F	Lupus
20	20 a 35	F	Lupus
21	20 a 35	F	Lupus
5	20 a 35	M	Lupus
11	20 a 35	M	Lupus
15	20 a 35	M	Lupus
4	51 a 65	F	Lupus
6	51 a 65	F	Lupus
9	51 a 65	F	Lupus
12	51 a 65	M	Lupus
16	51 a 65	M	Lupus
18	51 a 65	M	Lupus

Tabla 18. Ejemplo k-anonimato II. Fuente: elaboración propia.

Los errores habituales a la hora de aplicar la técnica vienen dados fundamentalmente por un número *k* insuficiente. Cuanto más alto es el valor *k,* menos riesgos de inferencia existen.

k-anonimato no impide los ataques de inferencia, pero el modo en que se aplique

puede reducir esta debilidad. Por el contrario, tienen un comportamiento muy satisfactorio frente a los ataques de singularización, dado que al menos *k* elementos compartirán los mismos atributos.

El riesgo de vinculabilidad es escaso y, en el caso de que esta tenga lugar, lo hará con *k* registros, por lo que la probabilidad es de *1/k*.

l-Diversity

Una mejora sobre k-anonimato es l-Diversity, que consiste en aplicar el método k-anonimato, pero garantizando que cada clase tiene como mínimo *l* valores, de manera que se evita que existan conjuntos de datos que tengan una variabilidad escasa de atributos.

En el ejemplo anterior, estaríamos ante una situación 1-Diversity, pues no existe variabilidad del atributo de enfermedad. Es decir, si se compromete la reidentificación, sería evidente concluir que el individuo singularizado padece lupus. Habrá que distribuir el conjunto de datos de manera que al menos en cada clase k-anonimato se disponga de valores diferentes en el atributo objetivo (en este caso, enfermedad). De este modo, mejora el comportamiento del conjunto de datos sobre potenciales ataques de inferencia.

Si tuviéramos una mayor diversidad en los diagnósticos en cada clase edad/sexo, sería más difícil concluir el diagnóstico. Este es un ejemplo de 4-anonimato y 2-diversidad (como mínimo cada clase tiene dos valores diferentes en diagnóstico).

Seudoidentificador	Edad	Sexo	Diagnóstico
10	0 a 10	F	Lupus
14	0 a 10	F	Lupus
17	0 a 10	F	ELA
19	0 a 10	F	ELA
1	0 a 10	M	Lupus
3	0 a 10	M	Lupus
7	0 a 10	M	Lupus
13	0 a 10	M	ELA
22	0 a 10	M	ELA
2	20 a 35	F	Lupus
8	20 a 35	F	Lupus
20	20 a 35	F	ELA
21	20 a 35	F	ELA
5	20 a 35	M	Lupus
11	20 a 35	M	Lupus
15	20 a 35	M	ELA
4	51 a 65	F	Lupus
6	51 a 65	F	Lupus
9	51 a 65	F	ELA
12	51 a 65	M	Lupus
16	51 a 65	M	Lupus
18	51 a 65	M	ELA

Tabla 19. Ejemplo l-Diversity. Fuente: elaboración propia.

t-Closeness

Es un perfeccionamiento de l-Diversity, consistente en la creación de clases de equivalencia de manera que la distribución de los atributos en toda la muestra no sea

alterada.

No exige que en cada clase existan *l* valores diferentes, sino que, además, la distribución de los valores en el atributo en el conjunto total de datos no debe verse alterada. Se aplica sobre todo cuando es muy importante para el estudio que se vaya a realizar con los datos mantener el atributo lo más próximo a la realidad, es decir, a los datos originales.

10.6. Herramientas de software

El proceso de reducción del riesgo es un tanto artesanal, pero existen herramientas que ayudan en este proceso. Podemos clasificar estas herramientas según su orientación:

- Manejo de identificadores directos a nivel de microdatos, herramientas que eliminan o enmascaran identificadores directos:
- Reducción del riesgo de reidentificación a partir de identificadores indirectos a nivel de microdatos:
- Manejo de datos agregados:

10.7. Riesgos asociados a las técnicas de anonimización

Antecedentes

Se ha demostrado que las técnicas de deidentificación para la anonimización de datos son insuficientes a la hora de garantizar la confidencialidad, pues la reidentificación de los datos es posible.

De hecho, Paul Ohm (profesor asociado de la University of Colorado Law School), entre otros científicos, defiende como axioma que es incompatible disponer de datos utilizables y al mismo tiempo que estos datos estén perfectamente disociados, teniendo en consideración los avances realizados en la investigación en el ámbito de la anonimización de datos.

De hecho, no hay dudas de que, a partir de datos que no necesariamente de forma directa identifican a los individuos, es posible identificar a los mismos.

Tras exponer algunos casos de la literatura científica, veremos con más claridad los riesgos que se ciernen sobre las iniciativas de datos abiertos y lo importante que es analizarla en cada caso antes de proceder a facilitar los datos al público o a los grupos de investigación.

Veamos algunos casos significativos que fortalecen esta convicción.

El caso AOL

En agosto de 2006 AOL anunció una nueva iniciativa llamada *AOL Research,* sumándose a las iniciativas de abrir los datos a la comunidad científica. Publicaron 20 millones de búsquedas realizadas por 650 000 usuarios en el buscador de AOL, correspondientes a tres meses de actividad. Esta fue una iniciativa ampliamente aplaudida por la comunidad de investigadores del comportamiento social en Internet.

Antes de hacer pública esta información, AOL trató de anonimizar los datos con objeto de proteger la privacidad de los usuarios y para ello había suprimido los datos más obvios de identificación como, por ejemplo: nombre de usuario y dirección IP. El fin era facilitar a los investigadores poder correlacionar búsqueda necesaria para el análisis de comportamiento y procedieron a sustituir estos datos por una cadena de números.

El análisis de la información permitió la identificación de individuos, lo cual supuso un gran escándalo en EE. UU. y la dimisión del CTO. Fue silenciada la iniciativa *AOL Research.*

El caso de código postal, fecha de nacimiento y sexo

En Massachusetts, una agencia gubernamental llamada Comisión Insurance Group (GIC), que años antes había comprado un seguro de salud para los empleados del estado, decidió (en la década de los 90) liberar registros de visitas al hospital para cualquier investigador que los solicitara sin coste alguno.

Ideas clave

Para anonimizar los datos se habían eliminado los campos que contienen el nombre, dirección, número de la seguridad social y otros identificadores. GIC asumía que con estos se garantizaba la privacidad de los pacientes, a pesar de que en los datos se incluían prácticamente un centenar de atributos por paciente. Junto a la visita al hospital se mantuvieron el código postal, la fecha de nacimiento y el sexo.

Cuando se liberaron los datos se garantizaba por las instituciones que estos estaban perfectamente anonimizados y la privacidad de los empleados públicos estaba garantizada. Una estudiante sabía que el gobernador residía en un municipio del estado denominado Cambridge. Cambridge entonces era una ciudad de más de 50 000 habitantes y con 7 códigos postales. Latanya Sweeney, que así se llamaba la estudiante, compro por solo 20 USD el censo electoral de la ciudad de Cambridge, que contenía, entre otros, la información del nombre, dirección, código postal, fecha de nacimiento y sexo de cada votante.

Cambiando los datos del censo electoral y los datos facilitados por el GIC no le resultó difícil acceder al conjunto de visitas hospitalarias y diagnósticos que había publicado GIC, referidos al gobernador del estado. En la ciudad solo seis personas compartían fecha de nacimiento, de ellos solo tres eran hombres y solo uno vivía en el código postal dónde residía el gobernador. Sweeney envió registros de salud del gobernador (incluyendo diagnósticos y prescripciones) a su despacho. Para ello solo necesitó cruzar una fuente pública con datos de identificación, con una fuente de datos teóricamente anonimizada.

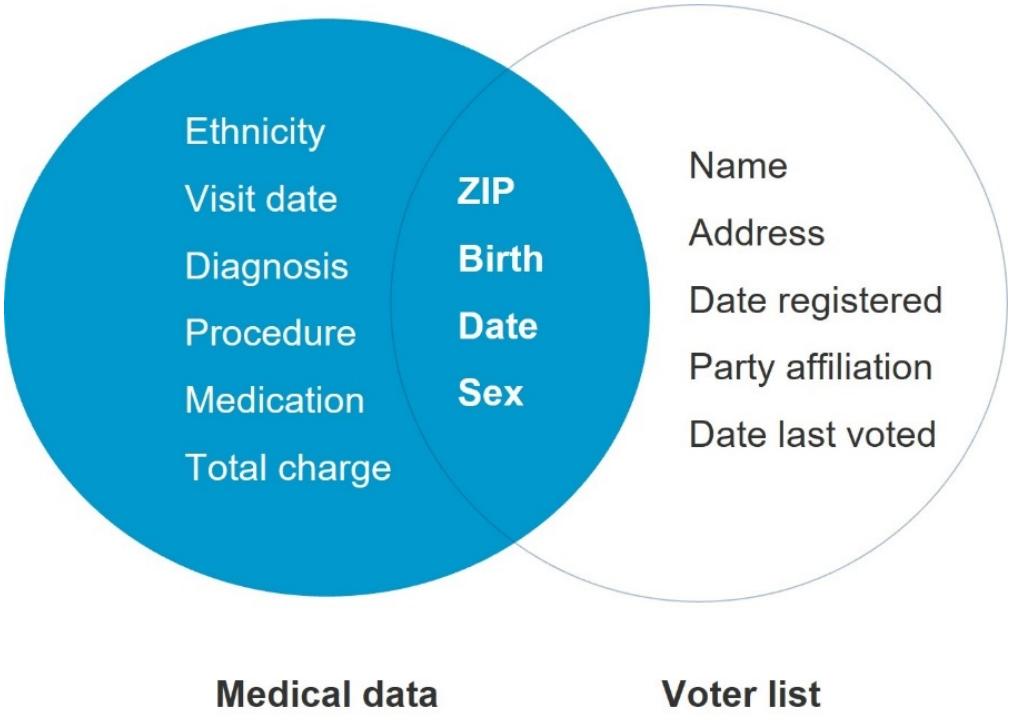


Figura 1. Cruce de datos. Fuente: elaboración propia.

Latanya Sweeney, profesora de Ciencias de la Computación, Tecnología y Política en la Universidad Carnegie Mellon, analizó el censo de los EE. UU. de 1996. En el año 2000 presentó un estudio para demostrar que en los EE. UU., simplemente con tres datos aparentemente inofensivos desde el punto de vista de su capacidad de identificar a las personas: código postal completo y la fecha nacimiento con el año incluido y el género o sexo, era posible identificar unívocamente al 87 % de la población de los EE. UU.

Para ello solo necesitó cruzar una fuente pública con datos de identificación con una fuente de datos teóricamente anonimizada.

Es decir, en un 87 % de los casos cualquier entidad u organización que tuviera estos datos podría identificar al individuo a partir de los datos del censo. Pero iba más lejos, el 57 % de la población era identificable con datos menos específicos como ciudad, fecha de nacimiento y sexo. Y hasta el 18 % con los datos de estado de residencia, fecha de nacimiento y sexo.

Los datos obtenidos de su estudio fueron:

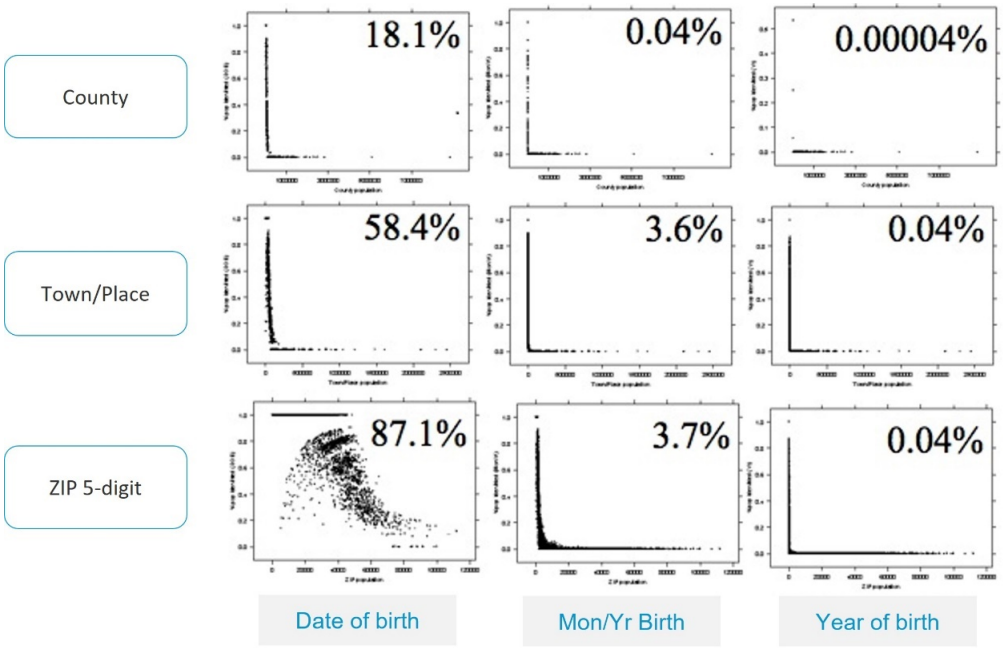


Figura 2. Resultados de estudios. Fuente: Sweeney, 2019.

Más tarde, Philippe Golle revisó el estudio de Latanya. Recalculó las estadísticas sobre el censo del año 2000 y contó solo con el código postal, la fecha de nacimiento y el sexo. La probabilidad de reidentificación era del 63 %; para el censo de 1996 (base del estudio de Latanya) era del 61 %.

	5 digit ZIP code	County
Year of birth	0,2 %	0,0 %
Year and month of birth	4,2 %	0,2 %
Year, month and day of birth	63,3 %	14,8 %

Tabla 20. Fraction of the US population uniquely identifiable by gender, location, date of birth.

Detengámonos un momento en estos datos perfectamente extrapolables a cualquier región que disponga de un censo que incorpore, junto a los datos de identificación de los individuos, los tres datos referidos (no en los porcentajes, pero sí en la posibilidad de reidentificación, pues un componente que tiene un impacto significativo es el grado de dispersión de la población).

CP	Fecha de nacimiento	Sexo	Enfermedad
28001	12/01/2005	M	Hepatitis C
28002	14/03/1987	H	Lupus
28005	25/05/1936	M	Fibromialgia
28004	22/01/1950	H	Hipertensión
28001	11/07/1947	M	Cáncer de colon

Tabla 21. Ejemplo conjunto datos (CP, nacimiento y sexo). Fuente: elaboración propia.

Si pensamos en el ámbito hospitalario y en un conjunto de datos deidentificado, tenemos que, lejos de considerar este conjunto de datos anonimizado, estaremos pensado en cuál es la probabilidad de identificar a un individuo, concluyendo:

- Que la desidentificación de los datos no origina un conjunto de datos anonimizado.
- Que es preciso aplicar técnicas adicionales para reducir o eliminar este riesgo.

El caso Netflix Prize

El 2 de octubre de 2006 Netflix, que en aquel momento era el más importante operador de alquiler de películas *online,* publicó un millón de documentos que revelaban la calificación de las películas realizadas por medio millón de clientes entre 1999 y 2005.

Cada registro contenía la siguiente información: la película, la puntuación asignada (de 0 a 5 estrellas) y la fecha de la valoración.

Antes de publicar los datos habían procedido a anonimizarlos, manteniendo exclusivamente un identificador para poder enlazar las valoraciones de un mismo usuario. Todo el resto de información había sido eliminada, por lo que en Netflix consideraron que el conjunto de los datos garantizaba la privacidad de los usuarios.

La finalidad de los datos era facilitar a los investigadores que participaban en un concurso que tenía por objeto definir un algoritmo de recomendaciones de películas para los clientes de Netflix para facilitar en su página una mejora de la experiencia del usuario con información adicional de recomendaciones del tipo «a quienes les ha gustado una película». Por ejemplo, a quienes les ha gustado *Minority Report* también les ha gustado *Matrix* (algo que vemos con frecuencia en los grandes portales de venta de productos o servicios). Los premios ofrecidos eran importantes,

por lo que captó el interés de la comunidad científica y estudiantil.

Un grupo de investigadores de la Universidad de Tejas liderado por Arvind Narayanan y el Profesor Vitaly Shmatikov identificaron que un atacante que solo conociera un poco sobre un suscriptor individual podría identificar fácilmente el registro de ese suscriptor si este estaba presente en el conjunto de datos y sus valoraciones, o al menos identificar un conjunto de datos dentro del cual estarían las recomendaciones de ese suscriptor.

El trabajo de investigación que publicaron incluía innumerables ejemplos de lo fácil que era identificar a los suscriptores. Para ello solo sería necesario conocer sus puntuaciones para seis películas no incluidas en el *top* 5.

Para mostrar estos resultados abstractos en ejemplos concretos, Narayanan y Shmatikov compararon los datos de calificación de Netflix obtenidos con datos de bases de datos de películas similares disponibles en Internet que publican también valoraciones (Internet Movie Database), lo que facilitaba la identificación de los subscriptores. En este contexto era más fácil identificar a aquellos cuyos gustos eran más singulares.

Posteriormente, la compañía lanzó un segundo concurso que incluía datos demográficos y de comportamiento con información como: edad, género, código postal, las calificaciones realizadas y las películas elegidas. A finales de 2009 los clientes de Netflix formalizaron una demanda colectiva contra la empresa por violación de su privacidad, lo que llevo a la desestimación del segundo concurso.

Técnicas de reidentificación de datos

Los casos vistos como ejemplo tienen un denominador común. En todos los casos la reidentificación de las personas se produjo gracias a contar con información adicional: vinculación de dos conjuntos de datos. Es posible vincular datos incluso en los casos en los que se eliminan los datos que identifican directamente y aquellos que podemos considerar cuasiidentificadores. Donde se encuentra un importante caladero de información es en Internet y las redes sociales.

En un proceso de reidentificación, tendremos:

- **Adversario:** los modelos de anonimización que se utilizan en el ámbito de la ciencia computacional se nutren, entre otros, de la teoría de juegos. Es un agente motivado, es decir, suficientemente interesado por la reidentificación, y actúa con base en el principio de esfuerzo-beneficio.

- **Información externa:** siempre que el adversario accede a una fuente de información, busca información externa con la que puede vincular la información.

- **Cruce de datos *(inner-joins)*:** se cruza la información de diferentes fuentes de datos, por lo que se incrementan los datos disponibles y las posibilidades de reidentificación. Se trata de conectar filas de un conjunto de datos con filas de otro conjunto de datos, casando datos que comparten o coinciden en los dos conjuntos de datos.

Por ejemplo, veamos este conjunto de datos:

CP	Fecha de nacimiento	Sexo	Enfermedad
28001	12/01/2005	M	Hepatitis C
28002	25/05/1936	H	Lupus
28005	25/05/1936	M	Fibromialgia
28004	22/01/1950	H	Hipertensión
28001	11/07/1947	M	Cáncer de colon

Tabla 22. Ejemplo reidentificación I. Fuente: elaboración propia.

Estarían deidentificados, pero si disponemos de un conjunto de datos como el siguiente:

Nombre	Dirección	Fecha de nacimiento
Juan	xxxxxxxxxx	13/12/1955
Pedro	xxxxxxxxxx	12/02/1955
Daniel	xxxxxxxxxx	10/01/1955
María	xxxxxxxxxx	25/05/1936
Jorge	xxxxxxxxxx	21/12/1955
Juan	xxxxxxxxxx	25/05/1936

Tabla 23. Ejemplo reidentificación II. Fuente: elaboración propia.

Vemos que la fecha de nacimiento está presente en los dos conjuntos de datos. Cruzando las tablas tenemos *inner-join* por **fecha.**

Nombre	Dirección	Fecha de nacimiento	CP	Fecha de nacimiento	Sexo	Enfermedad
Juan	xxxxxxxxxx	13/12/1955				
Pedro	xxxxxxxxxx	12/02/1955				
Daniel	xxxxxxxxxx	10/01/1955				
María	xxxxxxxxxx	25/05/1936	28002	25/05/1936	M	Fibromialgia
María	xxxxxxxxxx	25/05/1936	28005	25/05/1936	H	Lupus
Jorge	xxxxxxxxxx	21/12/1955				
Juan	xxxxxxxxxx	25/05/1936				
			28001	12/01/2005	M	Hepatitis C
			28004	22/01/1950	H	Hipertensión
			28001	11/07/1947	M	Cáncer de colon

Tabla 24. Ejemplo reidentificación III. Fuente: elaboración propia.

Nos identifica dos posibles candidatos. Si además sabemos que María es un nombre fundamentalmente de mujer y disponemos del sexo, podremos afirmar con alta probabilidad que María padece fibromialgia.

Este modelo expuesto de reidentificación responde al siguiente esquema:

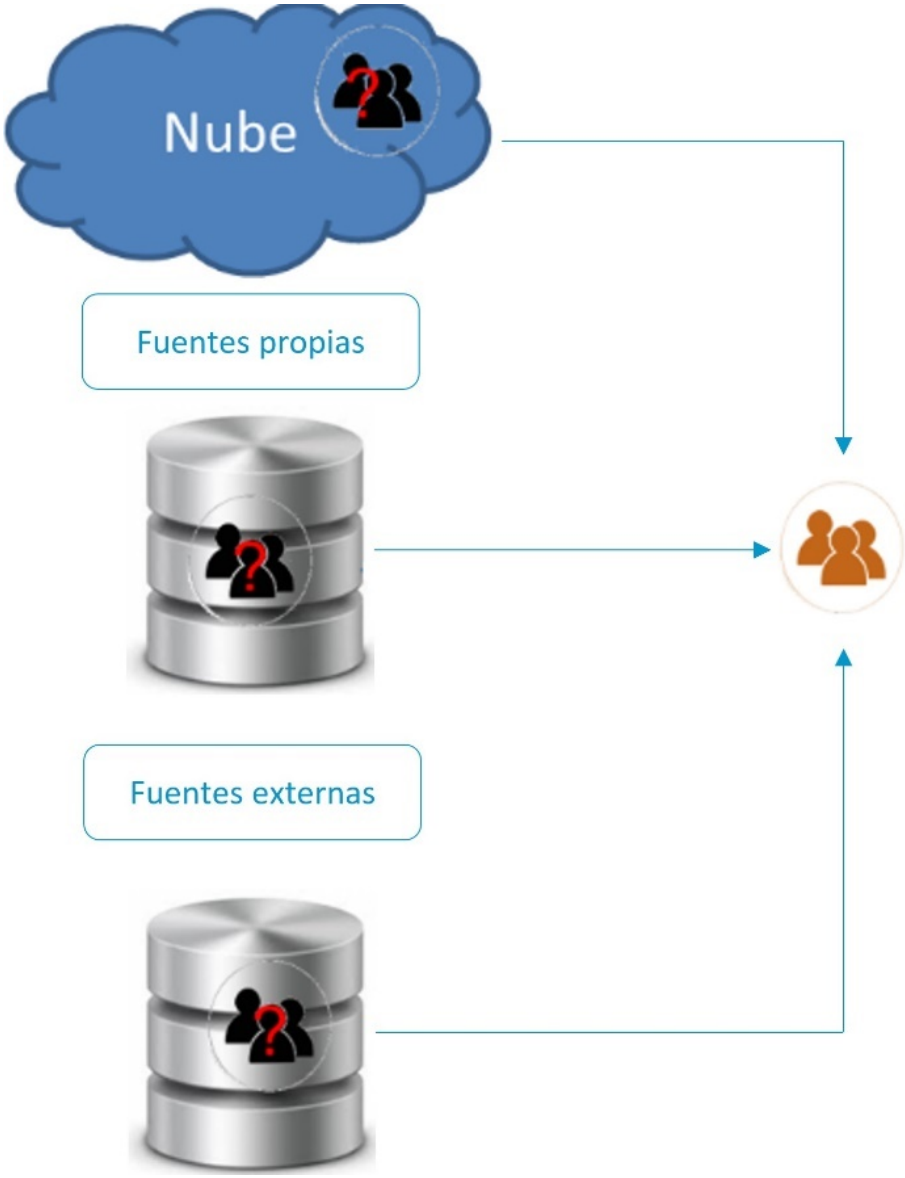


Figura 3. Esquema diferentes fuentes de datos. Fuente: elaboración propia.

Ideas clave

Es fácil determinar que el riesgo inherente de reidentificación de cualquier conjunto de datos hoy en día es muy superior al que el mismo conjunto de datos estaría sometido en los años 90 y muy probablemente muy inferior al que tendrá en los próximos años, si tenemos en cuenta que disponemos de:

- Tecnología que es capaz de tratar información no estructurada, facilitando vinculaciones de fuentes diversas.
- Elevadas capacidades de tratamiento para realizar operaciones.
- Internet y el uso extensivo de las redes sociales como fuentes de información.
- Hiperconectividad.
- Las iniciativas de datos abiertos.

Por otra parte, técnicas que en su momento fueron consideradas suficientes como la deidentificación de los datos hoy dejan de ser válidas aplicadas de forma independiente, por lo que requieren la combinación de otras que mitiguen sus riesgos inherentes.

A continuación, identificaremos los riesgos de algunas de las técnicas de anonimización más utilizadas.

Riesgos de las técnicas de anonimización

En opinión del WP29, en su «Opinion 05/2014 on Anonymisation Techniques», adoptada el 10 de abril de 2014, todas las técnicas tienen debilidades que se deben considerar seriamente antes de que una determinada técnica se utilice para diseñar un proceso de anonimización por parte del responsable de tratamiento.

Se deben tener en cuenta los **fines** que se persiguen con la anonimización, tales como la protección de la privacidad de los individuos cuando se hace público o accesible a terceros un conjunto de datos o el hecho de no facilitar el acceso a determinada pieza de información cuando se accede a un conjunto de datos.

Tras la evaluación determinaremos cuál es la mejor solución, que seguramente no será única, sino que tendremos que aplicar una combinación de técnicas que respondan al problema de una forma más eficaz atendiendo a las limitaciones de las técnicas de anonimización.

Según el grupo de trabajo del artículo 29 (WP29), hay tres cuestiones clave a la hora de valorar la robustez de una técnica:

- ¿Es posible destacar un individuo dentro del conjunto de datos, por ejemplo, por un valor característico de uno de sus atributos?
- ¿Sigue siendo posible vincular los registros relativos a un individuo con otras fuentes u obtener la identidad del individuo?
- ¿Se puede inferir la información relativa a una persona desde el conjunto de los datos?

Atendiendo a estas cuestiones, los **tres riesgos** de las técnicas de anonimización que propone el WP29 son:

- Señalamiento o singularización *(singling out):* corresponde a la posibilidad de aislar algunos o todos los registros que identifican a un individuo en el conjunto de datos.
- Vinculabilidad *(linkability):* la capacidad de vincular al menos dos registros referentes al mismo interesado o a un grupo de interesados (ya sea en la misma base de datos o en dos diferentes). Si un atacante puede establecer (por ejemplo, por medio de un análisis de correlación) que dos registros se asignan a un mismo grupo de personas, pero no pueden distinguir individuos en el grupo, la técnica proporciona resistencia contra *singling out,* pero no contra *linkability*.
- Inferencia *(inference):* deducir con significativa probabilidad el valor de un atributo de entre los valores de un conjunto de otros atributos.

Cada técnica no cumple con total garantía los criterios de eficacia de la anonimización, sino que presenta riesgos. Prácticamente podemos afirmar que habiendo microdatos hay riesgo de reidentificación.

Para facilitar una primera aproximación a esta visión de riesgo y conveniencia de las diferentes técnicas, se propone la siguiente tabla.

	Riesgo de singularización	Riesgo de vinculabilidad	Riesgo de inferencia
Seudonimización	Sí	Sí	Sí
Adición de ruido	Sí	Puede que no	Puede que no
Sustitución	Sí	Sí	Puede que no
Agregación y k-anonimato	No	Sí	Sí
l-Diversity/t-Closeness	No	Sí	Puede que no
Privacidad diferencial	Puede que no	Puede que no	Puede que no
Hash/Tokens	Sí	Sí	Puede que no

Tabla 25. Riesgo de reidentificación. Fuente: Art. 29 WP, 2014.

Consejos para reducir el riesgo de reidentificación

Desde el WP29, en su «Opinion 05/2014 on Anonymisation Techniques», se nos proponen un conjunto de recomendaciones para aplicar con independencia de la técnica utilizada para reducir el riesgo de identificación. Se aconseja, en general:

- No confiar en un enfoque de «liberar y olvidar», considerando siempre el riesgo residual de identificación que los datos tratados pueden tener, manteniendo por parte de los responsables de tratamiento las siguientes prácticas:
  - Identificar nuevos riesgos o reevaluar los riesgos periódicamente.
  - Evaluar si los controles para los riesgos identificados son suficientes y se ajustan en consecuencia.
  - Monitorizar y controlar el riesgo.
- Como parte de estos riesgos residuales es importante tener en cuenta la potencial identificación de la parte no anonimizada de un conjunto de datos (si la hay), especialmente cuando se combina con la parte anónima, además de las posibles correlaciones entre los atributos (por ejemplo, entre la ubicación geográfica y los datos del nivel de riqueza).

Estas recomendaciones se enlazan perfectamente con la **evaluación de impacto** en la protección de datos personales. Integra la incorporación de un proceso de evaluación continuo del riesgo inherente y de las medidas implantadas, lo que sería aplicable tanto a los tratamientos de datos personales como a los datos anonimizados o disociados.

Siguiendo con las recomendaciones del WP29, se deben establecer claramente los objetivos que deben alcanzarse a través del conjunto de datos anónimos, ya que desempeñan un papel clave en la determinación del riesgo de identificación.

Esto se une a la necesaria consideración de todos los elementos relevantes del contexto. Por ejemplo: la naturaleza de los datos originales, los mecanismos de control establecidos (incluidas las medidas de seguridad para restringir el acceso a las bases de datos), el tamaño de la muestra (características cuantitativas), la disponibilidad de recursos de información públicos (que puedan ser utilizados por los destinatarios), la previsión de liberación de datos a terceros (limitado, ilimitado en Internet, etc.).

Además, se debe prestar atención a los posibles atacantes, teniendo en cuenta el atractivo de los datos para los ataques dirigidos (de nuevo, la sensibilidad de la información y la naturaleza de los datos serán factores clave en este sentido).

En resumen, podemos concluir:

- La dificultad de conseguir conjuntos disociados de datos (perfectamente anonimizados).
- Los riesgos de un conjunto de datos anonimizados pueden verse alterados en el futuro, por ello se recomienda la reevaluación periódica de los mismos.

10.8. Principios a la hora de construir un data warehouse

Por último, y para terminar este tema, vamos a proponer un conjunto de principios que podemos seguir a la hora de construir un *data warehouse* que reduzca la exposición a un evento de revelación de datos de una forma sencilla. Estos principios son:

- **Separaciónfuncional:** facilita que el acceso a los datos se realice exclusivamente a aquellos para los que se está autorizado de acuerdo con las funciones desarrolladas, quedando imposibilitado el acceso a información adicional.

- **Agregación de datos:** siempre que sea posible y no interfiera en la finalidad del conjunto de datos es conveniente proceder a la agregación de datos de manera que se complique la posibilidad de reidentificar al individuo. Siempre debemos considerar que los registros no agregados suelen hacer más fácil la reidentificación de registro de datos agregados.

- **Deidentificación de los registros:** eliminación de identificadores directos e indirectos en la medida de lo posible. Por su parte, sustituir los identificadores por seudónimos introducirá una barrera adicional. Las expectativas sobre el nivel de anonimización de un conjunto de datos después de eliminar los identificadores directos o indirectos o sustituir estos por seudónimos no deben ser maximalistas, pues son evidentes los riesgos que aún tendrá el conjunto de datos sobre la potencial reidentificación del individuo. Por ejemplo, a través de la combinación de un conjunto de atributos. Estos son los casos de:

- Datos de edad.

- Geolocalizadores: código postal, datos de ubicaciones en mapas.

- Sexo.

- Información adicional: en una historia clínica podrías tener códigos de diagnósticos, sobre todo los menos frecuentes (como enfermedades raras).
- Información sobre educación: formación o titulaciones poco usuales.
- Información sobre profesión: profesiones poco usuales.
- Raza.
- Ingresos.
- Religión.
- Indicadores socioeconómicos.
- Y otros.

Para algunos de estos valores contaremos con una distribución relativamente uniforme para el conjunto de la población. La presencia de este dato no facilita la identificación del individuo, pero combinados disponen de una mayor capacidad en ese sentido.

Por ejemplo, el dato sobre el sexo no aporta mucho. Sin embargo, sexo con edad más geolocalización y diagnóstico clínico puede generar una mayor concreción sobre el potencial del individuo al que se refiere.

Requerirá un esfuerzo significativo eliminar esta información del conjunto de datos, analizando en cada caso la posibilidad de su eliminación o agregación, atendiendo al objeto del estudio y cómo esta acción puede perturbar las conclusiones del estudio y en qué manera. Aquí adquiere especial relevancia la finalidad del uso al que están destinado estos datos.

- **Aplicación de técnicas de anonimización** sobre microdatos: combinar las diferentes técnicas de anonimización con el objeto de reducir riesgos.

Para orientar qué técnica aplicar, podemos consultar la siguiente tabla.

Técnica	Cuándo usar	Tipo de atributo
Supresión de atributos.	Cuando los atributos no son necesarios.	Todos.
Supresión de registros.	Presencia de registros atípicos.	Aplica a todo el registro.
Enmascaramiento de caracteres.	Enmascarar algunos caracteres en un atributo proporciona suficiente anonimato. Conservamos partes del valor original.	Identificador directo.
Pseudonimización.	Debemos poder distinguir los registros, pero no debemos conservar ninguna parte del valor del atributo original.	Identificador directo.
Generalización.	Modificamos los registros para ser menos precisos pero útiles.	Todos.
Permutación de datos.	Aplicable cuando no es necesario analizar las relaciones entre los atributos en el nivel de registro.	Todos.
Perturbación de datos (aleatorización-adición de ruido).	Ligeras modificaciones en los atributos no alteran el estudio, reducen riesgo de reidentificación a través del acceso a valores reales.	Identificadores directos o datos de análisis.
Agregación de datos.	Los datos agregados son suficientes para la utilidad del estudio. No se necesitan presentar registros individualizados.	Identificadores indirectos.

Tabla 26. Técnicas de anonimización. Fuente: elaboración propia.

Ideas clave

Un **identificador directo** es un atributo que por sí solo identifica a un individuo (DNI, huella dactilar…); mientras que un **identificador indirecto** o cuasiidentificador es un atributo que por sí mismo no identifica, pero combinado con otra información puede que sí.

Debemos tener presente que aplicar técnicas de anonimización reduce la utilidad de los datos, se pierde carga de información.

Por tanto, elegir la técnica más adecuada va a requerir determinar qué utilidad necesitamos que mantengan los datos para el análisis o estudio que se vaya a realizar.

Por ejemplo, si un estudio requiere amplia precisión geográfica, tal vez no podamos generalizar el CP, incrementando posiblemente el riesgo de reidentificación.

10.9. Referencias bibliográficas

Article 29 Working Party. (2013). Opinion 05/2014 on anonymisation techniques [Archivo PDF].
https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Ley 14/2007, de 3 de julio, de Investigación biomédica. Boletín Oficial del Estado, 4 de julio de 2007, núm. 159, pp. 28826-28848.
https://www.boe.es/eli/es/l/2007/07/03/14

Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. Boletín Oficial del Estado, 14 de diciembre de 1999, núm. 298.
https://www.boe.es/eli/es/lo/1999/12/13/15/con

Real Decreto 1720/2007, de 21 de diciembre, por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal. Boletín Oficial del Estado, 19 de enero de 2008, núm. 17.
https://www.boe.es/eli/es/rd/2007/12/21/1720/con

Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento General de Protección de Datos). Diario Oficial de la Unión Europea L 119, 4 de mayo de 2016, pp. 1-88.
https://www.boe.es/doue/2016/119/L00001-00088.pdf

Sweeney, L. (2019). Latanya Sweeney, Ph.D. [Página web].
http://latanyasweeney.org/

Opinion 05/2014 on anonymisation techniques

Article 29 Working Party. (2013). Opinion 05/2014 on anonymisation techniques [Archivo PDF].
https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Documento del grupo de trabajo del artículo 29 (WP29) con consideraciones sobre las técnicas de anonimización de datos que facilita un entendimiento sobre los riesgos asociados a las mismas e incorpora recomendaciones para el tratamiento del riesgo residual.

Código de buenas prácticas para la gestión de riesgos derivados de los procesos de disociación

ICO. (2012). Anonymisation: managing data protection risk code of practice [Archivo PDF].

https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf

Documento de la ICO en inglés.

Página web del grupo de trabajo del artículo 29

European Comission. (s.f.). Data protection [Página web].

http://ec.europa.eu/justice/data-protection/index_en.htm

En la página web del grupo de trabajo del artículo 29 se puede acceder a toda la documentación. Son especialmente interesantes las opiniones que consolidan la visión interpretativa de la autoridad de protección de datos de la UE.

Test

1. La ley de protección de datos no se aplica sobre los datos disociados obtenidos tras un proceso de disociación.

A. Verdadero.

B. Falso.

2. ¿Cuáles de las siguientes técnicas no es de disociación-anonimización?

A. Asignación al azar.

B. Generalización.

C. Adición de ruido.

D. T-Combinación.

3. ¿Cuáles de los siguientes son riesgos asociados a las técnicas de anonimización?

A. Singularización.

B. Vinculabilidad.

C. Inferencia.

D. Todos los anteriores son correctos.

4. La generalización es una técnica del tipo:

A. Técnicas que implican la modificación de datos.

B. Técnicas que implican la reducción de atributos.

C. Técnicas que implican la supresión de datos.

D. Técnica de pseudoanonimización.

5. La Dra. Sweeney concluyó que en EE. UU. es posible identificar al 87 % de los ciudadanos unívocamente, a partir de los datos de:

A. El número de la seguridad social.

B. El número de licencia de conducir y el estado emisor.

C. El código postal completo, sexo y fecha de nacimiento con el año incluido.

D. El código postal completo, sexo, fecha de nacimiento con el año incluido y raza.

6. ¿Cuál de los siguientes riesgos no es de reidentificación?

A. Riesgo de singularización.

B. Riesgo de vinculabilidad.

C. Riesgo de inferencia.

D. Riesgo de polarización.

7. De entre las siguientes técnicas, ¿cuál ofrece mejores resultados en términos de menores riesgos de reidentificación?

A. Seudonimización.

B. Adición de ruido.

C. K-anonimato.

D. l-Diversity/t-Closeness.

8. De entre las siguientes técnicas, ¿cuál ofrece los peores resultados en términos de menores riesgos de reidentificación?

A. Seudonimización.

B. Adición de ruido.

C. K-anonimato.

D. l-Diversity/t-Closeness.

9. Teniendo en consideración el siguiente conjunto de datos, el resultado de aplicar k-anonimato, indica cuál es valor *k* del conjunto:

A. k = 2.

B. k = 3.

C. k = 4.

D. k = 5.

10. Teniendo en consideración el siguiente conjunto de datos, el resultado de aplicar l-Diversity, indica cuál es el valor *l* del conjunto:

A. l = 1.

B. l = 2.

C. l = 3.

D. l = 4.