

PROCEDIMIENTO DE DEPÓSITO DE DATOS DE INVESTIGACIÓN

RIMASALUD, REPOSITORIO INSTITUCIONAL DE LA CONSEJERÍA DE SANIDAD



This work is licensed under CC BY-NC-SA 4.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-sa/4.0/>

Contenido

Ciencia abierta y Datos de investigación	3
❖ La Ciencia Abierta.....	3
❖ Los Datos de investigación	3
❖ El Plan de Gestión de Datos (PGD).....	4
❖ Legislación vigente, límites legales y exenciones.	5
❖ Ventajas de depositar Datos en un Repositorio Institucional	7
DEPOSITA Datos de Investigación en RIMASalud	8
❖ Requisitos para depositar Datos de investigación en RIMASalud	8
❖ Los Principios FAIR	8
❖ Derechos de autor y licencias de uso	9
El Registro de Datos de investigación	10
1. Metadatos	11
2. Fichero de datos o <i>dataset</i>	11
3. El fichero “Readme.txt”	14
4. Licencia de distribución no exclusiva.	14
Anexo I. Técnicas de anonimización	15

RIMASalud es un espacio digital abierto que tiene como objetivo recoger, preservar y difundir la producción científica de los profesionales de la Consejería de Sanidad de la Comunidad de Madrid. El Repositorio proporciona mayor visibilidad e impacto, y cumple con los mandatos que suelen exigir los organismos financiadores de proyectos. Socialmente, RIMASalud permite el acceso y reutilización del conocimiento científico, contribuyendo a la disminución de la brecha digital de acceso a la información.

Ciencia abierta y Datos de investigación

❖ La Ciencia Abierta

La Ciencia Abierta (*Open Science*) es un movimiento internacional que tiene como objetivo la difusión del conocimiento científico, de forma libre y gratuita, así como la colaboración, transparencia y eficiencia de la investigación.

La Ciencia Abierta promueve el acceso abierto a las publicaciones (a través de repositorios y revistas *Open Access*), la evaluación en abierto (procesos de *peer review* y métricas del impacto de las publicaciones en abierto) y el acceso abierto a los datos de investigación que permitan su reutilización y reproducción.

❖ Los Datos de investigación

En la Estrategia Nacional de Ciencia Abierta (ENCA, 2023) se definen los datos de investigación como “todo aquel material que ha sido generado, recopilado, observado o registrado durante el ciclo de vida de un proyecto de investigación y que se utiliza como evidencia de un proceso de investigación, está reconocido por la comunidad científica y sirve para validar los resultados de la investigación y garantizar su reproducibilidad”. Los datos de investigación serán, entre otros, “resultados de experimentos, mediciones, observaciones resultantes del trabajo de campo, resultados de encuestas, grabaciones de muestras físicas, documentación, código, configuraciones de instrumentación, estadísticas e imágenes.”.

Por lo tanto, no todos los datos son datos de investigación:

Sí son datos de investigación	No son datos de investigación
☐ Datos tabulares (Excel, csv, etc)	☐ Notas o cuadernos de laboratorio
☐ Imágenes, vídeos o gráficos	☐ Comunicaciones con colegas o emails
☐ Textos o audios de entrevistas	☐ Borradores de trabajos
☐ Datos de experimentos o encuestas	☐ Notas de otras investigaciones
☐ Grabaciones de muestras físicas	☐ Muestras de laboratorio / físicas
☐ Estadísticas	☐ Notas o apéndices técnicos de proyectos



o Códigos de software, bases de datos, etc.	
---	--

❖ El Plan de Gestión de Datos (PGD)

El PGD es un documento “vivo” que se realiza desde el inicio de la planificación del proyecto de investigación y describe cómo serán gestionados los datos, recogidos o generados, durante y después de finalizar. Es un documento en evolución, cuyos apartados se van modificando y ampliando conforme evoluciona el propio proyecto.

La mayoría de las convocatorias de financiación de proyectos exigen la elaboración del PGD¹ y el depósito de los datos de investigación en un repositorio de acceso abierto seguro (*“tan abierto como sea posible, tan cerrado como sea necesario”*).

Qué información debe incluir un Plan de Gestión de Datos:

- o Presentación del proyecto de investigación en el que se inscriben los datos.
- o Descripción de los datos que se van a recoger o generar: tipología y procedencia.
- o Organización: cómo se van a recolectar o generar los datos. Procedimiento. o Autoría de los datos: quiénes realizarán la recolección o generación de los mismos. o Análisis de los datos.
- o Gestión de los datos: software, control de versiones, etc.
- o Preservación: dónde y cómo se van a almacenar los datos. Qué Repositorio.
- o Reutilización: cómo se van a compartir los datos y se va a facilitar el acceso a los datos generados. Licencias de uso.
- o Cuestiones éticas y legales.

Cómo elaborar un Plan de Gestión de Datos

Para realizar un PGD se recomienda acudir primero a la propia institución donde se vaya a desarrollar el proyecto de investigación para conocer si ya cuenta con directrices concretas y una plantilla.

Si el proyecto está dentro de una convocatoria y/o cuenta con financiación pública o de alguna institución privada, es igualmente imprescindible consultar primero qué requisitos se establecen para la elaboración del Plan de Gestión de Datos.

Si no existe una plantilla específica, se puede acudir a diversos servicios para la creación/gestión de PGDs²:

¹ H2020 (desde 2017) y Horizonte Europa (2021-27).

² Si el proyecto no está sometido a exigencias específicas de una convocatoria o reglamentación concretas, no tiene por qué incluir todos los apartados que aparecen en los distintos formularios disponibles.



- [DMPOnline del Digital Curation Centre](#): permite crear y gestionar en línea PGDs a través de un formulario al que se va respondiendo. Existe un checklist para comprobar que se ha incluido todo lo que debe incluir un PGD ([Checklist for a Data Management Plan](#)³)
- [Argos](#): plataforma de la Unión Europea para PGDs. Puede registrarse cualquier persona. Disponible en español.
- [PGDOnline](#): del Consorcio Madroño

Para conocer de forma gráfica en qué consiste un PGD se puede consultar la infografía de Rebiun: [10 pasos para elaborar un Plan de Gestión de Datos](#).

❖ Legislación vigente, límites legales y exenciones.

Legislación vigente

La normativa europea y su transposición a la legislación española promueven el acceso abierto a los datos de investigación para facilitar su accesibilidad, interoperabilidad y reutilización.

Normativa europea:

- [DIRECTIVA \(UE\) 2019/1024 DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 20 de junio de 2019](#) relativa a los datos abiertos y la reutilización de la información del sector público (versión refundida).
- [REGLAMENTO \(UE\) 2021/695 DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 28 de abril de 2021](#), por el que se crea el Programa Marco de Investigación e Innovación «Horizonte Europa», se establecen sus normas de participación y difusión, y se derogan los Reglamentos (UE) nº 1290/2013 y (UE) nº 1291/2013.

Normativa española:

- [Real Decreto-ley 24/2021](#), de 2 de noviembre, por el que se transpone la [DIRECTIVA \(UE\) 2019/1024 DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 20 de junio de 2019](#) relativa a los datos abiertos y la reutilización de la información del sector público.
- [Ley 17/2022](#), de 5 de septiembre, por la que se modifica la Ley 14/2011, de 1 de junio, de la Ciencia, la Tecnología y la Innovación

Otros documentos de referencia

- Programa Marco de la Unión Europea - [HORIZONTE EUROPA \(2021-2027\)](#)
- [Estrategia Nacional de Ciencia Abierta \(ENCA, 2023\)](#)

³ La Universidad de Cantabria ha traducido al español esta plantilla: Universidad_Cantabria_Guia_modelo_basico_PGD.pdf



Límites legales al acceso abierto a los datos

Aunque la Ciencia abierta promueva el acceso abierto a los datos de investigación, hay que tener en cuenta que existen límites legales en cuanto a protección de datos se refiere, para mantener el derecho a la confidencialidad, privacidad y protección de datos personales. Debe tenerse en cuenta tanto la legislación nacional como la que establecen las distintas administraciones públicas:

- [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).
- [Ley Orgánica 3/2018](#), de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. BOE, 6 de diciembre de 2018.
- [Orden SSI/81/2017, de 19 de enero](#), por la que se publica el Acuerdo de la Comisión de Recursos Humanos del Sistema Nacional de Salud, por el que se aprueba el protocolo mediante el que se determinan pautas básicas destinadas a asegurar y proteger el derecho a la intimidad del paciente por los alumnos y residentes en Ciencias de la Salud.
- De manera adicional, resulta pertinente prestar especial atención a las distintas guías, informes y circulares emitidas por la autoridad de control española, esto es la [Agencia Española de Protección de Datos \(AEPD\)](#), en tanto que constituyen instrumentos interpretativos y orientativos de gran relevancia para la correcta aplicación de la normativa vigente en materia de protección de datos.

Entre los límites legales al acceso abierto a los datos de cualquier investigación están⁴:

- La protección de los datos personales.
- La protección de recursos genéticos.
- Los derechos de propiedad industrial.
- La protección de datos relacionados con asuntos de seguridad.
- La protección de datos que puedan poner en riesgo el propio proyecto de investigación.

Exenciones

El [Reglamento europeo \(2016/679/UE\)](#) también protege el derecho al acceso abierto a los datos estableciendo exenciones para la recopilación de datos para la investigación:

- Artículo 5.1.b) del RGPD hace referencia al principio de limitación de la finalidad y a su vinculación con el interés público

⁴ Véase también el apartado de *Licencias restrictivas* en el documento “Deposita Datos de investigación en RIMASalud”.



- Artículo 6.1 del RGPD establece las bases jurídicas relativas al consentimiento.
- Artículo 9.2.j del RGPD dispone que “El tratamiento es necesario con fines de archivo en interés público, fines de investigación científica o histórica o fines estadísticos, de conformidad con el artículo 89, apartado 1, sobre la base del Derecho de la Unión o de los Estados miembros, que debe ser proporcional al objetivo perseguido, respetar en lo esencial el derecho a la protección de datos y establecer medidas adecuadas y específicas para proteger los intereses y derechos fundamentales del interesado”

❖ **Ventajas de depositar Datos en un Repositorio Institucional**

- Ofrece información científica de calidad, abierta y gratuita, a la ciudadanía.
- Facilita a cualquier persona el acceso a la investigación que se realiza con fondos públicos.
- Aporta transparencia a la investigación pública y fomenta la innovación.
- Permite a los investigadores el depósito de datos de forma gratuita.
- Aporta difusión y visibilidad a la investigación científica.
- Facilita la colaboración entre los usuarios de datos y sus creadores, y la posibilidad de generar nuevos datos.
- Permite medir el impacto de la producción científica de los investigadores.
- Facilita la reutilización de los resultados de la comunidad científica.

...y en RIMASalud:

- Contribuye a difundir y visibilizar la producción científica de los profesionales de la Consejería de Sanidad.
- Reduce los costes de suscripción a publicaciones científicas.
- Mide e informa sobre el impacto de la producción científica de los centros e instituciones integradas en el Sistema Sanitario Público madrileño.

DEPOSITA Datos de Investigación en RIMASalud

❖ Requisitos para depositar Datos de investigación en RIMASalud

- o Puede depositar cualquier investigador/a perteneciente a centros o servicios dependientes de la Consejería de Sanidad de la Comunidad de Madrid.
- o Los datos depositados deben cumplir con la legislación vigente.
- o Los datos de investigación deben ajustarse a los principios FAIR.
- o Los datos de investigación deben ser originales salvo que reutilicen datos ya compartidos bajo los principios FAIR.
- o El autor/a debe firmar una licencia de distribución no exclusiva¹.

❖ Los Principios FAIR

Son los principios que permiten que los datos de investigación sean localizables, accesibles, interoperables y reutilizables, en cumplimiento con los mandatos de la Ciencia abierta.

Los Repositorios que cumplen con los principios FAIR, como RIMASalud, garantizan los siguientes aspectos¹:

🚦 Localizable (Findable)

- F1. A cada registro (metadatos) se le asigna un identificador único y persistente.
- F3. Los (meta)datos de cada registro se indexan en un sistema con motor de búsqueda.
- F4. Los metadatos contienen un campo específico para el identificador del dataset.

🚦 Accesible (*Accesible*):

- A1. Los (meta)datos se pueden recuperar a través de su identificador mediante protocolo de comunicación estandarizado.
 - A1.1 El protocolo es abierto, gratuito y universalmente aplicable.
 - A1.2 El protocolo permite un procedimiento de autenticación y autorización, cuando sea necesario.
- A2. Los metadatos serán accesibles, incluso cuando los datos ya no están disponibles.

🚦 Interoperable (Interoperable):

- I1. Los (meta)datos utilizan un lenguaje formal, accesible, compartido y ampliamente aplicable para la representación del conocimiento.
- I2. Los (meta)datos usan vocabularios que siguen principios FAIR.

🚦 Reutilizable (Reusable)

- R1. Los (meta)datos disponen de atributos precisos y pertinentes.
- R1.3. Los (meta)datos cumplen con los estándares propios de una comunidad de un área de conocimiento determinada.

Los/as autore/as deben cumplir con los principios FAIR en las cuestiones siguientes:

✚ **Localizable (*Findable*)**

F2. Los datos se describen con metadatos enriquecidos.

Además de los campos obligatorios, se debe incluir una descripción lo más completa de los datos.

✚ **Interoperable (*Interoperable*):**

I1. Los (meta)datos utilizan un lenguaje formal, accesible, compartido y ampliamente aplicable para la representación del conocimiento.

Los metadatos deben incluir, preferiblemente, descriptores semánticos (MeSH, DeSC).

I3. Los (meta)datos deben incluir referencias calificadas a otros (meta)datos, cuando proceda.

✚ **Reutilizable (*Reusable*)**

R1.1. Los (meta)datos deben incluir una licencia de reutilización clara y accesible.

R1.2. Los (meta)datos deben estar asociados con su procedencia.

Recuerda: cuando hayas finalizado el proceso de Autoarchivo de tus datos, es conveniente que compruebes la calidad del registro en relación a los principios FAIR. Te indicamos algunas páginas que te permiten comprobarlo:

- [F-UJI](#)
- [FAIR eva](#)
- [FAIR Data Self-Assessment Tool](#)

❖ **Derechos de autor y licencias de uso**

Lo/as autore/as deberán elegir una licencia de uso para sus datos que, al mismo tiempo que conserve los derechos de propiedad intelectual, permita cumplir con los requisitos de acceso abierto para su reutilización.

Previamente, los autores deben tener en cuenta qué tipo de acuerdo firman con los editores de revistas que tengan ya políticas propias de difusión de datos de investigación vinculados a artículos científicos. Por ello es fundamental que, antes de elegir el lugar de publicación, los autores revisen convenientemente las políticas editoriales de cada revista a este respecto.

Licencias OpenData Commons

Son licencias de uso específicas para datos. Dentro del marco de la legislación de propiedad intelectual española se recomienda elegir entre:

- **Open Data Commons Attribution License (ODC-By):**
Licencia que permite difundir, reutilizar y adaptar los datos siempre que exista el reconocimiento a *la autoría del conjunto de datos originales y a sus términos de uso*.
- **Open Data Commons Open Database License (ODbL):**

Licencia que permite difundir, reutilizar y adaptar los datos siempre que exista el reconocimiento de la autoría de los datos originales, se mantenga la misma licencia de uso en obras derivadas y se garantice la disponibilidad de una copia en acceso abierto en versiones futuras o adaptaciones.

Para conocer las características completas de cada una de estas licencias, puedes visitar la web de Open Knowledge International: <https://opendefinition.org/licenses/>

Licencias Creative Commons

Son modelos de contrato que permiten ofrecer con carácter público la utilización de los documentos originales por terceros.

Horizonte Europa recomienda el uso de una de estas dos licencias¹:

- CC BY, licencia con reconocimiento de autoría.
- CC0 1.0., licencia que ofrece los datos en dominio público sin restricciones.

Actualmente RIMASalud utiliza licencias CC 4.0.

Licencias restrictivas

Existen circunstancias legales que obligan a limitar el acceso abierto a los datos, como ya se ha explicado en el apartado más arriba, “Legislación vigente, límites legales y excepciones”: cláusulas de confidencialidad, protección de información personal, protección de recursos genéticos, etc.

Para estos casos existen licencias específicas de carácter restrictivo. Sin embargo, en RIMASalud solo se podrán depositar aquellos datos de investigación que sí permitan cumplir con los principios FAIR de Reutilización y que puedan dotarse con licencias de uso Creative Commons u OpenDataCommons.

El Registro de Datos de investigación

Los registros de datos de investigación en RIMASalud deben constar de 4 elementos imprescindibles:

1. Metadatos
2. Fichero/s de datos o *dataset*
3. Fichero Readme.txt
4. Licencia de distribución no exclusiva



1. Metadatos

La adecuada descripción de los datasets a través de los metadatos es la mejor forma de conseguir su accesibilidad, recuperación, visibilidad y preservación. Por eso, además de los campos obligatorios, es recomendable enriquecer los metadatos aportando la mayor información posible, mejorando así su calidad y potenciando los principios FAIR.

En la opción de Autoarchivo, RIMASalud irá guiando a los autores en el proceso de cumplimentar los metadatos. Para conocer el proceso completo y detallado acceda a la [“Guía de Autoarchivo”](#).

2. Fichero de datos o dataset

i. Tipología de datos.

Los datos que se pueden subir a RIMASalud son:

- Textos
- Imágenes/fotografías
- Videos
- Grabaciones sonoras
- Bases de datos
- Datos tabulares
- Estadísticas

ii. Formatos de almacenamiento según tipología de datos:

Se recomienda que el formato elegido sea de código abierto sin software propietario (ej. PDF/A mejor que Word, ASCII mejor que Excel, MPEG-4 mejor que QuickTime, TIFF o JPEG2000 mejor que GIF o JPG, etc.). Por eso es conveniente, antes de iniciar el Autoarchivo, convertir los datos a dichos formatos.

La documentación y especificaciones del software utilizado para el formato deben estar disponibles gratuitamente en Internet, de forma que cualquier persona tenga acceso sin límite ni restricciones por derechos de propiedad intelectual.

En caso de que no sea posible, se elegirán formatos ampliamente utilizados en el momento de la difusión de los datos, para facilitar la reutilización, la usabilidad y la preservación a largo plazo.

Los formatos recomendados para cada tipo de documento son:

Tipología documental	Formatos recomendados
Textos	PDF/A, ASCII, XML, HTML, JSON, TXT, RTF
Imágenes/fotografías	TIFF, JPEG 2000, PDF, DNG, GIF, BMP, SVG
Videos	MPEG, AVI, MXF, MKV
Grabaciones sonoras	WAVE, AIFF, MP3, MXF, FLAC

Bases de datos	XML, CSV, JSON
Datos tabulares	CSV, TXT, TSV
Estadísticas	DTA, POR, SAS, SAV
Formato comprimido	ZIP

Para ampliar información sobre formatos admitidos y recomendados están disponibles estas páginas:

- Library of Congress de Estados Unidos:
<https://www.loc.gov/preservation/resources/rfs/>
- UK Data Service:
<https://www.ukdataservice.ac.uk/managedata/format/recommended-formats.aspx>

iii. Límite de tamaño de los ficheros de datos.

Los archivos subidos a RIMASalud no deben superar el límite de 1GB. En caso de ficheros más grandes, deberá realizarse previamente bien una partición del mismo bien una conversión a un fichero comprimido.

iv. Preparación del dataset.

Antes de subir los datos de investigación a RIMASalud hay que seguir una serie de pasos:

- o **Seleccionar** los datos a depositar, conforme al Plan de Gestión de Datos previamente elaborado.
- o **Depurar** los datos.
Antes de subir los datos es necesario realizar una depuración exhaustiva para eliminar duplicados, corregir errores o añadir datos incompletos.

También se debe tener en cuenta la coherencia y uniformidad de los datos para que puedan ser reutilizados (mismas unidades de medición, mismos parámetros, mismo lenguaje y abreviaturas, etc.).

En caso de datos tabulares, cada variable se expresará en una columna independiente. Y cada columna tendrá indicado su título. Asimismo, cada observación distinta de la variable se expresará en una fila diferente. Si existen múltiples tablas, deberá existir una columna en cada tabla que permita enlazarlas.

Para ayudar en la labor de depuración se puede utilizar la herramienta OpenRefine:
<https://openrefine.org/>

- **Anonimizar** los datos:

Si el conjunto de datos contiene algún tipo de información personal, ésta deberá anonimizarse antes de cerrar el documento y subirlo al Repositorio para garantizar y preservar la privacidad y el derecho a la protección de datos de carácter personal (ver más arriba el apartado de Límites legales).

Al mismo tiempo habrá que asegurarse de que, después del proceso de anonimización, el conjunto de datos mantenga la calidad, utilidad y accesibilidad necesarias para que sea reutilizable, estableciendo un equilibrio con la protección de datos. El lema fundamental de los principios FAIR es: “**tan abierto como sea posible, tan cerrado como sea necesario**”.

Una de las herramientas actuales que facilitan esta labor es Amnesia:
<https://amnesia.openaire.eu/>

Indicamos a continuación la diferencia entre **datos anonimizados** y **datos seudonimizados**:

La **seudonimización** consiste en el tratamiento de datos personales de tal manera 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, es decir, la seudonimización permite la reidentificación de los sujetos.

Por otra parte, la **anonimización** tiene como finalidad eliminar o reducir al mínimo los riesgos de reidentificación de los datos anonimizados manteniendo la veracidad de los resultados del tratamiento de los mismos, es decir, en el proceso de anonimización se debe producir la ruptura de la cadena de la identificación de las personas, evitando que las mismas se puedan identificar.

Por ello, se concluye que los datos seudonimizados son más vulnerables a la identificación que los datos anonimizados y en consecuencia, el tratamiento de los datos seudonimizados deberá cumplir con la normativa vigente en materia de protección de datos, al igual que los datos originales, dado que la normativa vigente (LOPDGDD/RGPD) los sigue considerando datos personales. La diferencia de considerar un dato anonimizado o seudonimizado es que en el primer caso no se pueden establecer asociaciones con otros datos y fuentes de información que puedan llegar a la reidentificación del sujeto.

En base a lo anterior, en **RIMASalud** los **datos de investigación** deben estar siempre **anonimizados** para que no se posible en ningún caso una reidentificación de los interesados.

Sobre las distintas técnicas de anonimización véase el Anexo I.

○ **Nombrar** los archivos:

Los archivos deben estar nombrados de manera lógica. El nombre debe reflejar el contenido e incluir suficiente información para hacerlo unívoco. La mejor práctica es:

- Mantener una coherencia interna y seguir siempre el mismo criterio de denominación para cada fichero.
- Utilizar nombres significativos y breves.
- Identificar la actividad o proyecto en el nombre del archivo: puede contener el acrónimo del proyecto, el nombre de los investigadores, el tipo de datos y el lugar de estudio.
- Se aconseja indicar la versión (v01, v02...): enumerar las versiones de los archivos es útil para indicar las revisiones o la edición de archivos, especialmente en los proyectos de colaboración. Se recomienda que cada versión constituya un registro independiente (ver Guía de Autoarchivo: Relaciones entre registros).
- Indicar el tipo de fichero: conservar las extensiones de 3 letras específicas a códigos de aplicaciones como .wrl, .mov, o .tif.
- Evitar el uso de caracteres especiales o espacios (mejor utilizar el guión bajo).
- Si se utilizan abreviaturas, deben explicarse en el fichero Readme.txt.
- Si aparece la fecha, utilizar un formato estandarizado (ej. AAAAMMDD). ISO 8601.
- Si existen varios archivos debe utilizarse una numeración secuencial con ceros a la izquierda para garantizar el orden: 01, 02...10

Ejs. 02_Surveydaten2021_V02_20210517_FER.tif
20230712_DOEProject_Ex1Test1_Data_Garcia_v03.xlsx

3. El fichero “Readme.txt”

Además del archivo o archivos que constituyan los datos de investigación, cada registro deberá incluir obligatoriamente un Fichero Readme.txt que tendrá como finalidad documentar los datos, incluyendo tanto información general como específica, para permitir su mejor reutilización.

A continuación, facilitamos una [plantilla modelo](#) para elaborar el fichero Readme.txt.

4. Licencia de distribución no exclusiva.

Los autores deben firmar esta licencia que les permitirá, a un mismo tiempo, preservar y difundir los datos en el Repositorio y mantener otros usos o vías de difusión, siempre y cuando no contravengan las licencias editoriales que hayan sido firmadas.

Por esta razón, se recomienda a los autores que publiquen en revistas de acceso abierto que permitan el archivo en repositorios, así como conservar la versión final de sus trabajos revisados por pares y no firmar licencias de uso exclusivas con los editores.

La licencia de distribución no exclusiva del Repositorio Institucional para Datos de Investigación puede descargarse clicando [aquí](#).



Anexo I. Técnicas de anonimización

- Eliminación de datos identificativos directos del paciente excepto el CIPA (si bien este dato se tratará conforme exponemos a más adelante). Estos datos identificativos deben, asimismo, ser eliminados también de los datos no estructurados, ya que pueden darse casos en los que se identifiquen a los pacientes o a los familiares.
 - Enmascaramiento de caracteres. Consiste en la **sustitución de algunos caracteres del valor de datos por un símbolo** consistente (por ejemplo, * o X). Normalmente se aplica solo a algunos caracteres parte del atributo y su uso puede ser más adecuado en identificadores directos, como un código postal o DNI. Dependiendo del tipo de atributo, se puede decidir **reemplazar un número fijo de caracteres o un número variable**.
- o Anonimización de datos clínicos: siguiendo la guía de anonimización del Comité Europeo de Protección de Datos (CEPD) y de la AEPD, se aplicarán las siguientes medidas recogidas en el Dictamen 05/2014 sobre técnicas de anonimización:
- **Aleatorización:** La aleatorización es una familia de técnicas que modifican la veracidad de los datos a fin de eliminar el estrecho vínculo existente entre los mismos y la persona. Si los datos se hacen lo suficientemente ambiguos, no podrán remitir a una persona concreta. La aleatorización por sí sola no reduce la singularidad de cada uno de los registros, ya que estos pueden obtenerse a partir de un único interesado, pero sí puede proteger contra ataques o riesgos de inferencia. Además, se puede combinar con técnicas de generalización para obtener mayores garantías de privacidad. Es posible que haya que aplicar otras técnicas para garantizar que un registro no sirve para identificar a una persona.
 - **Ruido:** la técnica de adición de ruido es especialmente útil cuando los atributos pueden causar un importante efecto adverso en las personas. Consiste en modificar los atributos del conjunto de datos para que sean menos exactos (incluyendo por ejemplo datos falsos), conservando no obstante su distribución general. Al tratar un conjunto de datos, cualquier observador supondrá que los valores son exactos, pero esto solo es cierto hasta cierto punto. Se recomienda incluir entre un 5% y 10% de ruido en un proceso de anonimización.
 - **Generalización:** La generalización es la segunda familia de técnicas de anonimización. Este enfoque generaliza o diluye los atributos de los interesados modificando las respectivas escalas u órdenes de magnitud (por ejemplo, sustituyendo una ciudad por una región, o una semana por un mes). Aunque la generalización pueda ser efectiva para descartar la singularización, no permite obtener una anonimización eficaz en todos los casos; en concreto, es necesario aplicar enfoques cuantitativos específicos y complejos para impedir la vinculabilidad y la inferencia.
 - **Diversidad L:** es una técnica similar al anonimato k, pero en lugar de aplicarlos a un único campo se aplica a varios grupos de campos. Tienen el objetivo de impedir que

un interesado sea singularizado cuando se le agrupa junto con, al menos, un número L de personas. Para lograrlo, los valores de los atributos se generalizan hasta el punto de que todas las personas acaban compartiendo el mismo valor.

Aplicando las dos técnicas combinadas (ruido + diversidad I), se reduce la singularización, vinculabilidad y la inferencia. No obstante, es importante recordar que tal y como transmite el CEPD y la AEPD, la anonimización no implica una máxima garantía sobre la reidentificación de los datos por parte de terceros, existiendo siempre la posibilidad de que terceros malintencionados pudieran, a futuros, identificar a los sujetos.

En ese sentido, se debe tener presente que, debido a las distintas técnicas de anonimización que se lleven a cabo, cabe la posibilidad de que la posterior reidentificación que realice de un ciudadano, contenga algún “dato falso” de los incluidos por el proveedor. Si bien es una posibilidad mínima, debe conocerse que se trata de un riesgo a asumir.

Para la correcta aplicación de estas técnicas deberá seguirse la Guía básica de anonimización publicada por la AEPD: <https://www.aepd.es/documento/guia-basica-anonimizacion.pdf>