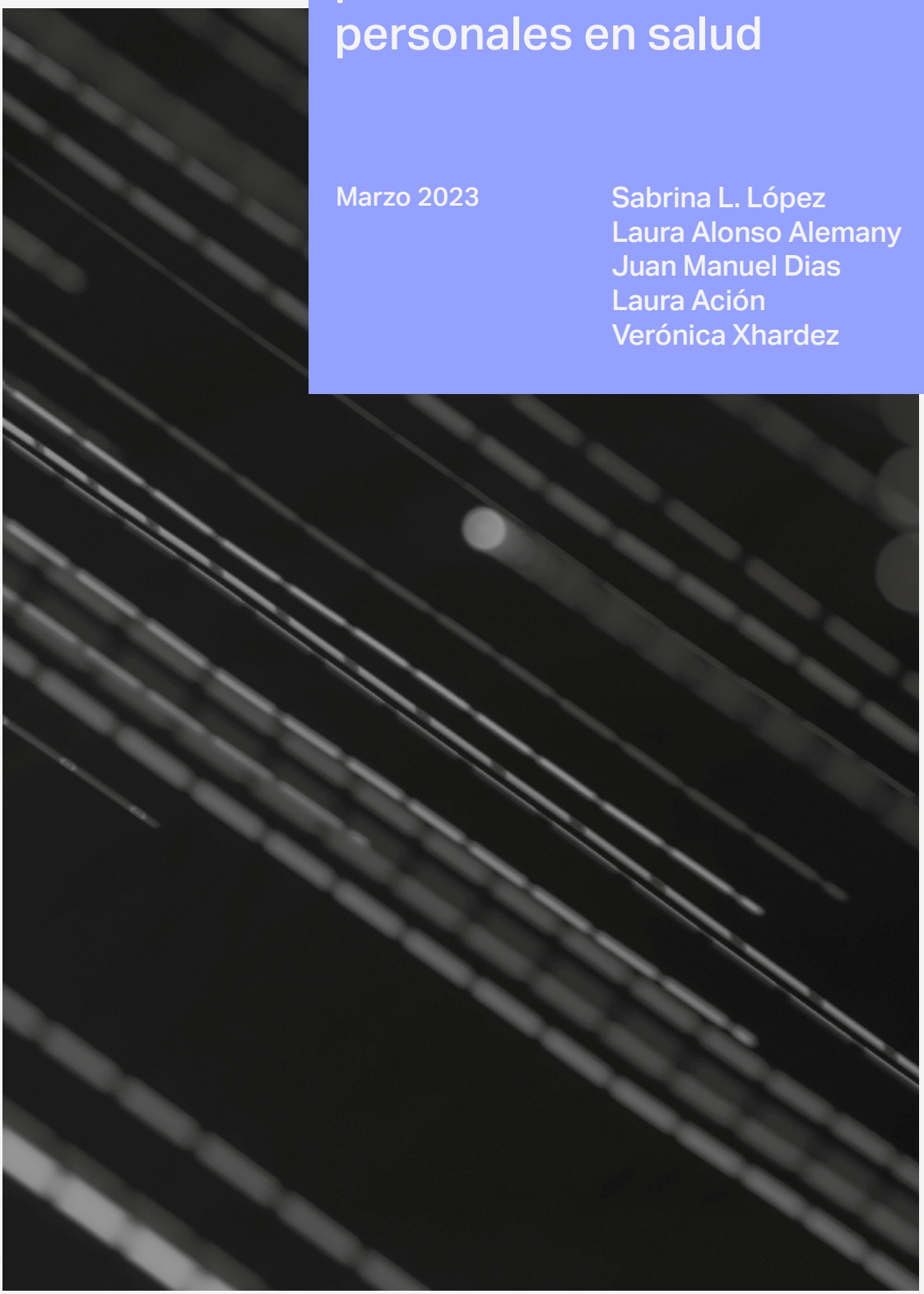
fundar

CIECTI
Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación

ARPHAI
Gestión epidemiológica basada en inteligencia artificial y ciencia de datos

Guía práctica para la protección de datos personales en salud

Datos



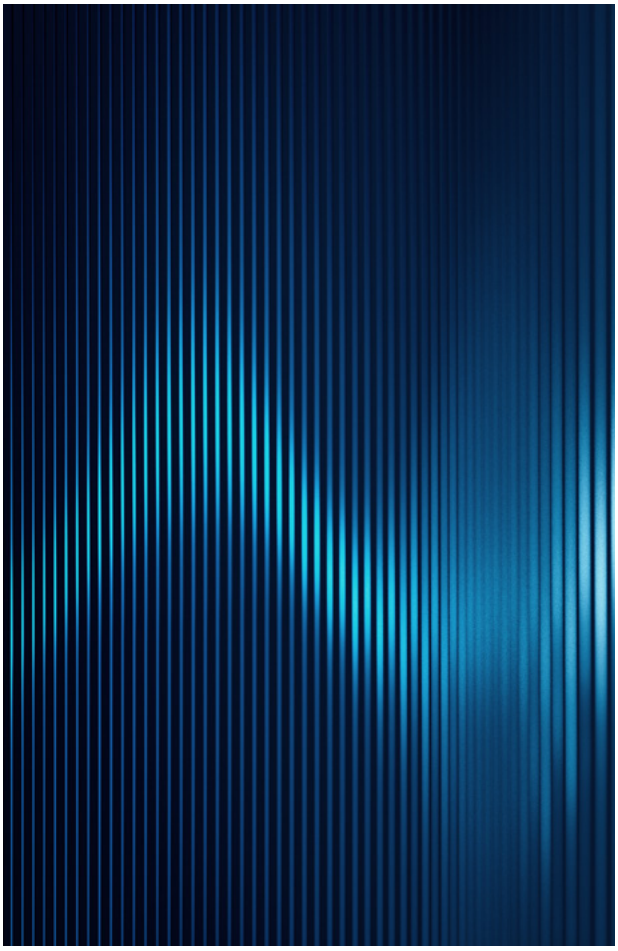
Marzo 2023

Sabrina L. López
Laura Alonso Alemany
Juan Manuel Dias
Laura Ación
Verónica Xhardez

Introducción

En la actualidad existe gran cantidad de información digitalizada de diversos dominios. Por esto, las disciplinas de extracción de información y análisis de datos son áreas de estudio que han crecido y demandan mucho interés. En el ámbito del cuidado de la salud existen numerosos documentos de diversos tipos (informes radiológicos, historias clínicas, informes de ecocardiogramas, etc). Son documentos que proveen información muy valiosa para la detección y caracterización de enfermedades, así también como datos relativos al paciente: edad, género, residencia, etc. Esta información puede usarse para estudios epidemiológicos y descriptivos.

No obstante, los datos de salud de las personas son considerados datos sensibles[1]. El concepto de "dato sensible" incluye todos aquellos datos que puedan causar discriminación o estigmatización, aun de manera potencial (Pérez Ponte)[2]. Por esta razón es clave el tratamiento que hagamos sobre estos datos para evitar que la información pueda ser relacionada a una persona.



→

Objetivo

Con este documento pretendemos brindar un conjunto de recomendaciones, metodologías y herramientas para el tratamiento responsable de los datos de salud. En tal sentido, nos detendremos en distintos escenarios de uso y sugeriremos en cada contexto de aplicación un tratamiento adecuado de los datos.

1 Para ampliar sobre este tema sugerimos consultar el siguiente documento elaborado conjuntamente entre ARPHAI y FUNDAR: Datos digitales de salud: ¿parte de la solución y del problema?

2 Ver más en Protección de datos personales y el tratamiento de los datos de la salud.

→ 3

? →

A quién va dirigido

A personas que trabajan con datos de salud y quieren mejorar las garantías de protección de datos personales. No presentamos recomendaciones de seguridad en sistemas de información, sino que nos centramos en los procesos de anonimización de estos datos, especialmente datos estructurados.

Este documento te puede ayudar si sos:

1 → Una persona a cargo del sistema informático de una institución, a cargo de tareas como administrar una base de datos, resolver problemas técnicos o actualizar el sistema de gestión, o si aportás tu experiencia en la decisión de cómo implementar mejoras en el sistema, opinás sobre diferentes alternativas y muchas veces implementás la solución elegida.

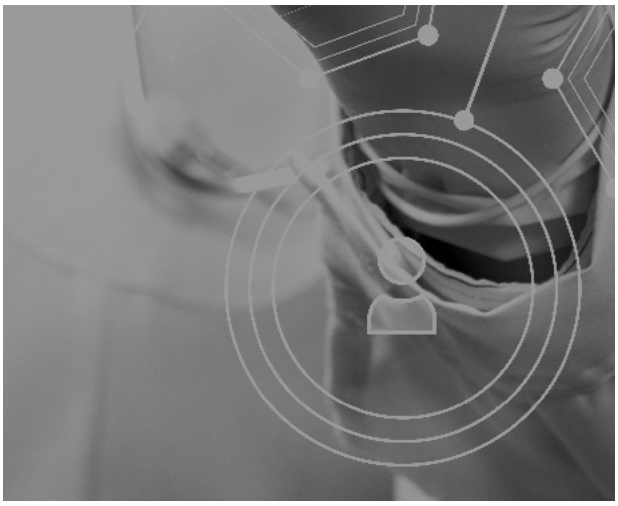
2 → Una persona en un pequeño equipo de sistemas, que hace seguimiento de las diferentes herramientas que funcionan en tu institución, relevando problemas y limitaciones, evaluando diferentes herramientas y creando algunas nuevas, implementando y dando seguimiento a las soluciones propuestas.

3 → Una persona especializada en el tratamiento de datos de salud, con especial atención a garantías de anonimización y tratamiento de datos no estructurados, o a cargo del intercambio de datos entre diferentes dependencias e instituciones.

→

También será un aporte para quienes trabajan en el ámbito de la salud y necesitan compartir información hacia fuera de sus instituciones de pertenencia. Cabe aclarar que en este documento no nos detendremos en describir bajo qué condiciones es posible compartir la información, ya que damos por sentado que esas condiciones se cumplen[3], en particular y sobre todo en relación al consentimiento de las personas titulares de los datos.

3 Ley 25.326 de PROTECCIÓN DE LOS DATOS PERSONALES, artículo 11 (cesión)



1 →

¿Cómo y por qué anonimizar?

Es preciso anonimizar datos de salud cuando necesitamos garantizar **que no se pueda identificar a una persona junto con información sensible**, por ejemplo información sobre su historia de salud. Por lo tanto, debemos garantizar que no se puede identificar a una persona con la información que disponibilicemos. Se puede identificar a una persona con datos como su número de DNI, su nombre y apellido, su dirección, su teléfono. Todos estos datos deberán someterse a un proceso de anonimización antes de que se disponibilice información sensible asociada a una persona, como por ejemplo su historia clínica.

A veces **la combinación de informaciones hace posible identificar** a una persona de una forma que sería imposible si estos datos se dieran por separado. Por ejemplo, la fecha de nacimiento, el código postal y el género resultaron suficientes para identificar los nombres de las personas (consultando información electoral y otras fuentes) (Sweeney et al., 2013). De hecho, en otro trabajo, Sweeney (2000) estimó que un 87% de la población de los Estados Unidos podía identificarse en forma inequívoca en función de esos tres datos (Yankelevich, 2021).

Ante la duda de si se puede identificar a una persona con cierta información, conviene aplicar el **principio precautorio** y no disponibilizar la información. Por ejemplo, en el caso de enfermedades poco frecuentes, es posible que se pueda identificar a la única persona de una localidad con esa enfermedad, incluso si no se asocia a ningún dato personal.

Finalmente, en todos los casos, es imprescindible tener planificada la **destrucción de los datos** una vez que han sido utilizados para el fin por el que fueron solicitados.

2 →

Principios imprescindibles
(antes de anonimizar)

Los datos de salud tienen que estar en un sistema seguro, con garantías, y respaldo. También con una estructura de permisos según el rol de usuario, contraseñas, encriptación y servidores seguros.

Como mencionamos previamente, a veces la combinación de ciertos datos e informaciones hace posible identificar a una persona de una forma que sería imposible si esos datos se dieran por separado. Por esta razón, reducir la disponibilización de información al mínimo imprescindible para satisfacer una necesidad de información aumenta las garantías de no identificación. Este principio impacta en el diseño del sistema de información subyacente, ya sea en el registro, en el tipo de pedidos que se pueden hacer y, principalmente, en la información que se facilita al responder a un pedido.

Merece especial atención el compartir datos, o disponibilizar datos que se recolectaron con un objetivo primario (por ejemplo, asistencial) para otros fines (por ejemplo, investigación, planificación). En un contexto distinto al primario, pueden darse interacciones con otras fuentes de información o con otras aplicaciones que resulten en efectos no deseados, ya sea de identificación, discriminación u otros. Por esta razón, es muy importante evaluar cada nuevo escenario de uso de los datos con la complejidad necesaria para resguardar la información sensible.

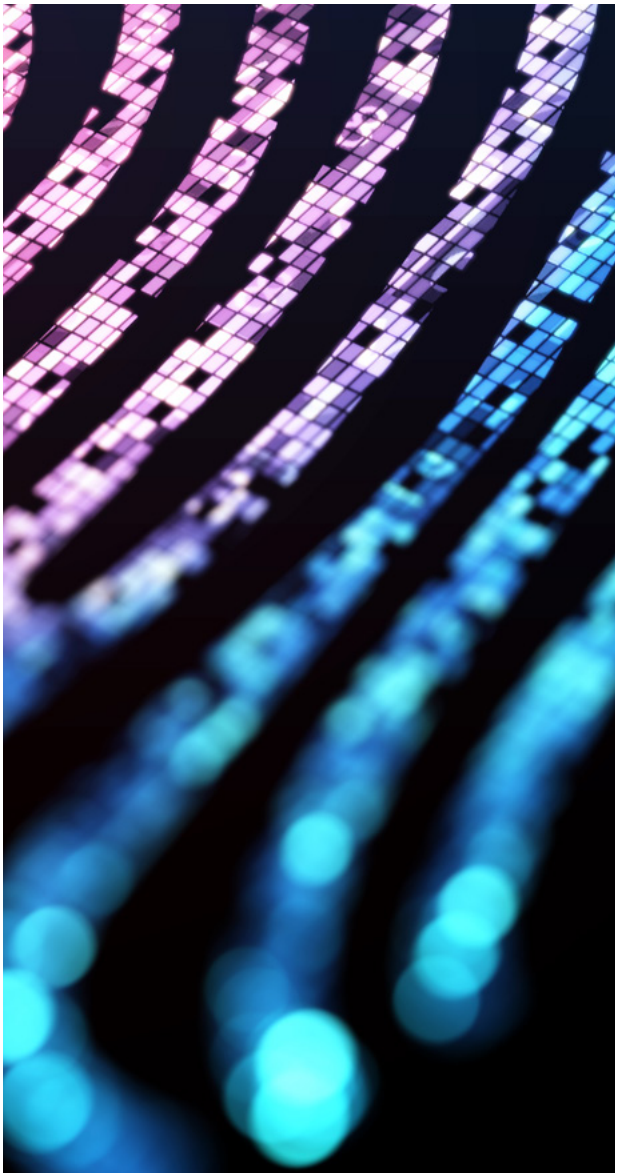
3

¿Qué información identifica a una persona?

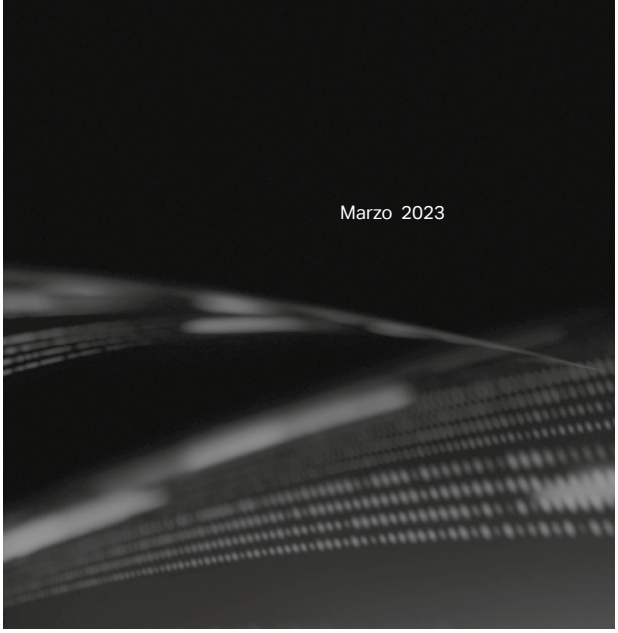
Se denominan identificadores directos a aquellos datos que permiten identificar de forma directa a una persona, por ejemplo a través de un documento de identificación oficial como es el número de DNI, el nombre completo, dirección, correo electrónico o el código de historia clínica. Pero hay otras informaciones que también pueden contribuir a la identificación de las personas en registros de atención médica, a los que se denomina identificadores indirectos, como son la edad, género, relaciones familiares, zona de origen o de residencia, números de teléfono, matrícula, etc[4]. Incluso pueden contribuir a la identificación de una persona las informaciones sobre prácticas médicas realizadas, diagnóstico y síntomas, sobre todo si son poco frecuentes. También existen datos descriptivos de los pacientes que pueden ser únicos, como por ejemplo, la vascularidad del ojo, que es un dato que se usa para prevenir y diagnosticar enfermedades degenerativas muy frecuentes como la retinopatía diabética.

Este tipo de información puede encontrarse en campos estructurados de las historias clínicas electrónicas (HCE), por dar un ejemplo. Si queremos usar este tipo de datos se debe realizar procedimientos para eliminar información que no sea estrictamente necesaria en orden de satisfacer una necesidad determinada. Si bien este trabajo se concentra en campos estructurados, los campos de texto libre en registros de atención presentan una mayor dificultad.

Por esta razón, si se estima necesario el acceso a datos de texto libre, por ejemplo, porque contienen datos sobre atenciones que no se encuentran en los datos estructurados, se deberá realizar un proceso de detección y reemplazo de identificadores, ya sea manual o automático, a modo de minimizar los riesgos de identificación de los usuarios.



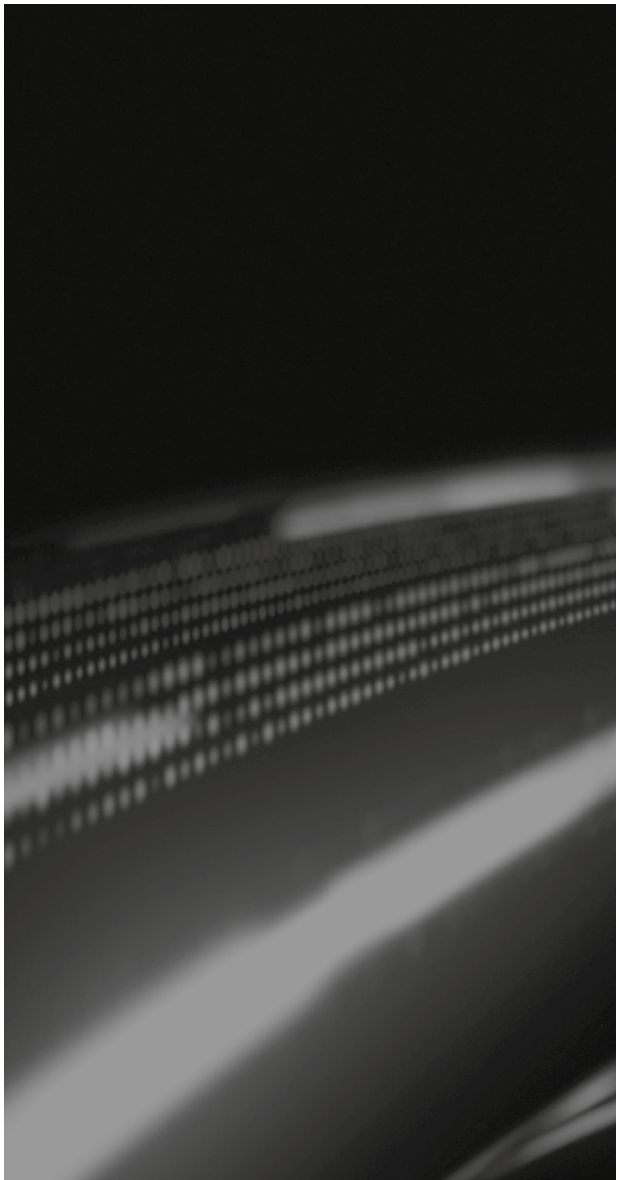
4 Aquí pueden encontrarse los lineamientos de HIPAA, la ley de Estados Unidos para protección de datos personales en salud.



4

Pero, en concreto, ¿cómo anonimizamos?

¿Qué es anonimización? Hay muchos trabajos sobre el tema, entre ellos un documento de FUNDAR. Además, en la actualidad, está en agenda la propuesta de anteproyecto de ley de protección de datos personales: nos parece una buena oportunidad tomar de ese trabajo dos definiciones: datos personales sensibles y anonimización.



Datos personales sensibles

Aquellos que se refieran a la esfera íntima de su Titular, o cuya utilización indebida puedan dar origen a discriminación o conlleve un riesgo grave para este. De manera enunciativa, se consideran sensibles los datos personales que puedan revelar aspectos como origen racial o étnico; creencias o convicciones religiosas, filosóficas y morales; afiliación sindical u opiniones políticas; datos relativos a la salud, discapacidad, a la preferencia u orientación sexual, datos genéticos o datos biométricos dirigidos a identificar de manera unívoca a una persona humana.

Anonimización

La aplicación de medidas dirigidas a impedir la identificación o reidentificación de una persona humana, sin esfuerzos o plazos desproporcionados o inviables, teniendo en cuenta factores como los costos y el tiempo necesario para la identificación o reidentificación de la persona a la luz de la tecnología disponible en el momento del tratamiento.

De acuerdo con el uso que se va a hacer de la información solicitada, los datos a compartir tendrán ciertas características que a su vez definirán las estrategias y herramientas a utilizar para desidentificarlos. A continuación mencionamos algunos escenarios posibles.

Escenario 1

Microgestión

Una pediatra pide datos de sus pacientes al departamento de registros de HCE de una Clínica. Pide los datos de la última consulta que realizaron, cuál fue la fecha y motivo de dicha consulta.

Identificador	Edad en la consulta	Fecha de la consulta	Motivo consulta	Motivo de consulta
persona 1	2	xx/xx/xxxx	fiebre	gripe
persona 2	1	xx/xx/xxxx	control niño sano	bajo peso

Para este escenario pueden darse dos situaciones. Por un lado, si los datos que solicita corresponden a consultas con el mismo pediatra, no será necesario realizar anonimización, ya que el mismo pediatra es la fuente de datos. Por otro lado, si los datos que solicita corresponden a todas las consultas que han realizado los pacientes, con cualquier especialidad, tampoco será necesaria la anonimización porque el pediatra ya ha identificado al usuario del cual solicita la información, a partir de su consulta. Para este caso se deberá justificar la necesidad de la información, para preservar el principio de reducir la disponibilización de información al mínimo imprescindible, y su acceso a los datos quedará registrado.

Escenario 2

Mesogestión

El director de un hospital necesita saber cómo planificar los recursos necesarios para las urgencias: cuántos médicos dedicar en cada día y de qué especialidades, según la época del año, día de la semana, condiciones climatológicas, etc. Para ello, solicita los datos de pacientes que concurrieron a la guardia en los últimos 5 años.

Identificador	Edad en la consulta	Fecha de la consulta	Especialidad	Motivo de consulta
persona 1	3	xx/xx/xxxx	pediatría	dolor de oído
persona 2	31	xx/xx/xxxx	ginecología	dismenorrea

En este caso se proveerá la información sin los campos de la Base de Datos estructurada que tienen datos identificatorios. Por defecto, tampoco se proveerá el campo de texto libre de los registros de atención, ya que también puede contener datos identificatorios de los usuarios. Si alguno de estos datos, como por ejemplo la zona de residencia, resulta necesario para la planificación, se deberá justificar su necesidad y quedará registrado el acceso a los datos.

Si se estima necesario el acceso a datos de texto libre, se deberá realizar un proceso de detección y reemplazo de identificadores, ya sea manual o automático, a modo de minimizar los riesgos de identificación de los usuarios.

Escenario 3

Macrogestión

Un cargo jerárquico del Ministerio de Salud requiere datos masivos de prevalencia de ciertos síntomas para identificar brotes de epidemias tempranamente. Este tipo de datos se encuentra típicamente en los campos de texto libre de los registros de atención. En este caso, el mejor escenario posible sería aplicar procesos de extracción de información sobre los textos, de forma de identificar la presencia de los síntomas de interés, y devolver esa información en lugar del texto completo. En el caso de que no fuera posible realizar este tipo de proceso, la recomendación es idéntica que en el escenario de Mesogestión.

Síntoma	Municipio	Edad	Fecha de la consulta
fiebre	Florencio Varela	38	xx/xx/xxxx
odinofagia	Florencio Varela	40	xx/xx/xxxx

Escenario 4

Investigación

Un grupo de investigación solicita información al Ministerio de Salud local sobre personas con cáncer de cuello de útero atendidas en los últimos 5 años, fecha de diagnóstico y prácticas realizadas. Presentan un protocolo aprobado de un comité de ética. En este caso también se recomienda aplicar procesos de extracción de información sobre los textos, de forma de identificar la presencia de los síntomas de interés, y devolver esa información en lugar del texto completo. En el caso de que no fuera posible realizar este tipo de proceso, la recomendación es idéntica que en el escenario de Mesogestión.

ID_persona	Edad	Fecha diagnóstico	Práctica
persona 1	38	xx/xx/xxxx	pap
persona 2	40	xx/xx/xxxx	biopsia
persona 3	50	xx/xx/xxxx	cesárea

Escenario 5

Datos abiertos

La organización quiere publicar bases de datos donde consten edad, cobertura de salud, género y localización geográfica de las personas atendidas en el último año.

Edad en la consulta	Cobertura	Género	Barrio	Edad
0 a 2 años	Sin cobertura	F	A	23
3 a 6 años	Obra Social	M	B	21

En este escenario, la recomendación es generar datos sintéticos con las mismas propiedades estadísticas que los datos originales, pero nunca publicar datos originales, aún con un proceso de anonimización, ya que no tenemos garantías de que esos procesos no permitan la reidentificación de las personas.

Escenario 6

Acceso a la información pública

Todos los organismos del Estado[5] cuentan con información que se presume pública. Cualquier persona puede solicitar dicha información como pedido de Acceso a la Información Pública sin que sea necesario justificar el pedido. En este sentido podemos mencionar el principio de apertura, presente en artículo 1 de la Ley de Derecho de Acceso a la Información Pública 27.275:

Apertura: la información debe ser accesible en formatos electrónicos abiertos, que faciliten su procesamiento por medios automáticos que permitan su reutilización o su redistribución por parte de terceros.

La Dirección Nacional de Epidemiología e Información Estratégica recibe un pedido de información sobre cantidad de casos notificados de HIV, agrupado por sexo, jurisdicción y año[6].

Sexo	Jurisdicción	Año	Cantidad de Casos HIV
Mujer	Chubut	2010	42
Mujer	Chubut	2011	65
Mujer	Chubut	2012	39

En estos casos lo que se evalúa para otorgar la información es el grado de agregación y la posibilidad de que ese tipo de información pueda o no quedar disponible para su reutilización, de acuerdo a su sensibilidad. En algunas oportunidades se suelen excluir combinaciones con un número menor a un cierto número de casos (por ej. 10). Típicamente, el texto libre y los registros individualizados en general no pueden ser objeto de pedidos de acceso a la información pública.

5 La Ley 27275 de DERECHO DE ACCESO A LA INFORMACIÓN PÚBLICA en su artículo 7 detalla los organismos alcanzados por la Ley Nacional. Tener en cuenta además las normativas provinciales

6 Ejemplo tomado del portal de Datos Abiertos, del dataset Notificación de casos de VIH por sexo y jurisdicción

Herramientas para anonimización de datos

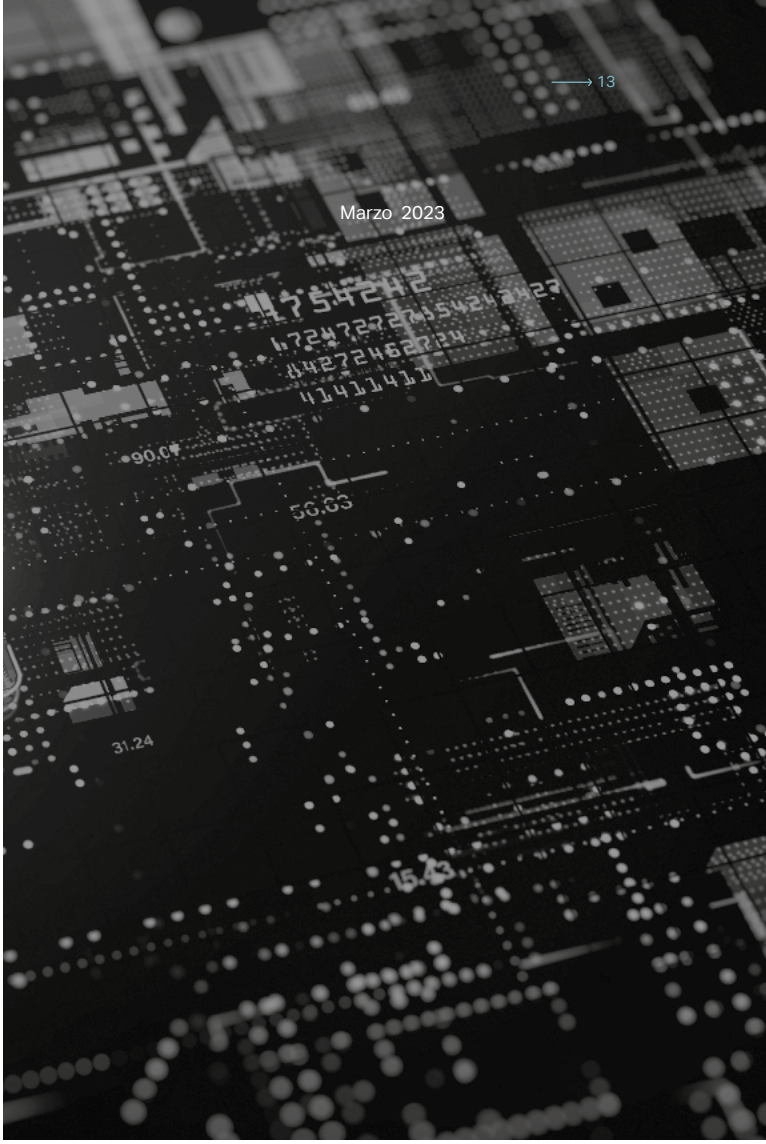
Las distintas técnicas de anonimización que se presentan a continuación nos permiten, por un lado, reducir al máximo posible los riesgos que representa el tratamiento de datos de carácter personal; por otro lado, identificar y ocultar información sensible para que pueda ser reutilizada con distintos fines (políticas públicas, investigación y desarrollo) evitando la vulneración de los derechos de protección de datos de las personas.

Antes de aplicar este tipo de herramientas es importante conocer cuáles son sus alcances y limitaciones y además saber el impacto que pueden tener en el valor analítico de los datos según cada contexto de uso.

En la mayoría de los casos la anonimización absoluta de los datos es prácticamente imposible de lograr. Más allá de esto, es necesario que la identificación de una persona represente un esfuerzo tan grande que sea prácticamente imposible para quien intente hacerlo. En este sentido, para poder utilizar los datos a los fines que nos hayamos propuesto debe darse un equilibrio entre la protección de los mismos y mantener parte de su carácter original, de forma que sigan resultando útiles para que se puedan analizar.

En el documento de Fundar, Anónimos pero no tanto (Yankelevich, 2021), se propone una metodología de gestión eficiente de datos sin poner en riesgo la privacidad, donde el uso de las herramientas que proponemos a continuación queda enmarcado en un flujo de trabajo completo.

Para proteger los identificadores directos en datos estructurados podemos recurrir a su eliminación o en caso de que sean imprescindibles, el *hashing*. El caso del texto libre en datos masivos es más complejo, pero también hay tratamientos posibles.



Hashing (datos estructurados)

Se denomina así al proceso que transforma un dato en otra serie de caracteres de longitud fija, sin importar la longitud original de dicho dato. Existen diversos algoritmos que permiten hacer esta transformación. En el ejemplo, se toma como dato de entrada la columna 'ID_persona' (que luego se elimina), y se utiliza el algoritmo SHA-1 (*Secure Hash Algorithm* 1)[7] para generar la columna 'ID_persona hasheado'.

7 Valores obtenidos en aquí

Fila	ID_persona	ID_persona hasheado	Edad	Fecha diagnóstico	Práctica
1	~~persona 1~~	7c12fba06f5e69b-d65964683d16cb-676f7ebe818	35	xx/xx/xxxx	pap
2	~~persona 1~~	7c12fba06f5e69b-d65964683d16cb-676f7ebe818	35	xx/xx/xxxx	biopsia
3	~~persona 2~~	c95adf64428042ae-d1054aee22824e-dc79f3ec74	39	xx/xx/xxxx	biopsia
...	~~...~~	...	...	...	...
n	~~persona n~~	c4290c48a7b-4919f94d2801d-b4e86b5f5e8b17c6	58	xx/xx/xxxx	biopsia

Anonimización

Una parte importante de los datos que podrían colaborar a mejorar el cuidado del paciente o la investigación médica clínica, es texto libre (en formato no estructurado) por lo que la información es de difícil acceso tanto para las personas como para los sistemas automatizados[8].

A continuación presentamos un ejemplo de procesamiento con una herramienta de anonimización para diferentes entidades que a su vez se puede combinar con expresiones regulares para mejorar su desempeño.

Sin embargo, es importante remarcar que ni estos procesos ni los manuales, garantizan la completa eliminación de identificadores directos e indirectos en el texto libre. Por lo tanto, la información de texto libre deberá ser considerada siempre como información sensible y restringir su acceso a los mínimos estrictamente necesarios, siempre con registro de los accesos.

Procesamiento	Salida
Ninguno	'Nombre: Cecilia Grierson Edad: 24 años Tel: 274534988 DNI: 12345678. Antecedentes: diabetes madre y padre'
Scubadub	'Nombre: {{NAME}} Edad: 24 años {{NAME}}: 27453-4988 DNI: 12345678. {{NAME}}: diabetes madre y padre'

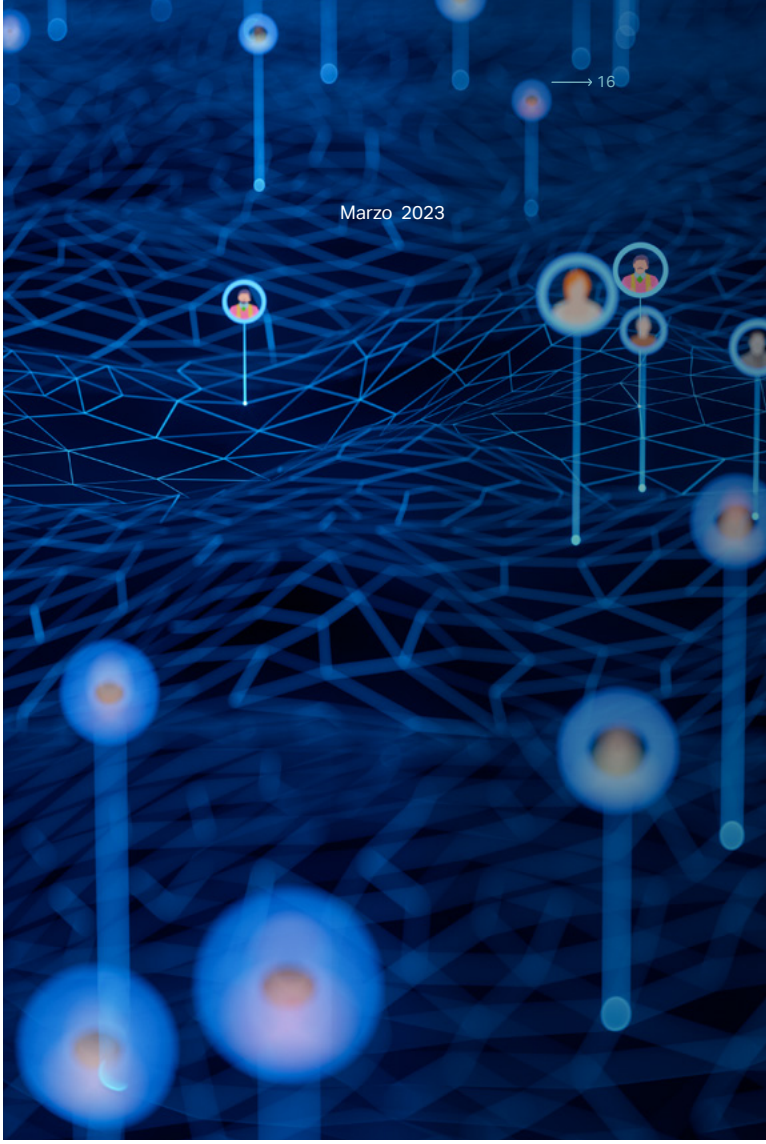
En relación a los identificadores indirectos, lo que se busca es que los datos en su conjunto cumplan con ciertas propiedades como la *k*-anonimidad, *l*-diversidad entre otras posibles definiciones. Independientemente de la propiedad elegida, lo que se busca es una métrica que permita cuantificar y comparar el resultado entre diversos métodos.

8 Las herramientas desarrolladas en español son limitadas y no se desempeñan bien en el dominio médico

k-anonimidad

Se dice que los datos satisfacen la *k-anonimidad* si y sólo si la combinación de valores de los identificadores indirectos que figuran en una tabla que se desea anonimizar aparecen al menos k veces (Sweeney, 2002). Si k = 10, cada combinación de variables (clase de equivalencia) debe tener al menos 10 registros. Esto previene que un ataque logre identificar un sólo caso dentro de ese grupo.

Para lograr esta condición, se emplean la generalización y la supresión que se explican a continuación.



Generalización

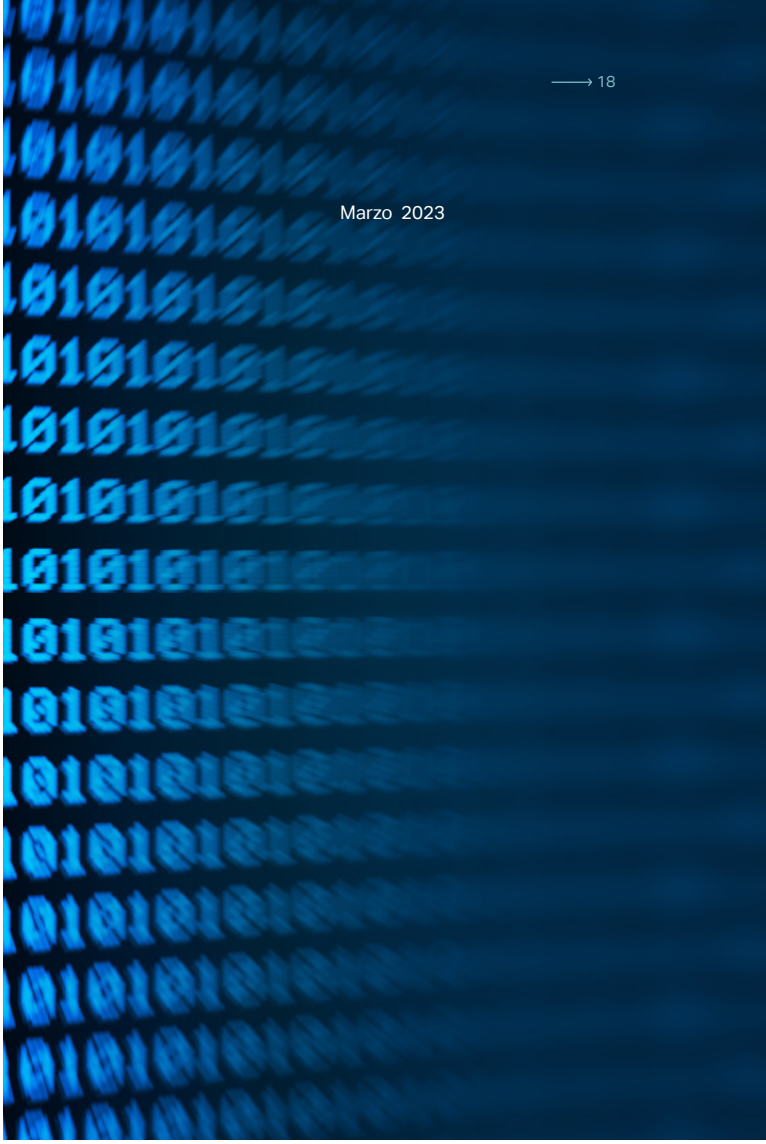
Consiste en reemplazar los valores puntuales por valores agrupados, menos específicos. Por ejemplo, una persona que aparece en los registros con la edad puntual de 38 años se la reemplaza por un intervalo que va desde los 30 a los 40 años. La columna 'edad agrupada' es el resultado del proceso de generalización a partir del valor 'edad' que es eliminado del conjunto de datos. Lo que se busca es reducir la granularidad de los datos, dificultando o incluso imposibilitando la recuperación de los valores puntuales asociados con un individuo.

Fila	ID_persona	Edad	Edad agrupada	Fecha diagnóstico	Práctica
1	persona 1	~~35~~	30-39	xx/xx/xxxx	pap
2	persona 1	~~35~~	30-39	xx/xx/xxxx	biopsia
3	persona 2	~~39~~	30-39	xx/xx/xxxx	biopsia
...	...	~~...~~	~~...~~	...	...
n	persona n	~~58~~	50-59	xx/xx/xxxx	biopsia

Supresión

La supresión consiste en remover los casos atípicos que por su baja frecuencia no pueden ser unidos en una clase de equivalencia ya sea porque son únicos en su clase (por ej. sólo una persona con edad en el rango de 50-59) o porque al generalizar implicaría ampliar una categoría de forma tal que se vuelve poco informativa (se amplía la categoría 40-49 a 40-60). La fila de la persona con edad extrema (58 años), se remueve del conjunto de datos.

Fila	ID_persona	Edad	Fecha diagnóstico	Práctica
1	persona 1	35	xx/xx/xxxx	pap
2	persona 1	35	xx/xx/xxxx	biopsia
3	persona 2	39	xx/xx/xxxx	biopsia
...	...	...	...	...
~~n~~	~~persona n~~	~~58~~	~~xx/xx/xxxx~~	~~biopsia~~



Perturbación de datos

Este método para anonimizar conjuntos de datos se puede aplicar a las entradas de datos numéricos, y consiste en reemplazar el valor de los atributos con valores aleatorios o partir de un listado, alterando los conjuntos de datos con un valor y una operación específicos. Este método modifica el conjunto de datos original mediante el uso de métodos de redondeo y ruido aleatorio.

Fila	ID_persona	Edad	Edad perturbada	Fecha diagnóstico	Práctica
1	persona 1	~~35~~	37	xx/xx/xxxx	pap
2	persona 1	~~35~~	34	xx/xx/xxxx	biopsia
3	persona 2	~~39~~	36	xx/xx/xxxx	biopsia
...	...	~~...~~	...	...	...
n	persona n	~~58~~	56	xx/xx/xxxx	biopsia

Intercambio de datos

Esta técnica lo que busca es alterar el orden o posición original de los elementos de un conjunto de datos ordenado. Se logra introduciendo una distorsión aleatoria en el conjunto de microdatos, que si bien mantiene el detalle y la estructura de la información original, no permite que los valores de los atributos puedan ser vinculados a sus registros originales.

Los microdatos son todos aquellos datos sobre las características de las unidades de estudio de una población -individuos, hogares, instituciones y/o empresas. En particular los datos anonimizados, son los registros individuales que se modifican para suprimir los identificadores directos e indirectos, a fin de minimizar los principales riesgos que conlleva su publicación.

Datos originales

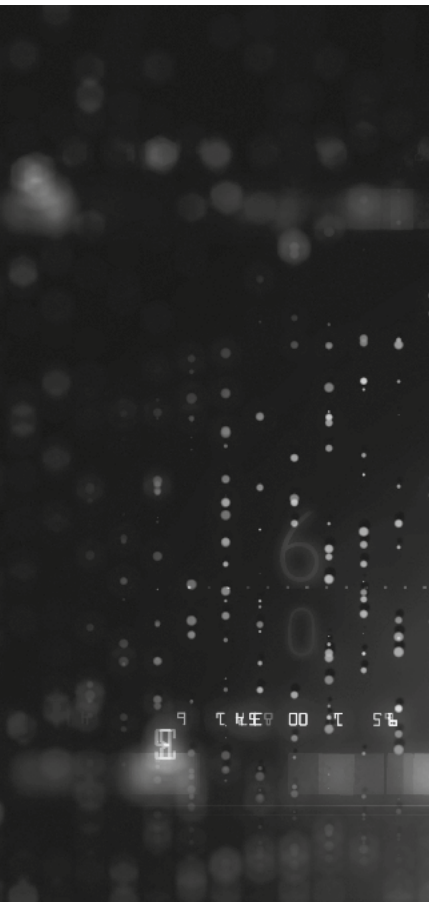
Fila	ID_persona	Edad	Fecha diagnóstico	Práctica
1	persona 1	35	xx/xx/xxxx	pap
2	persona 2	37	xx/xx/xxxx	biopsia
3	persona 2	39	xx/xx/xxxx	hisopado
4	persona 4	58	xx/xx/xxxx	extracción

Datos de intercambio

Fila	ID_persona	Edad	Fecha diagnóstico	Práctica
4	persona 1	58	xx/xx/xxxx	biopsia
1	persona 4	37	xx/xx/xxxx	extracción
2	persona 2	39	xx/xx/xxxx	pap
3	persona 3	35	xx/xx/xxxx	extracción

Datos sintéticos

Una alternativa a la anonimización de los datos para su uso secundario es la generación de registros sintéticos realistas. La posibilidad de contar con datos que poseen las características estadísticas y propiedades temporales de los originales permite obtener los mismos resultados analíticos a partir de datos efectivamente anónimos. Sin embargo, es importante tener en cuenta que esta es un área de incipiente desarrollo, por lo que aún dista de ser fácilmente aplicable.



Conclusiones

El presente documento constituye una guía práctica con recomendaciones para el tratamiento de datos personales en distintos escenarios de uso de la información, haciendo foco en el sector público de salud. La anonimización es el proceso de convertir los datos con el objetivo de que no sea posible identificar a las personas, y constituye una herramienta para mitigar al máximo posible los riesgos de trabajar masivamente con datos de carácter personal. En escenarios de uso de información sobre la salud de las personas la tarea de anonimizar es crucial, ya que además de tratar con datos personales tratamos con datos sensibles.

Para estos contextos de uso particular, se presentan una serie de técnicas que permiten ocultar la información sensible para difundirla sin vulnerar los derechos a la protección de datos de las personas. En escenarios en los cuáles dudamos de si se puede identificar a una persona con cierta información, conviene aplicar el principio precautorio y no disponibilizar la información.

Entendemos que el uso de la información pública procesada adecuadamente es un hito necesario para la elaboración de políticas públicas basadas en evidencia en el ámbito de la salud; para lograr ese objetivo es necesario evaluar cada nuevo escenario de uso de los datos con la complejidad necesaria para resguardar la información sensible. Por último, podemos identificar grandes avances en la sensibilización sobre estos problemas y un buen número de organizaciones ya implementan diferentes herramientas para cubrir vulnerabilidades y/o disponibilizar información de forma segura. Pretendemos seguir aportando a ese objetivo llegando con estas recomendaciones a los diversos perfiles que trabajan con datos en el sector salud.

Recursos

Hashing

- "Introducción al hash como técnica de seudonimización de datos personales" de la Agencia Española de Protección de Datos Personales.

- Guía práctica sobre uso de librería hashlib (https://docs.python.org/3/library/hashlib.html) en español https://fund.ar/wp-content/uploads/2022/07/Fundar_guia_practica_de_proteccion_de_datos_.pdf

K-anonimidad

- Guía práctica sobre uso de librería sdcMicro (https://www.rdocumentation.org/packages/sdcMicro/versions/5.6.0) en inglés: https://sdcpractice.readthedocs.io/en/latest/sdcMicro.html

Texto libre

- Guía práctica sobre uso de librería scrubadub (https://scrubadub.readthedocs.io/en/stable/) en español: https://fund.ar/wp-content/uploads/2022/07/Fundar_guia_practica_de_proteccion_de_datos_.pdf

- Prototipo de anonimizador de texto libre en español adaptado al dominio médico: https://github.com/instituciones-abiertas/anonimizacion-texto-libre

Referencias

• Ley Nº 25.326. Protección de los datos personales. Buenos Aires, Argentina, 4 de octubre de 2000.

• Ley Nº 27.275. Derecho de acceso a la información pública. Buenos Aires, Argentina, 14 de septiembre de 2016.

• Luvini, P. (2022). Guía práctica para la protección de datos. https://fund.ar/publicacion/guia-practica-para-la-anonimizacion-de-datos/

• Pérez Ponte, M. (2017). Protección de datos personales y el tratamiento de los datos de la salud. Diccionario Enciclopédico de la Legislación Sanitaria Argentina. https://salud.gob.ar/dels/entradas/proteccion-de-datos-personales-y-el-tratamiento-de-los-datos-de-la-salud

• Health Insurance Portability and Accountability Act. Pub. L. No. 104-191, § 264, 110 Stat.1936.

• HIPAA (2007), How Can Covered Entities Use and Disclose Protected Health Information for Research and Comply with the Privacy Rule?, National Institutes of Health, US Department of Health and Human Services.

• Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely, Data Privacy Working Paper 3, Carnegie Mellon University, Pittsburgh.

• Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of

• Uncertainty, Fuzziness and Knowlege-Based Systems, 10(5), 557–570.

• Sweeney, L., Abu, A., and Winn, J. (2013). "Identifying Participants in the Personal Genome Project by Name". White Paper 1021-1, Data Privacy Lab, Universidad de Harvard.

• Yankelevich, D. (2021). Anónimos pero no tanto. Fundar. https://fund.ar/publicacion/anonimimos-pero-no-tanto/

Acerca del equipo autoral

Sabrina Laura López

Licenciada y doctora en Ciencias Biológicas por la UBA. Postdoctoranda en el Instituto de Cálculo UBA-CONICET. Integrante del equipo de Uso Responsable de Datos del proyecto ARPHAI.

Laura Alonso Alemany

Doctora en Lingüística (UB), investigadora en Procesamiento del Lenguaje Natural, profesora en Ciencias de la Computación (UNC) y miembro del equipo de Ética en Inteligencia Artificial de la Fundación Via Libre. Miembro del equipo de Fenotipado y co-directora del equipo Uso Responsable de Datos del proyecto ARPHAI.

Juan Manuel Dias

Licenciado en Sociología por la UBA y maestrando en Estadística de la UNTREF. Es egresado de la carrera de ciencia de datos de la EANT y de la Diplomatura de Ciencias Sociales Computacionales de la UNSAM.

Laura Ación

Licenciada en Ciencias Biológicas por la UBA. Magíster en Salud Pública y Doctora en Bioestadística por la Universidad de Iowa (EEUU). Investigadora Adjunta del CONICET en el Instituo de Cálculo UBA-CONICET. Co-directora del equipo de Uso Responsable de Datos del proyecto ARPHAI.

Verónica Xhardez

Licenciada en Ciencias Antropológicas (UBA), Magíster en Ciencias Políticas y Sociología (Flacso) y Doctora en Ciencias Sociales (UBA). Invesitgadora en el CIECTI y Coordinadora Técnica del Proyecto ARPHAI.

Dirección ejecutiva: Martín Reydó

Coordinación editorial: Gonzalo Fernández Rozas

Diseño: Micaela Nanni

Revisión institucional: Juliana Arellano

Fundar es un centro de estudios y diseño de políticas públicas que promueve una agenda de desarrollo sustentable e inclusivo para la Argentina. Para enriquecer el debate público es necesario tener un debate interno: por ello lo promovemos en el proceso de elaboración de cualquiera de nuestros documentos. Confiamos en que cada trabajo que publicamos expresa algo de lo que deseamos proyectar y construir para nuestro país. Fundar no es un logo: es una firma.

El CIECTI (Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación) es un espacio abierto a la participación de otras entidades del mundo científico, productivo y social, que persigue el objetivo de generar y consolidar las capacidades institucionales para diseñar, implementar, monitorear y evaluar políticas en CTI. El proyecto "Gestión epidemiológica basada en inteligencia artificial y ciencia de datos" (ARPHAI) fue liderado por CIECTI durante mas de 2 años con la finalidad de desarrollar herramientas tecnológicas basadas en inteligencia artificial y ciencia de datos que permitan mejorar la toma de decisiones de salud pública preventiva.

Modo de citar

López, S.; Alonso Alemany, L.; Dias, J.M.; Ación, L. y Xhardez, V. (2023).. Guía práctica para la protección de datos personales en salud. Buenos Aires: Fundar. Disponible en https://www.fund.ar

Guía práctica para la protección de datos personales en salud / Laura Ación... [et al.].-
1a ed.- Ciudad Autónoma de Buenos Aires : Fundar , 2023.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-987-48985-3-1

1. Análisis de Datos. 2. Bases de Datos. 3. Sistemas de Gestión de Bases de Datos.
I. Ación, Laura.
CDD 005.8076

ISBN 978-987-48985-3-1

CIECTI
Centro Interdisciplinario de Estudios en Ciencia, Tecnología e Innovación

ARPHAI
Gestión epidemiológica basada en inteligencia artificial y ciencia de datos

fundar